Google Developer Group
Editable University Name

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

GDGoC INU AI Part Paper Seminar

AI Member 양승빈

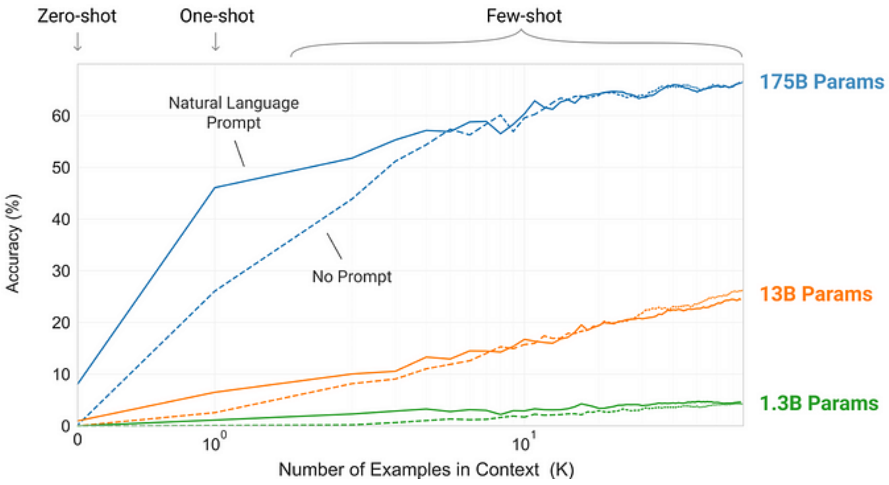
INDEX

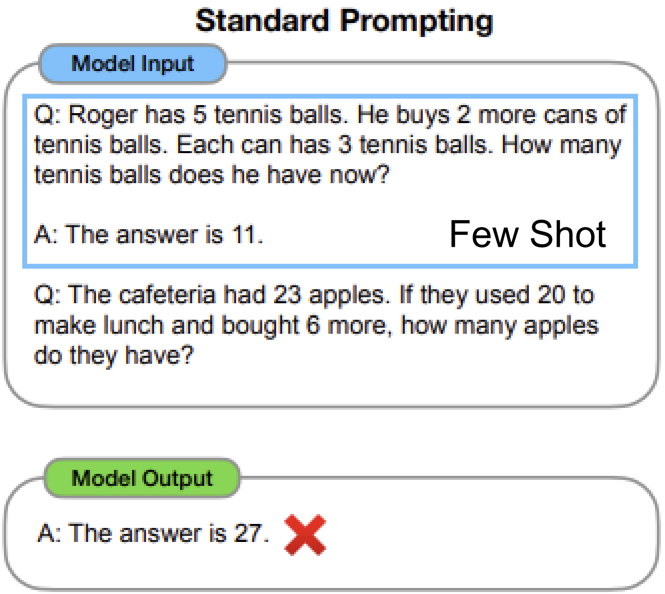
0. Abstract
1. Introduction
2. CoT Prompting
3. Arithmetic Reasonig
4. Commonsense Resoning
5. Symbolic Reasoning
6. Conclusions

0. Abstract

- (목적) Chain-of-Thougt를 통해 LLM의 추론 능력을 어떻게 향상시킬 수 있는지 탐구
- (결과) Chain-of-Thougt Prompting은 산술, 상식, 기호 추론 작업 성능 향상

1. Introduction





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.

Few Shot

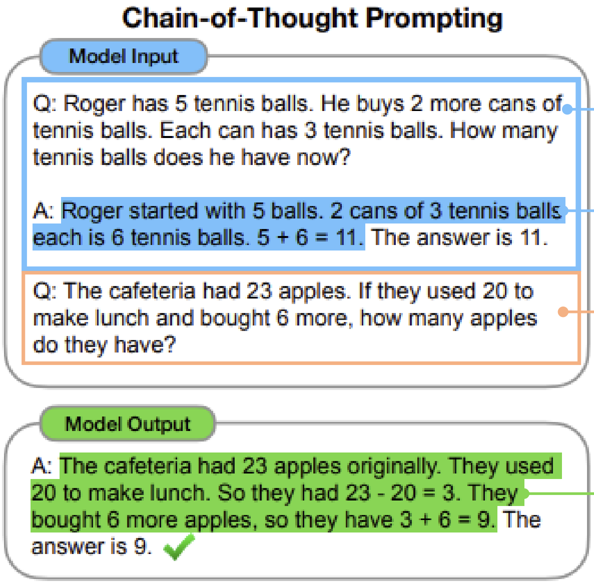
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Few-shot: 적은 수의 예시로 새로운 작업을 학습하고 수행할 수 있도록 하는 방식
(Format: 질문+답)

