간단 논세

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

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

2024.10.11 유하영

previous research

언어 모델의 규모를 확장 -> 성능 향상 -> 모든 task에서 항상 좋은 성능 X
(산술 /상식 /기호적 추론 task)

1. natural language rationales(이론적 근거) 함께 생성

고품질의 rationales를 생성하는 것은 비용이 많이 든다

2. prompt를 통한 In-context few-shot learning

사용된 전통적인 few-shot prompting 방법은 추론 능력 작업에서 성능이 낮다

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

Table 3.1: Performance on cloze and completion tasks. GPT-3 significantly improves SOTA on LAMBADA while achieving respectable performance on two difficult completion prediction datasets.

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

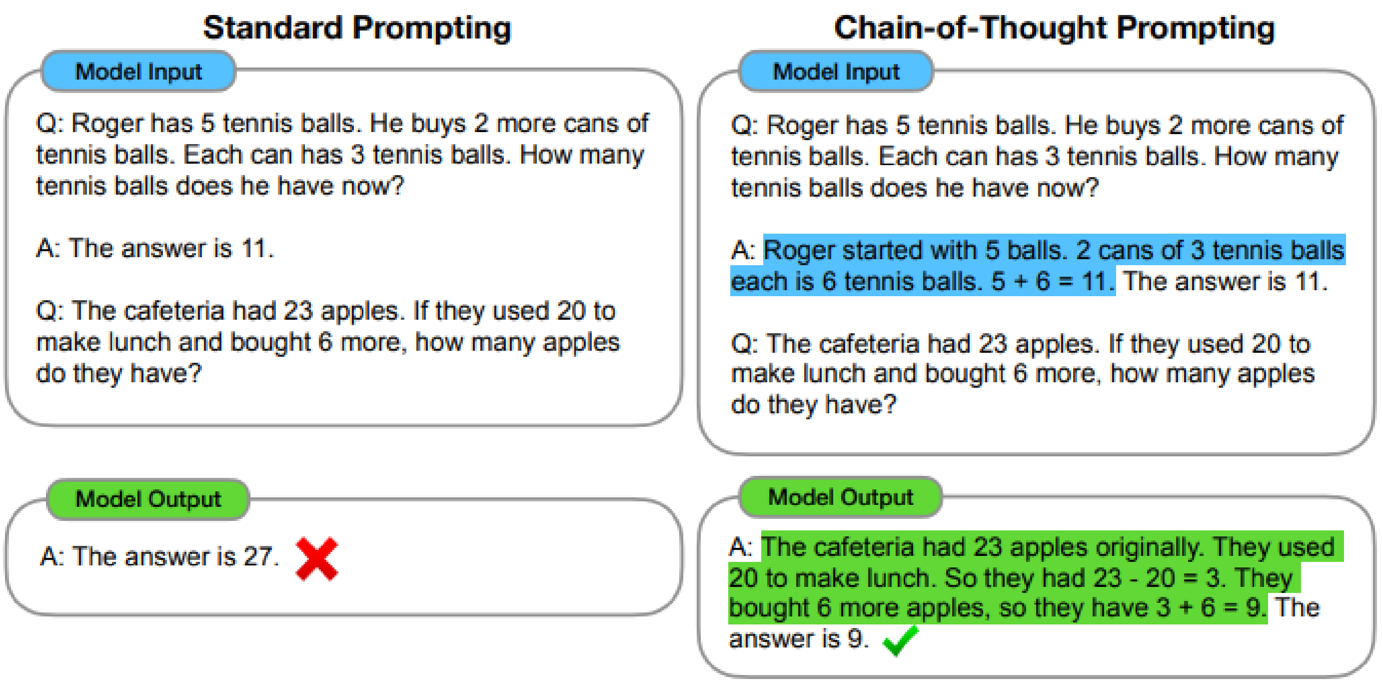
	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Table 3.5: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

Language Models are Few-Shot Learners. 2020

Chain-Of-Thought Prompting

LLM이 복잡한 문제를 해결할 때, **중간 추론 단계를 스스로 생성하여 최종 답을 도출**



1. 문제를 푸는 과정을 단계별로 분해해 추론
   -> 각 단계에서 추가적인 연산 할당
(multi-step problems into intermediate steps)

2. 틀린답 생성시 어느부분에서 잘못 추론했는지 debug가 가능
-> 중간 단계의 추론 과정이 모두 명시되기 때문에, 모델이 잘못된 답을 내놓더라도 **어느 단계에서 오류가 발생했는지**를 쉽게 파악할 수 있다.

3. CoT는 대부분의 task와 LM에 적용 가능

Experimental Setup

Math Word Problems (free response)

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.

