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

Сергей Лоптев

Доклад

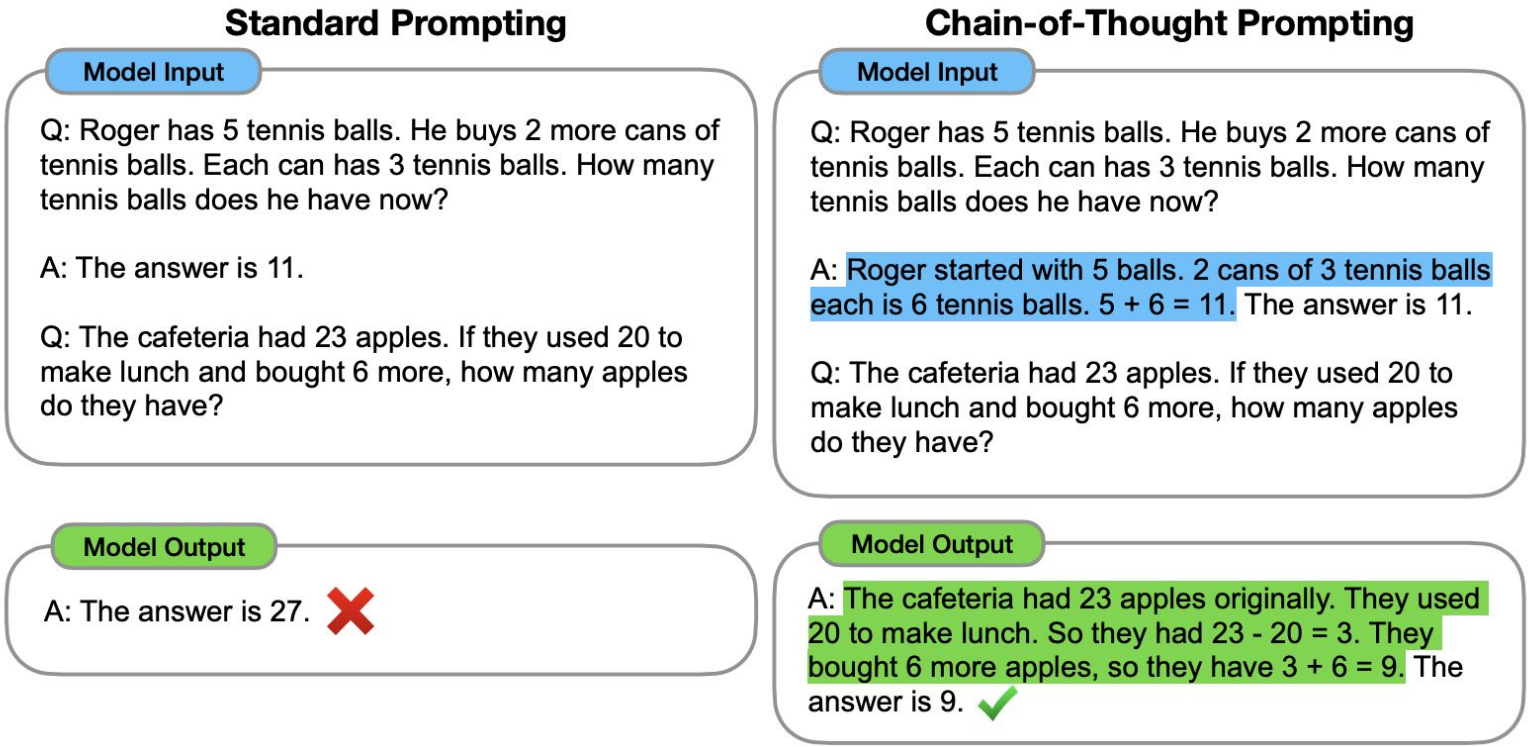
Общая идея

- Хотим улучшить мыслительные способности больших языковых моделей в few-shot постановке
- Хотим этого добиться без обучения чего-либо, а лишь используя более продвинутые подсказки (prompt)
- Давайте попробуем изменять не входы модели, а выходы: добавим к ответам некоторую последовательность мыслей, приводящую к данному ответу (chain of thought)