1. Introduction



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. ✓

Few Shot

Chain of Thought

Question

Effect of chain of Thought

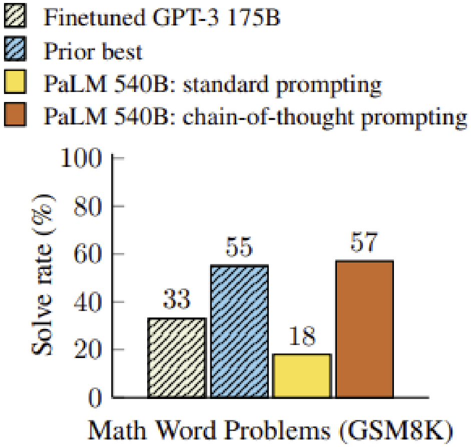
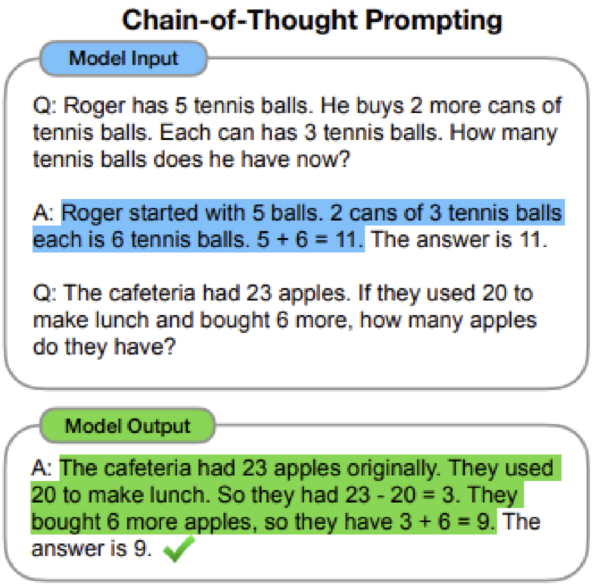


Figure 2: PaLM 540B uses chain-of-thought prompting to achieve new state-of-the-art performance on the GSM8K benchmark of math word problems. Finetuned GPT-3 and prior best are from Cobbe et al. (2021).

2. Chain-of-Thought Prompting

- Properties



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. ✓

- 다단계 문제 해결
- 해석 가능성
- 범용성
- 기존 모델 활용 가능

3. Experimental Setup

구분	LLM	Prompting	Bechmark
Arithmetic	• GPT-3: 350M, 1.3B, 6.7B, 175B • LaMDA: 422M, 2B, 8B, 68B, 137B • PaLM: 8B, 62B, 540B • UL2: 20B • Codex	• Standard: few-shot • Chain-of-Thought: CoT	• GSM8K • SAVMP • ASDiv • AQUA • MAWPS
Commonsense			• CSQA • SrategyQA • Date • Sport • SayCan
Symbolic			• Last letter concatenation • Coin flip

3. Experimental Setup

- Benchmark

분류	기법	설명
Arithmetic	GSM8K (Grade School Math 8K)	초등학생 수준 수학 단어 문제 해결 능력 평가
	SAVMP (Solving Arithmetic Verbal Math Problems)	수학 어휘 문제 해결 능력 평가
	ASDiv (Addition and Subtraction word problems with DIversity)	덧셈과 뺄셈 문제 해결 능력 평가
	AQUA (Arithmetic Questions with Ambiguous Numbers)	상식적 추론 및 수학적 능력 조합 평가
	MAWPS (Math Word Problem Sets)	다양한 난이도 수학 단어 문제 해결 능력 평가
Commonsense	CSQA (Common Sense Question Answering)	상식적인 질문 답변 능력 평가
	StrategyQA	다단계 추론 요구 질문 답변 능력 평가
	Date	날짜 및 시간 정보 이해 및 처리 능력 평가
	Sport	스포츠 관련 질문 답변 능력 평가
	SayCan	도구 사용 및 계획 능력 평가
Symbolic	Last letter concatenation	단어 마지막 글자 연결 능력 평가
	Coin flip	동전 던지기 관련 확률적 추론 평가