Math Word Problems (multiple choice)

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).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

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

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

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

수동으로 8개의 few-shot 예시를 작성

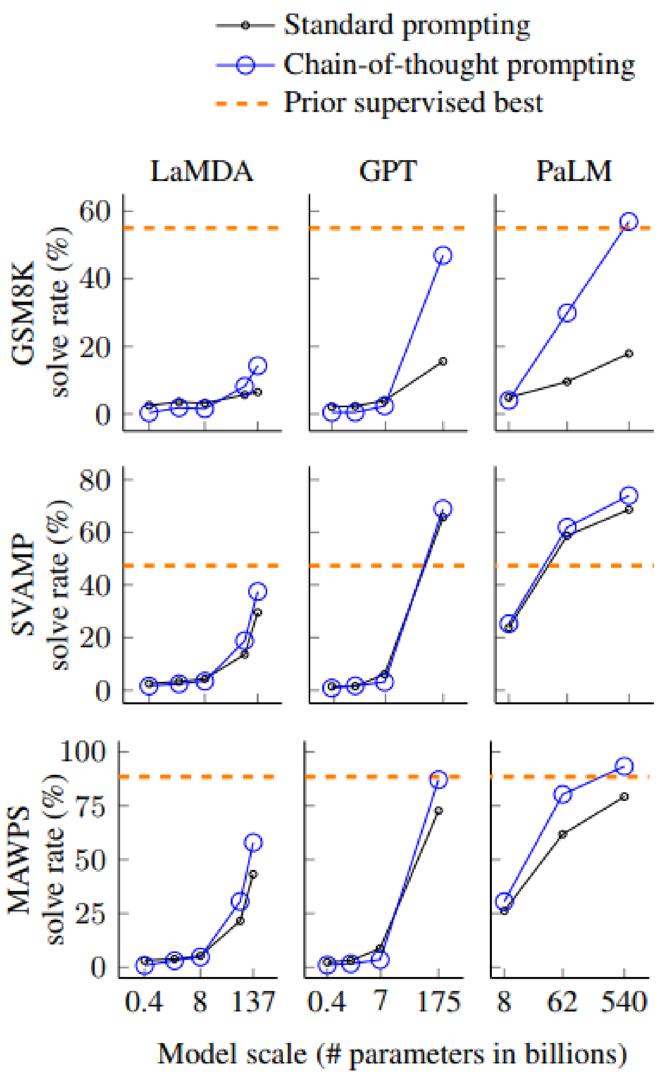
Figure 3: Examples of ⟨input, chain of thought, output⟩ triples for arithmetic, commonsense, and symbolic reasoning benchmarks. Chains of thought are highlighted. Full prompts in Appendix G.

