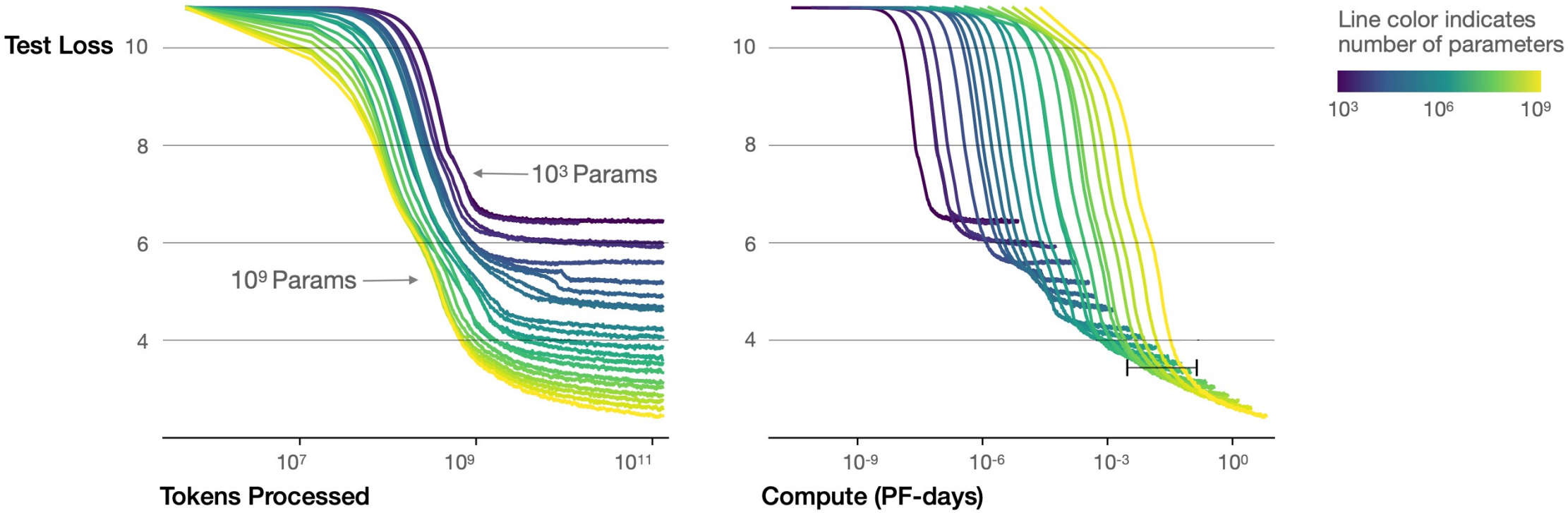
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Google Research, Brain Team

Presented by Michael Ungersböck (April 17th, 2024)

Scaling up LLMs

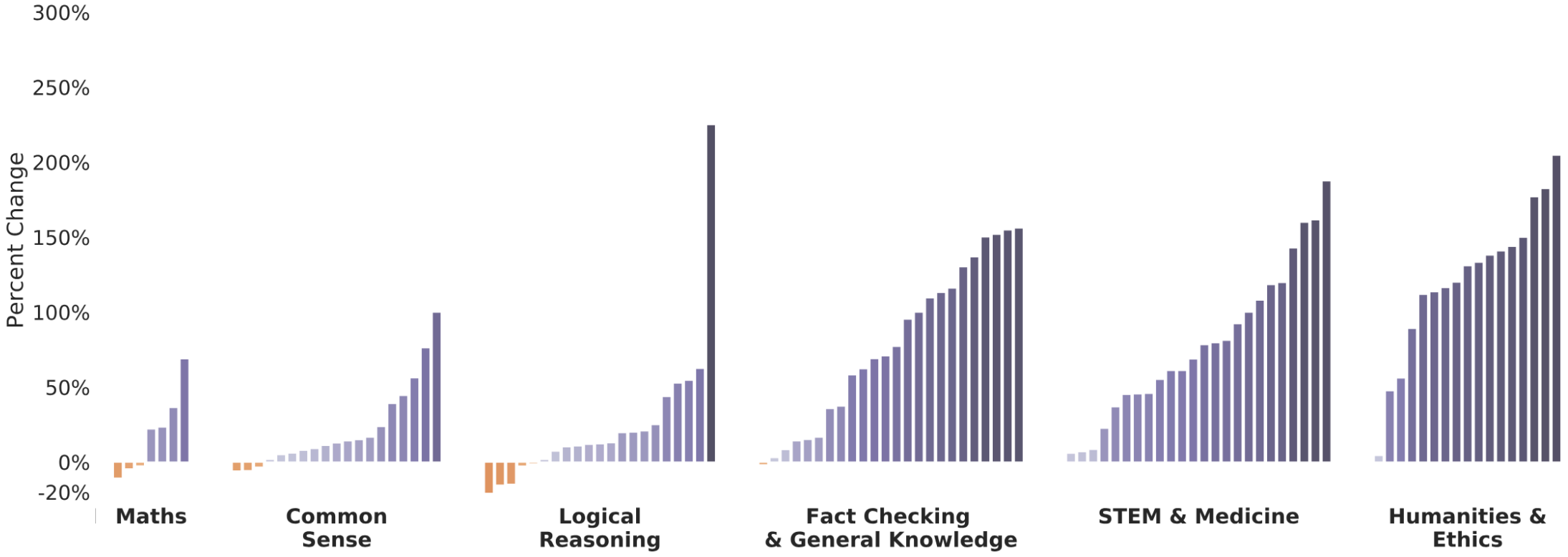
- PF-day: $10^{15} * 24 * 3600$ floating point operations