3. Arithmetic Reasoning

- Examples

Q. There are 15 trees is 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?

<MATH WORD PROBLEMS>

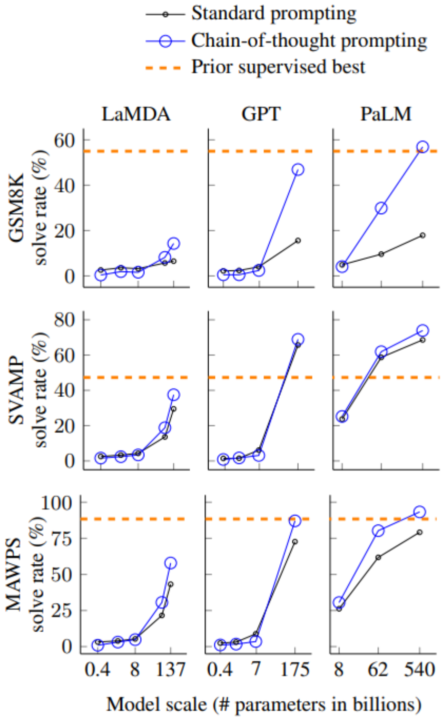
A. There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21−15=6. The answer is 6.

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

<AQUA ALGEBRAIC WORD PROBLEMS >

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).

3. Arithmetic Reasoning



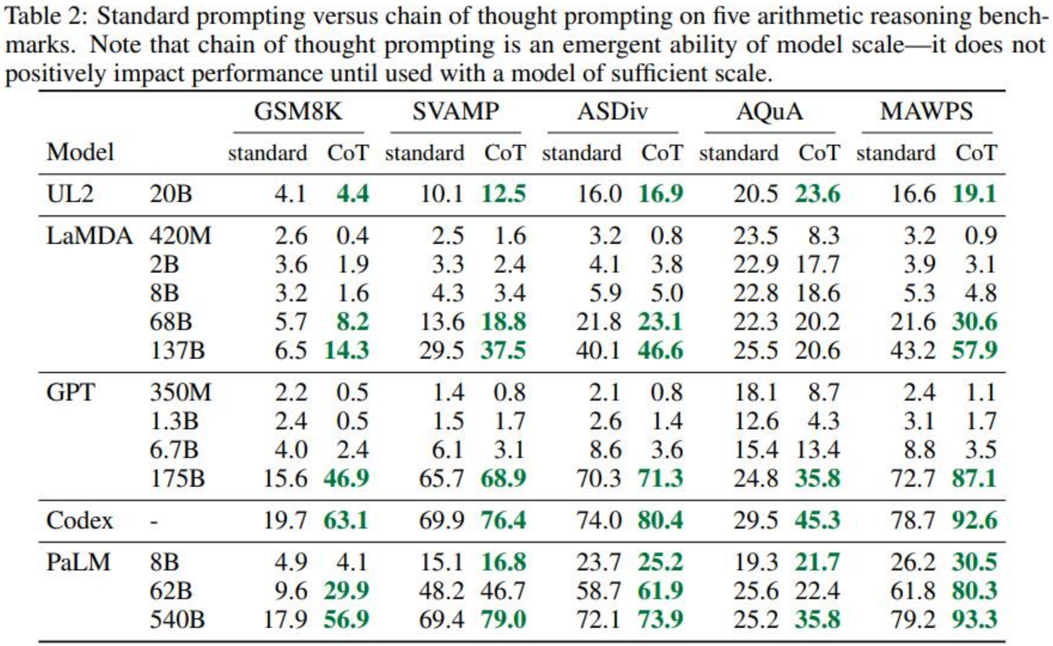


Table 2: Standard prompting versus chain of thought prompting on five arithmetic reasoning benchmarks. Note that chain of thought prompting is an emergent ability of model scale—it does not positively impact performance until used with a model of sufficient scale.