Chain-of-Thought

- Представьте, что вы в третьем классе решаете задачу по математике в несколько действий
- Вероятно, ваш мыслительный процесс будет последовательным, типа:
- “After Jane gives 2 flowers to her mom she has 10 ... then after she gives 3 to her dad she will have 7 ... so the answer is 7.” – Chain of Thought
- В данной работе авторы пробуют использовать идею этих мыслительных процессов в языковых моделях, и показывают что она успешно работает для достаточно больших языковых моделей.

Пример


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.

Почему это хорошая идея

- Chain-of-Thought позволяет моделям разлагать многошаговые задачи по шагам, то есть выделять дополнительные вычисления для задач, требующих большего количества рассуждений.
- Chain-of-Thought добавляет моделям интерпретируемости, помогая понять, как модель дошла до ответа (в том числе, это немного упрощает отладку).
- Рассуждения с помощью Chain-of-Thought могут быть использованы для математических задач, задач с применением здравого смысла, и задач, требующих манипуляцию символами; потенциально это применимо ко всем языковым задачам.
- Chain-of-Thought могут быть использованы для любых достаточно больших языковых моделей во few-shot постановке, без дополнительного обучения.

Эксперименты

Авторы проводят эксперименты на трех типах языковых задач:

- Математические задачи (arithmetic reasoning, math word problems)
- Задачи, связанные с применением здравого смысла (commonsense reasoning)
- Задачи, связанные с символьными манипуляциями (symbolic reasoning)

Разберём эти эксперименты.

Математические задачи: выборки

- **GSM8K (OpenAI)** – 8.5K лингвистически разнообразных математических задач для начальной школы, написанных людьми.
- **SVAMP (Microsoft)** – 1K математических задач – одна задача может быть представлена в нескольких вариациях (5+3=8 или 8-3=5)
- **ASDiv** – 2.3K математических задач, разнообразных как лингвистически, так и по типам задач
- **AQuA (DeepMind)** – 254 алгебраических задач с мультивыбором (посложнее чем начальная школа)
- **MAWPS** – 3.3K задач, состоит из нескольких выборок с разными типами задач

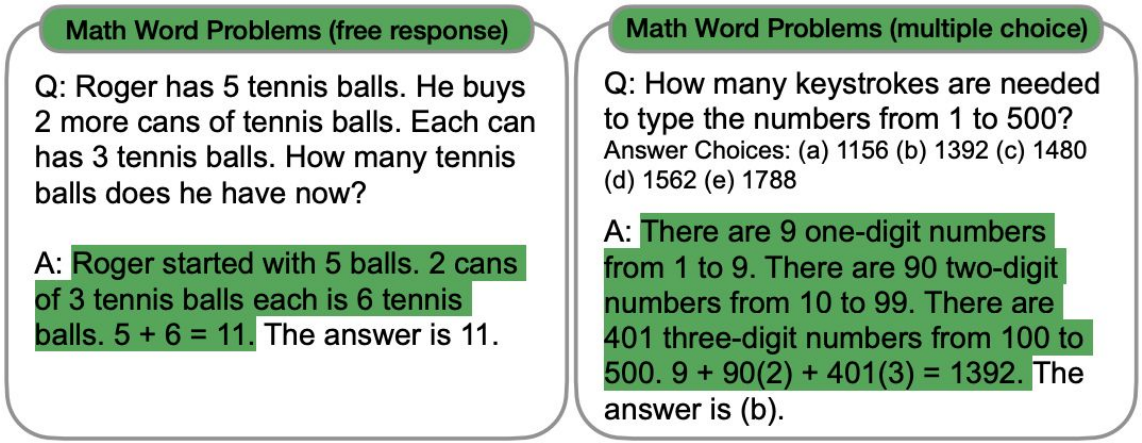
Математические задачи: примеры

Table 12: Summary of math word problem benchmarks we use in this paper with examples. N: number of evaluation examples.