(Kaplan et al., 2020)

Scale and Performance

- smaller performance gains for challenging reasoning tasks



(Rae et al., 2021)

Idea 1: Generating intermediate steps

Question: The speed at which a man can row a boat in still water is 25 km/h. If he rows downstream, where the speed of current is 11 km/h, what time will he take to cover 80 meters?

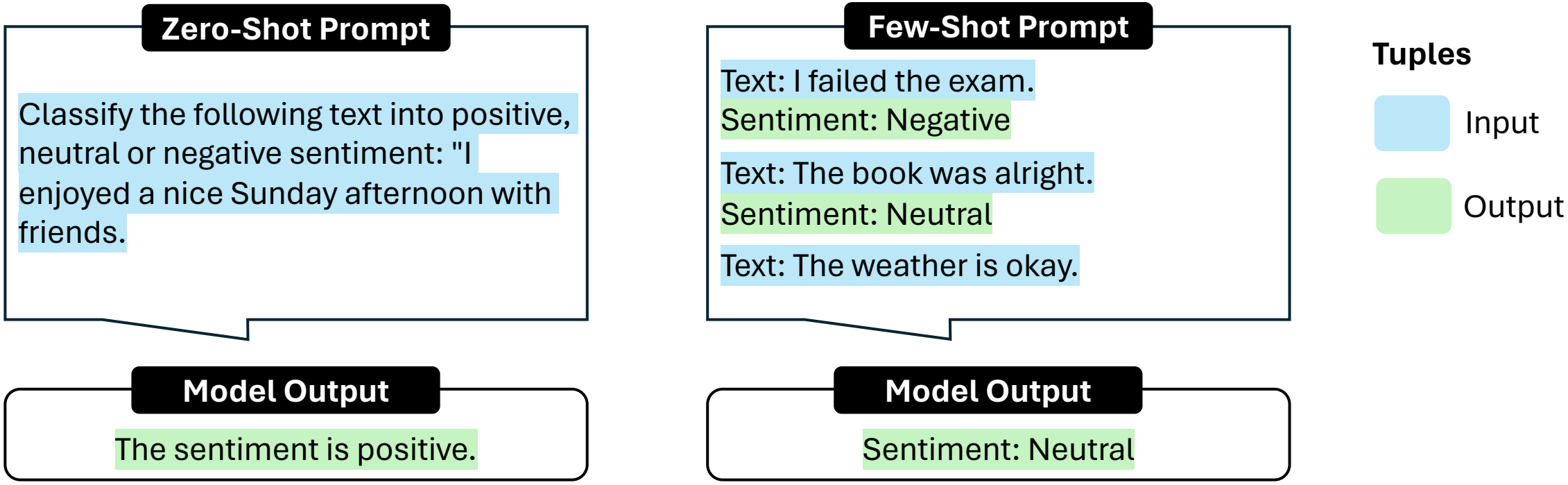
Options:
A) 18 s
B) 27 s
C) 26 s
D) 12 s
E) 8 s

Rationale: Total downstream speed: 25 +11 = 36 km/h = 36000 m/h. 36000 / (60*60) = 10 m/s. Hence time taken to cover 80 m = 80 / 10 = 8 seconds. → Answer: E

✓ improves arithmetic reasoning

✗ costly to create training sets with rationales

Idea 2: Few-Shot Prompting



- ✓ demonstrates different tasks without any finetuning
- ✗ performs poorly on reasoning tasks

Chain-of-Thought Example

Few-Shot Prompt

Roger has 5 tennis balls and buys 2 more cans. Each can has 3 tennis balls. How many tennis balls does he have now?

The answer is 11.

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

Model Output

The answer is 27.