		GSM8K		SVAMP		ASDiv		AQuA		MAWPS	
Model		standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	4.1	**4.4**	10.1	**12.5**	16.0	**16.9**	20.5	**23.6**	16.6	**19.1**
LaMDA	420M	2.6	0.4	2.5	1.6	3.2	0.8	23.5	8.3	3.2	0.9
	2B	3.6	1.9	3.3	2.4	4.1	3.8	22.9	17.7	3.9	3.1
	8B	3.2	1.6	4.3	3.4	5.9	5.0	22.8	18.6	5.3	4.8
	68B	5.7	**8.2**	13.6	**18.8**	21.8	**23.1**	22.3	20.2	21.6	**30.6**
	137B	6.5	**14.3**	29.5	**37.5**	40.1	**46.6**	25.5	20.6	43.2	**57.9**
GPT	350M	2.2	0.5	1.4	0.8	2.1	0.8	18.1	8.7	2.4	1.1
	1.3B	2.4	0.5	1.5	1.7	2.6	1.4	12.6	4.3	3.1	1.7
	6.7B	4.0	2.4	6.1	3.1	8.6	3.6	15.4	13.4	8.8	3.5
	175B	15.6	**46.9**	65.7	**68.9**	70.3	**71.3**	24.8	**35.8**	72.7	**87.1**
Codex	-	19.7	**63.1**	69.9	**76.4**	74.0	**80.4**	29.5	**45.3**	78.7	**92.6**
PaLM	8B	4.9	4.1	15.1	**16.8**	23.7	**25.2**	19.3	**21.7**	26.2	**30.5**
	62B	9.6	**29.9**	48.2	46.7	58.7	**61.9**	25.6	22.4	61.8	**80.3**
	540B	17.9	**56.9**	69.4	**79.0**	72.1	**73.9**	25.2	**35.8**	79.2	**93.3**

Google Developer Group
Incheon National University

3. Arithmetic Reasoning

- Ablation study

- (목적) Prompting에 의한 성능 향상이 CoT에 의한 것인지, 아니면 다른 요인에 의한 것인지 확인

- (방법) Equtaion only / Variable only / CoT 별 Prompting 진행

- (결과) 다른 요인이 CoT에 의한 성능 향상에 영향을 미치지 않음

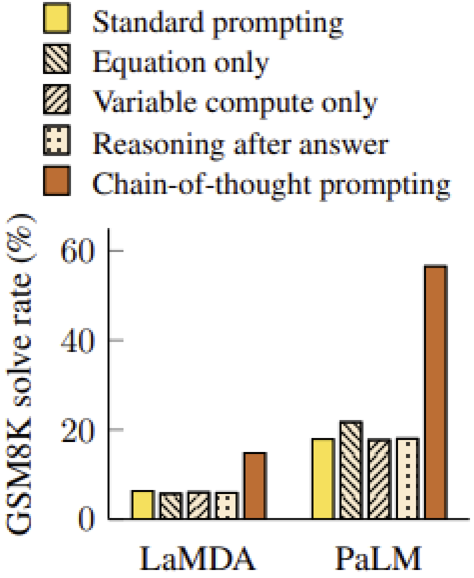


Figure 5: Ablation study for different variations of prompting using LaMDA 137B and PaLM 540B. Results for other datasets are given in Appendix Table 6 and Table 7.

3. Arithmetic Reasoning

- Robustness of Chain of Thought

- (목적) Prompting이 다양한 조건에서도 안정적으로 작동하는지 검증

- (방법) Different of annotators, exemplars, order of examplars, language models 별 성능 비교