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?

Математические задачи: подсказки

- Базовые подсказки (baseline prompts): просто берём из выборок как было
- Наш эксперимент: добавляем перед ответом chain-of-thought



Математические задачи: модели

- Модели:
  - **GPT-3 (OpenAI)**
  - **LaMDA (Google)**, усредняем результаты по 5 случайным сидам
  - **PaLM (Google)** – максимально 540B параметров, самая большая модель из списка
  - **UL2 20B (Google)**
  - **Codex (OpenAI)**
- Сэмплирование происходит жадно (greedy decoding)

Математические задачи: результаты

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**

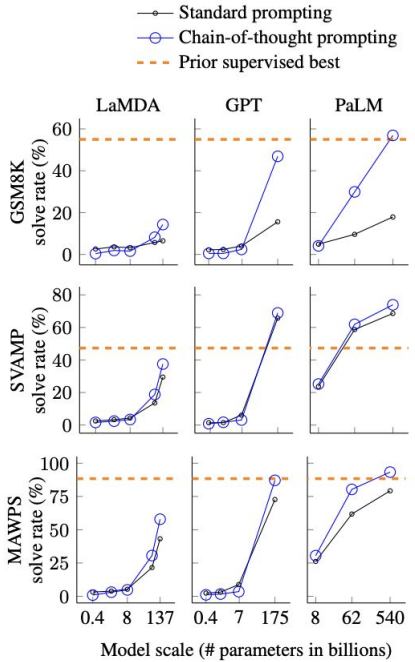


Figure 4: Chain-of-thought prompting enables large language models to solve challenging math problems. Notably, chain-of-thought reasoning is an emergent ability of increasing model scale. Prior best numbers are from Cobbe et al. (2021) for GSM8K, Jie et al. (2022) for SVAMP, and Lan et al. (2021) for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

Математические задачи: результаты

Table 1: Chain of thought prompting outperforms standard prompting for various large language models on five arithmetic reasoning benchmarks. All metrics are accuracy (%). Ext. calc.: post-hoc external calculator for arithmetic computations only. Prior best numbers are from the following. *a*: Cobbe et al. (2021). *b* & *e*: Pi et al. (2022), *c*: Lan et al. (2021), *d*: Piękos et al. (2021).

	Prompting	GSM8K	SVAMP	ASDiv	AQuA	MAWPS
Prior best	N/A (finetuning)	55[a]	57.4[b]	75.3[c]	37.9[d]	88.4[e]
UL2 20B	Standard	4.1	10.1	16.0	20.5	16.6
	Chain of thought	4.4 (+0.3)	12.5 (+2.4)	16.9 (+0.9)	23.6 (+3.1)	19.1 (+2.5)
	+ ext. calc	6.9	28.3	34.3	23.6	42.7
LaMDA 137B	Standard	6.5	29.5	40.1	25.5	43.2
	Chain of thought	14.3 (+7.8)	37.5 (+8.0)	46.6 (+6.5)	20.6 (-4.9)	57.9 (+14.7)
	+ ext. calc	17.8	42.1	53.4	20.6	69.3
GPT-3 175B	Standard	15.6	65.7	70.3	24.8	72.7
(text-davinci-002)	Chain of thought	46.9 (+31.3)	68.9 (+3.2)	71.3 (+1.0)	35.8 (+11.0)	87.1 (+14.4)
	+ ext. calc	49.6	70.3	71.1	35.8	87.5
Codex	Standard	19.7	69.9	74.0	29.5	78.7
(code-davinci-002)	Chain of thought	63.1 (+43.4)	76.4 (+6.5)	80.4 (+6.4)	45.3 (+15.8)	92.6 (+13.9)
	+ ext. calc	65.4	77.0	80.0	45.3	93.3
PaLM 540B	Standard	17.9	69.4	72.1	25.2	79.2
	Chain of thought	56.9 (+39.0)	79.0 (+9.6)	73.9 (+1.8)	35.8 (+10.6)	93.3 (+14.2)
	+ ext. calc	58.6	79.8	72.6	35.8	93.5