Results

수학문제

문맥과 구조를 파악해야하는 수학문제

언어이해 및 수학추론능력

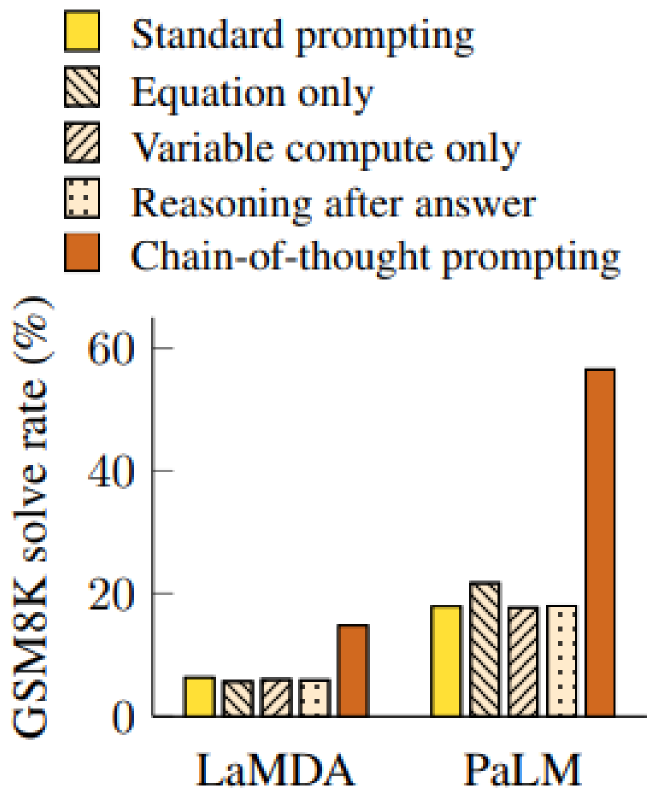


모델의 스케일이 커질수록 CoT 성능이 높아짐

CoT는 복잡한 문제에 적용했을 때, 성능이 더 높았다
쉬운 문제일땐 Standard prompting에 비해 성능이 감소

PaLM 540B -> SOTA

Ablation Study

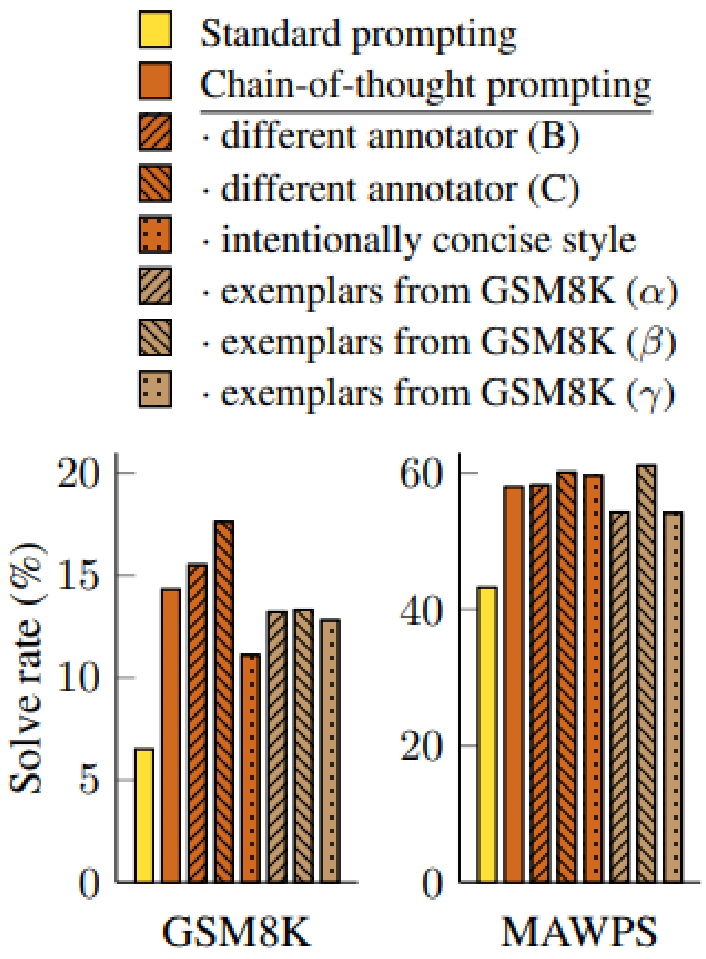


1. Equation only
- CoT를 자연어 대신 수식만 제공
- 복잡한 task에서는 성능이 좋지 않았지만, 조금 더 쉬운 task에서는 좋은 성능을 보임

2. Variable compute only
- CoT가 성능이 잘 나오는 이유가 Standard prompting 보다 토큰 수가 늘어나 연산량이 증가한 것이 원인이 아닐까 생각
- Standard prompting에서도 '…'을 넣어 CoT 때와 토큰 수를 똑같이 맞춰줌
- 여전히 CoT가 성능이 좋았다.

3. Chain of thought after answer
- CoT prompt를 answer 이후에 넣고 추론
- 기존의 standard prompting과 성능이 유사
  -> 추론 과정 중간에 CoT prompting을 사용이 도움이 됨을 증명

Robustness of Chain of Though



- **다양한 작성자에 의한 CoT 사용**

작성자의 스타일이나 문제에 따라 약간의 성능 차이, 기존 방식에 비해 향상된 성능

- **예시 순서의 민감성**

예시 순서를 바꾸더라도 강력한 성능 유지

Conclusions

CoT를 이용한 prompting은 ‘산술, 상식, 기호적 추론’ task에서 좋은 성능을 낸다.

chain-of-thought prompting does not positively impact performance for small models, and only yields performance gains when used with models of **~100B parameters.**

cf) Auto-CoT

Auto-CoT

1.Manual-CoT
= 본 논문의 CoT / 높은 성능

2. Zero-shot-CoT
"Let's think step by step"을 프롬프트에 추가하여
모델이 자동으로 중간 추론과정을 생성.
/ 성능이 떨어질 수 있다

=> Auto-CoT 제안
1.자동화된 프롬프트 생성(중간 추론 단계)
2.다양한 추론 경로 생성
그중 가장 신뢰성 있는 답변 선택 즉, 모델이 각기 다른 방식으로 문제를 풀어보게 하고, 결과 중에서 가장 신뢰성이 높은 답을 선택.
→ 수학적 추론과 상징적 추론 에서 기존 CoT보다 더 나은 성능을 보임

Table 1: Accuracy (%) of different sampling methods. Symbol † indicates using training sets with annotated reasoning chains.

Method	MultiArith	GSM8K	AQuA
Zero-Shot-CoT	78.7	40.7	33.5
Manual-CoT	**91.7**	46.9	35.8†
Random-Q-CoT	86.2	47.6†	36.2†
Retrieval-Q-CoT	82.8	**48.0†**	**39.7†**

수학(산술)/수학(일상적인 시나리오)/대수학 단어(알고리즘적)

➔ 복잡한 데이터셋에서 높은 성능을 보임

https://arxiv.org/abs/2210.03493

cf)

- The COT COLLECTION: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning (2023)

  CoT Fine-tuning 도입

- Zero-Shot Chain-of-Thought Reasoning Guided by Evolutionary Algorithms in Large Language Models (2024)

  Zero-shot-CoT에서 동일한 CoT 프롬프트를 모든 작업에 적용하는 대신,
  다양한 프롬프트를 동적으로 생성하여 성능을 개선