- (결과) 다양한 조건 속에서도 안정적인 성능 향상을 보여주고 있음

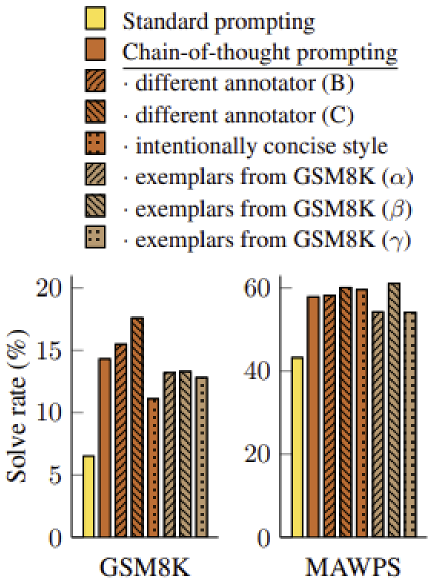
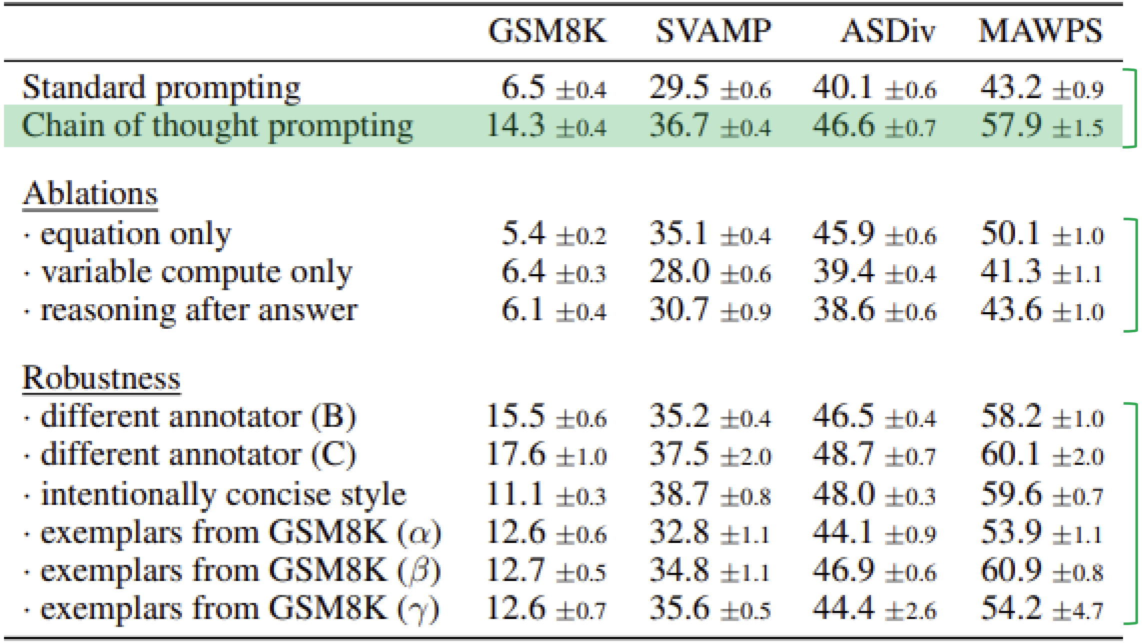


Figure 6: Chain-of-thought prompting has variance for different prompt examples (as expected) but outperforms standard prompting for various annotators as well as for different exemplars.

3. Arithmetic Reasoning

- Ablation study & Robustness of Chain of Thought



	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

CoT Prompting 성능 향상

CoT 추론에 의한 성능 향상

특정 조건에 지나치게 의존적이지 않음

4. Commonsense Reasoning

- Examples

Q. What do people use th absorb extra ink from a fountain pen?
(a) shirt poket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

A. The answer must be an item that can absorb ink. Of the above choices, only blotters are used to absorb ink. So the answer is (e)

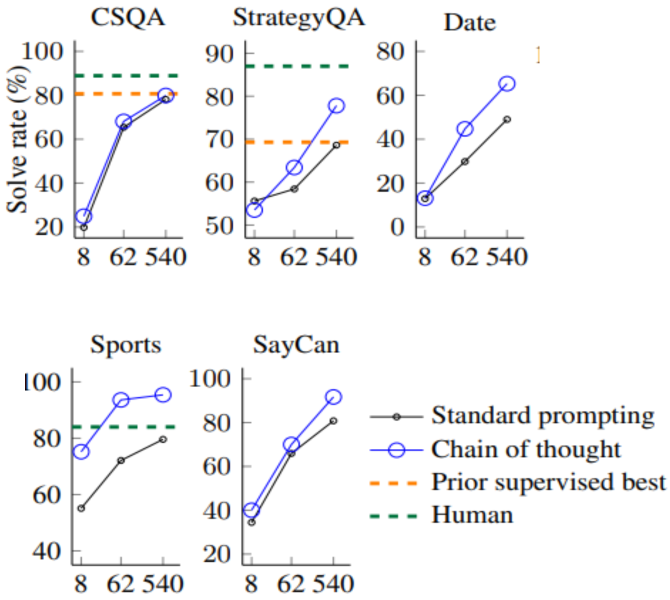
<CSQA>

Q. 2015 is coming in 36 hours. What is the data one week from today in MM/DD/YYYY?

A. If 2015 is coming in 36 hours, then it is coming in 2 days. 2 days before 01/01/2015 is 12/30/2014, so today is 12/30/2014. So one week from today will be 01/05/2015. So the answer is 01/05/2015.