Задачи про здравый смысл: выборки

- **CSQA** – 200K общих вопросов, требующих знания о мире
- **StrategyQA** – 2.8K вопросов, требующих от моделей несколько последовательных рассуждений для ответа
- Две выборки из **BIG-Bench (Google)**:
  - **Date Understanding** – требует от моделей извлечения дат из текста
  - **Sports Understanding** – требует от моделей определить, вероятно ли некоторое предложение, связанное со спортом
- **SayCan (Google)** – отображает инструкции, данные естественным языком, в инструкции для робота

Задачи про здравый смысл: примеры

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?

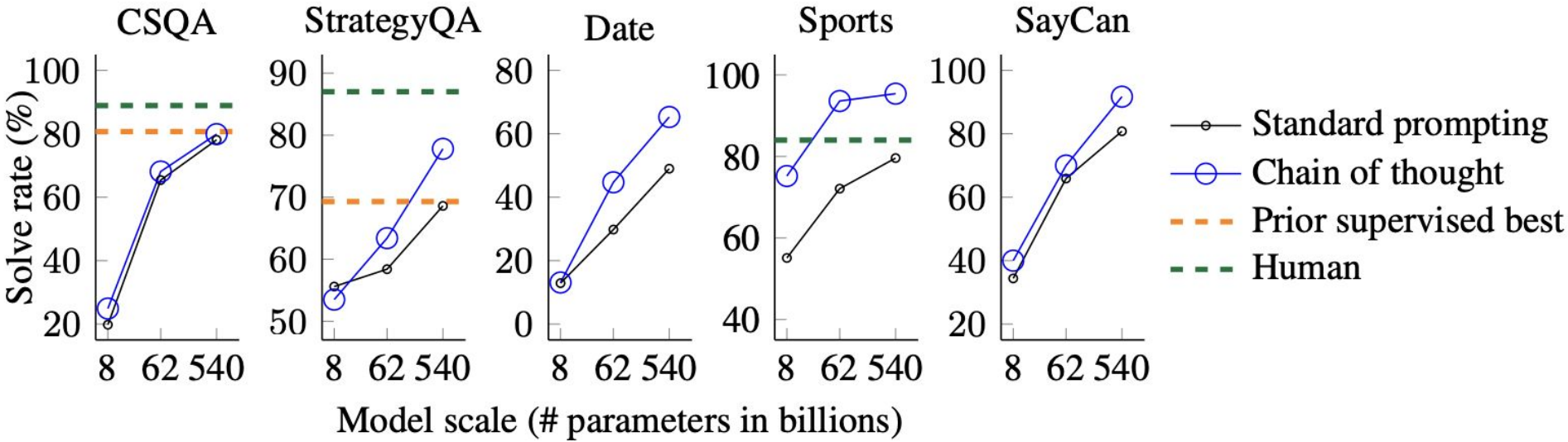
Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Задачи про здравый смысл: результаты

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**

Задачи про здравый смысл: PaLM vs prior best



Символьные задачи: выборки

Рассматриваются два типа задач:

- Last Letter Concatenation: требует от модели сконкатенировать последние буквы в словах, образующих имя. Имена взяты с namecensus.com.
- Coin Flip: просит модель сказать, какой стороной вверх лежит монета после того, как несколько людей переворачивают или не переворачивают её.

Для обеих задач также тестируется работоспособность вне домена, то есть с увеличенной длиной входа (length generalization).

Символьные задачи: примеры

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.