<Date understanding>

4. Commonsense Reasoning



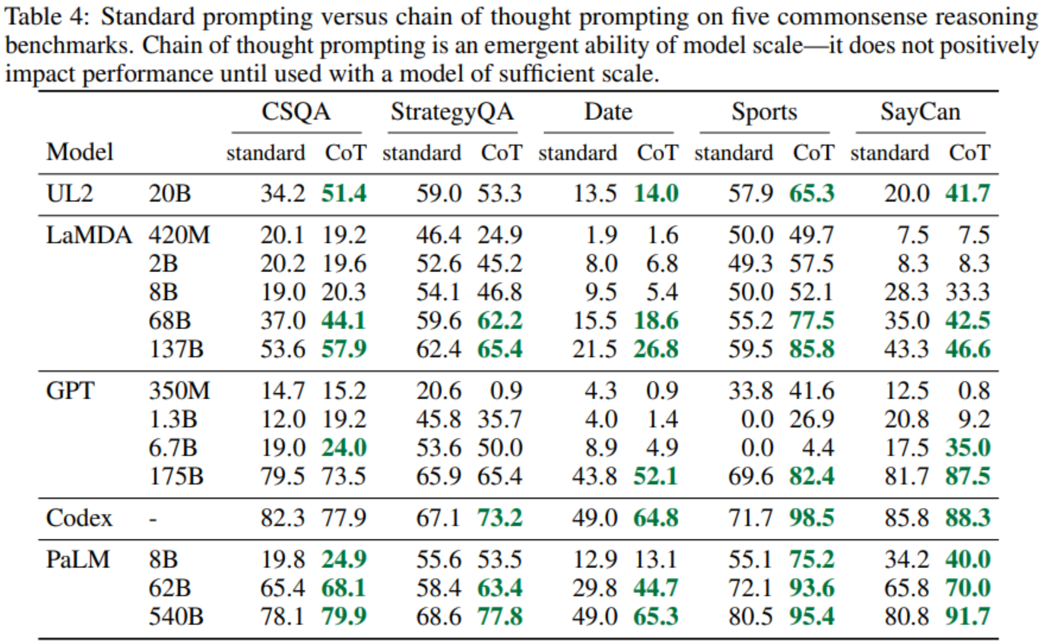


Table 4: Standard prompting versus chain of thought prompting on five commonsense reasoning benchmarks. Chain of thought prompting is an emergent ability of model scale—it does not positively impact performance until used with a model of sufficient scale.

		CSQA		StrategyQA		Date		Sports		SayCan	
Model		standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	34.2	**51.4**	59.0	53.3	13.5	**14.0**	57.9	**65.3**	20.0	**41.7**
LaMDA	420M	20.1	19.2	46.4	24.9	1.9	1.6	50.0	49.7	7.5	7.5
	2B	20.2	19.6	52.6	45.2	8.0	6.8	49.3	57.5	8.3	8.3
	8B	19.0	20.3	54.1	46.8	9.5	5.4	50.0	52.1	28.3	33.3
	68B	37.0	**44.1**	59.6	**62.2**	15.5	**18.6**	55.2	**77.5**	35.0	**42.5**
	137B	53.6	**57.9**	62.4	**65.4**	21.5	**26.8**	59.5	**85.8**	43.3	**46.6**
GPT	350M	14.7	15.2	20.6	0.9	4.3	0.9	33.8	41.6	12.5	0.8
	1.3B	12.0	19.2	45.8	35.7	4.0	1.4	0.0	26.9	20.8	9.2
	6.7B	19.0	**24.0**	53.6	50.0	8.9	4.9	0.0	4.4	17.5	**35.0**
	175B	79.5	73.5	65.9	65.4	43.8	**52.1**	69.6	**82.4**	81.7	**87.5**
Codex	-	82.3	77.9	67.1	**73.2**	49.0	**64.8**	71.7	**98.5**	85.8	**88.3**
PaLM	8B	19.8	**24.9**	55.6	53.5	12.9	13.1	55.1	**75.2**	34.2	**40.0**
	62B	65.4	**68.1**	58.4	**63.4**	29.8	**44.7**	72.1	**93.6**	65.8	**70.0**
	540B	78.1	**79.9**	68.6	**77.8**	49.0	**65.3**	80.5	**95.4**	80.8	**91.7**

Google Developer Group
Incheon National University

5. Symbolic Reasoning

- Examples

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.

<LAST LETTER CONCATENATION >

Q. A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up?

<COIN FLIP>

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.

5. Symbolic Reasoning

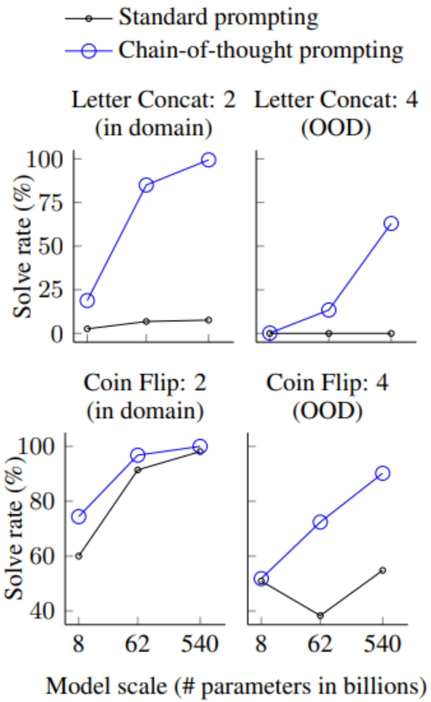
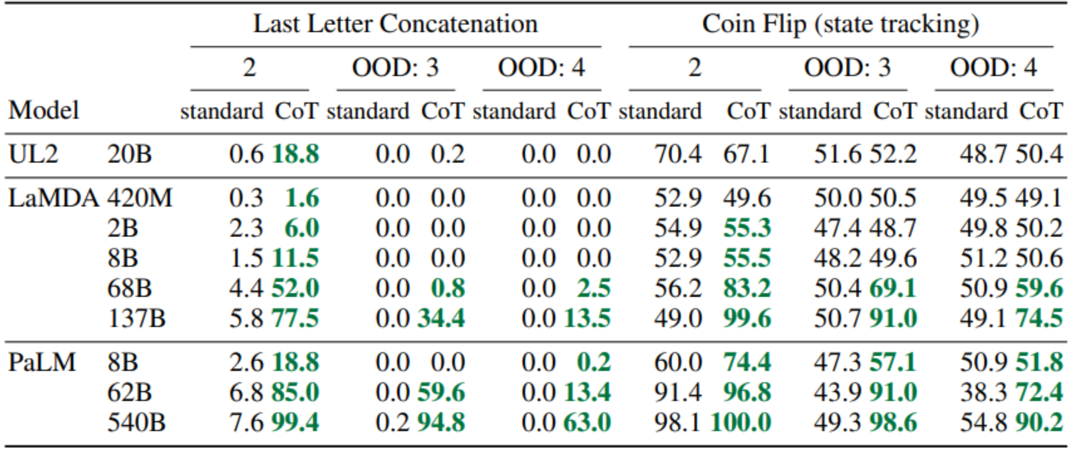


Table 5: Standard prompting versus chain of thought prompting enables length generalization to longer inference examples on two symbolic manipulation tasks.



		Last Letter Concatenation						Coin Flip (state tracking)					
		2		OOD: 3		OOD: 4		2		OOD: 3		OOD: 4	
Model		standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	0.6	**18.8**	0.0	0.2	0.0	0.0	70.4	67.1	51.6	52.2	48.7	50.4
LaMDA	420M	0.3	**1.6**	0.0	0.0	0.0	0.0	52.9	49.6	50.0	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**

6. Conclusions

특징

- **Prompting은 언어 모델의 추론 능력을 향상시키는 간단하고 광범위하게 적용 가능한 방법**
  - Prompting을 통해 다단계 추론 동작을 이끌어 낼 수 있음.
  - Prompting은 모델의 규모가 커짐에 따라 효과적으로 적용

한계

- 다만, 실제 신경망이 추론하는지 알 수 없고, 추론의 정확성을 보장하지 않아 오류가 발생할 수 있음

Q&A