Символьные задачи: результаты

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**

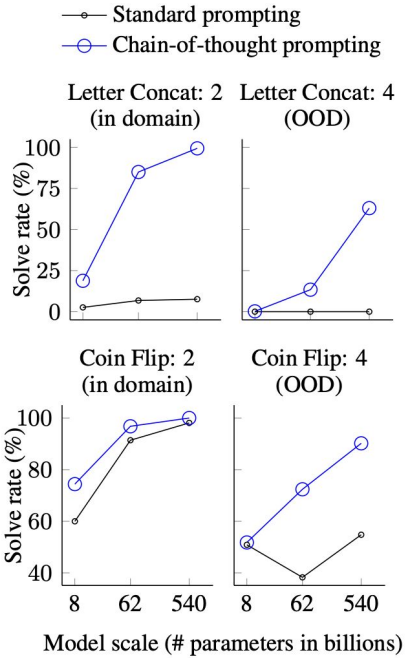


Figure 8: Using chain-of-thought prompting facilitates generalization to longer sequences in two symbolic reasoning tasks.

Ablation study на CoT

- Авторы провели эксперимент со вставкой другой вспомогательной информации в выводы, например:
  - Только математическое уравнение, приводящее к результату, например:
    5 + 2 * 3 = 11. The answer is 11.
  - Только информация о сложности задачи, то есть количество точек, равное длине уравнения, например:
    …….. The answer is 11.
  - Возможно, CoT только помогает достать нужное знание из модели, а не генерировать мысли. Попробуем положить CoT после ответа, например:
    The answer is 11. Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11.

Ablation study: результаты

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

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

Ablation study: результаты

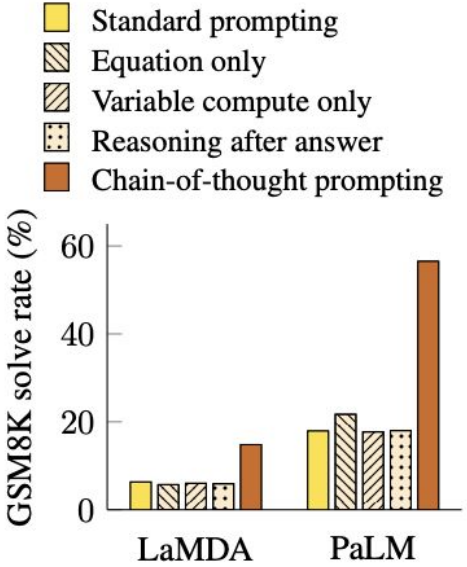


Figure 5: Ablation study for different variations of prompting using LaMDA 137B and PaLM 540B. Results for other datasets are given in Appendix Table 6 and Table 7.

Устойчивость к разным аннотациям

- Также авторы провели эксперимент на робастность модели к разным аннотаторам и примерам, были взяты следующие экспериментальные выборки:
  - Два дополнительных аннотатора, помимо основного
  - Основным аннотатором дополнительно написаны CoT в более лаконичном стиле
  - Дополнительно взяли по 3 случайных набора по 8 примеров из обучающей выборки GSM8K, размеченные краудсорсингом
- Также посмотрели на производительность в зависимости от количества примеров (shots)

Робастность CoT: результаты

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

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

Робастность CoT: результаты

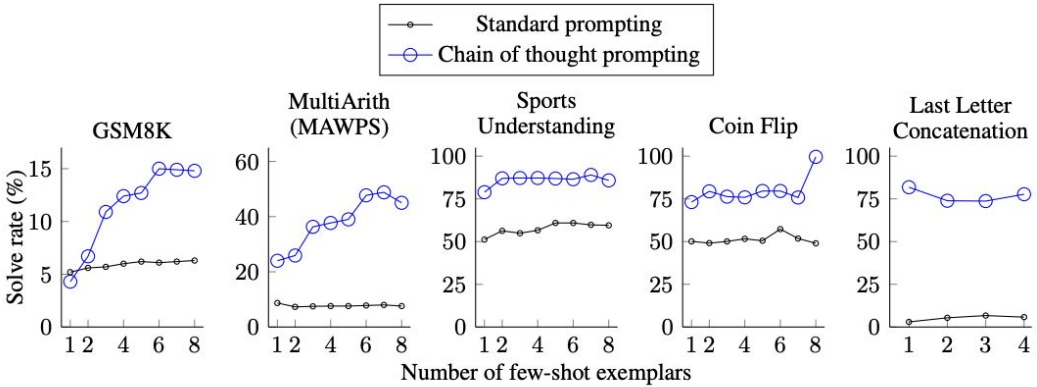


Figure 11: The improvement of chain of thought prompting over standard prompting appears robust to varying the number of few-shot exemplars in the prompt.

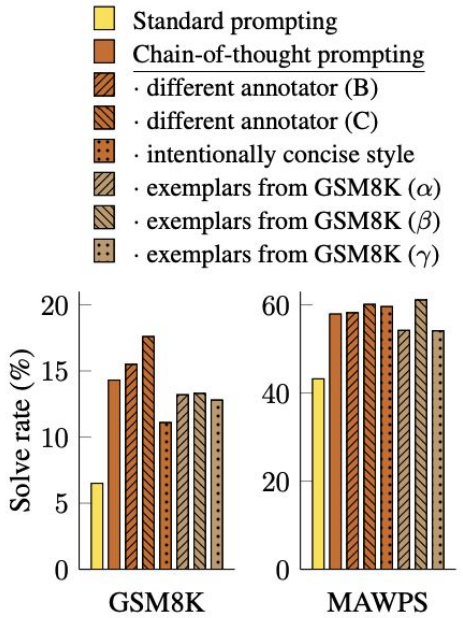


Figure 6: Chain-of-thought prompting has variance for different prompt examples (as expected) but outperforms standard prompting for various annotators as well as for different exemplars.

Почему хорошо работает с большими моделями

- Авторы анализируют, почему CoT так хорошо работает только с большими (50B+) моделями
- Классифицировали 45 ошибок, сделанных PaLM 62B, и посмотрели, сколько исправляет PaLM 540B:

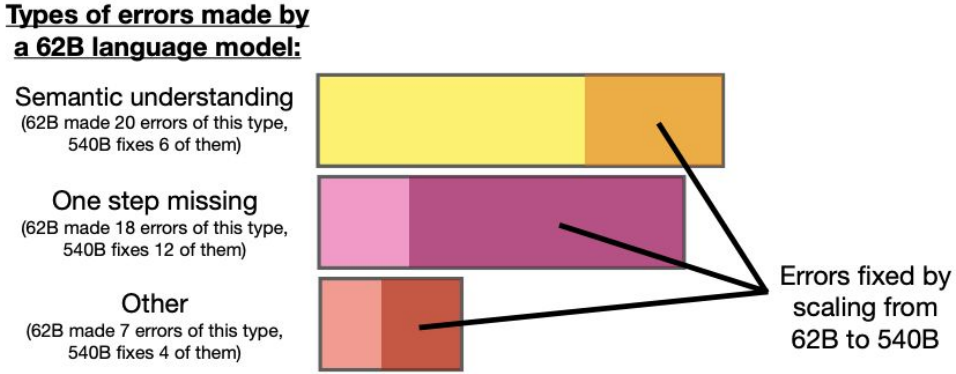


Figure 9: Error analysis of 45 problems that PaLM 62B got incorrect. These errors were categorized that semantic understanding, one step missing, and other. The other category includes hallucinations, repetitive outputs, and symbol mapping errors. Scaling PaLM to 540B fixed a substantial portion of errors in all categories.

Почему плохо работает с маленькими моделями

- Маленькие модели плохо справляются даже с простыми задачами отображения (e.g. symbolic reasoning)
- Маленьким моделям присущи слабые арифметические способности, они сильно зависят от размера модели
- Маленькие модели периодически не генерируют финальный ответ, или из-за повторений, или из-за того, что логика не дошла до него

Выводы

- Техника chain-of-thought заключается в добавлении последовательностей мыслей к выходам модели в примерах во few-shot постановке
- Эта техника помогает моделям более точно отвечать на исследуемые задачи посредством уточнения "мыслительного процесса" модели
- В том числе, она бьёт SOTA результаты на математических и некоторых common sense задачах
- Основные ограничения – размер модели. Хорошо работает только для достаточно больших моделей