Chain-of-Thought Prompt

Roger has 5 tennis balls and buys 2 more cans. Each can has 3 tennis balls. How many tennis balls does he have now?

Roger started with 5 tennis balls. 2 cans of 3 balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

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

Model Output

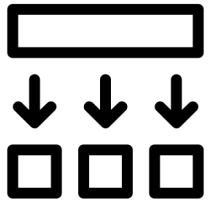
The cafeteria had 23 apples. They used 20 for lunch, so they had 23 – 20 = 3. They bought 6 apples more so 3 + 6 = 9. The answer is 9.

Triples

- Input
- CoT
- Output

Chain-of-Thought Prompting

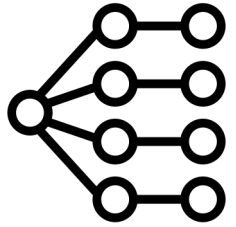
- new way to structure prompts: [Input, CoT, Output]
- series of natural language reasoning steps



decompose problems

universally applicable

evaluate behavior

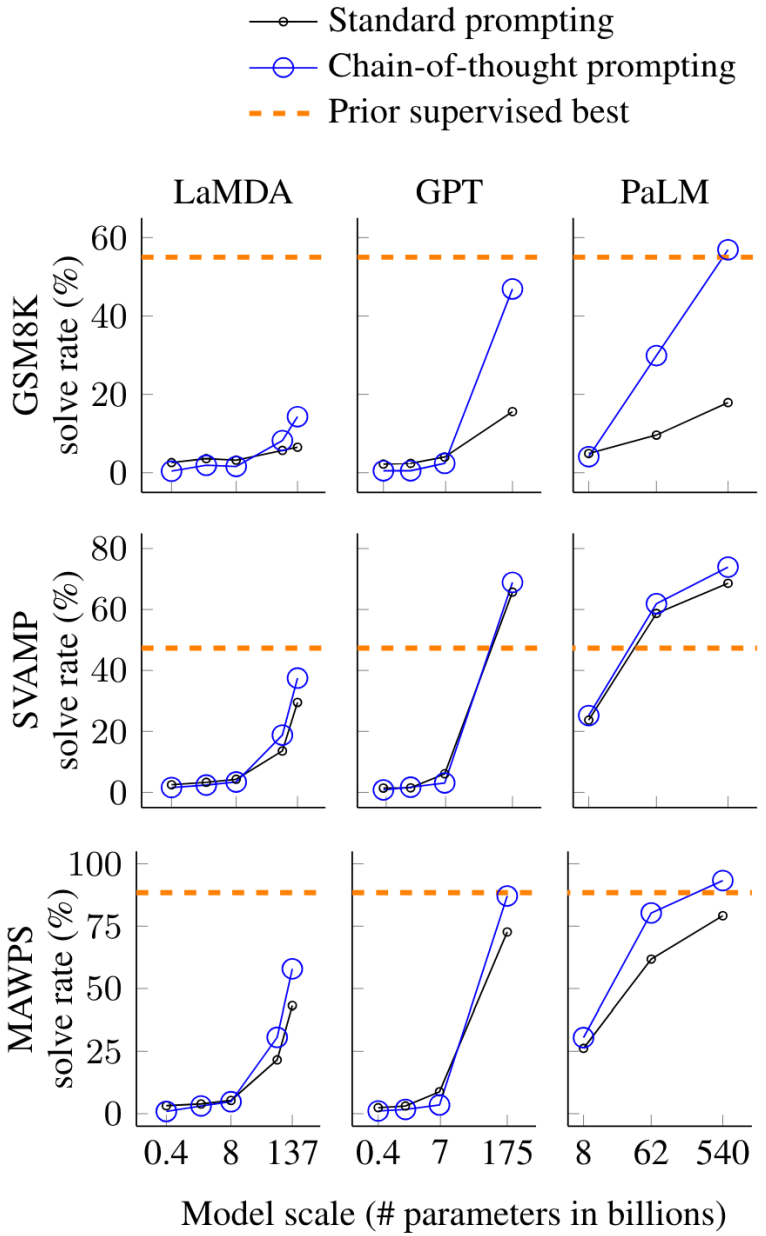


use existing LLMs

Arithmetic Reasoning

Setup & Results

- **benchmarks**
  - GSM8K, SVAMP, ASDiv, MAWPS, AQuA
- **prompting**
  - standard: 8-shot
  - chain-of-thought: manually annotated
- **LLMs**
  - GPT-3, LaMDA (5 seeds), PaLM
  - OpenAI Codex, UL2



(Wei et al., 2022)

Arithmetic Reasoning

Ablation Study

Equation only

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

23 - (5*3) = 8. The answer is $8.

Variable compute

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

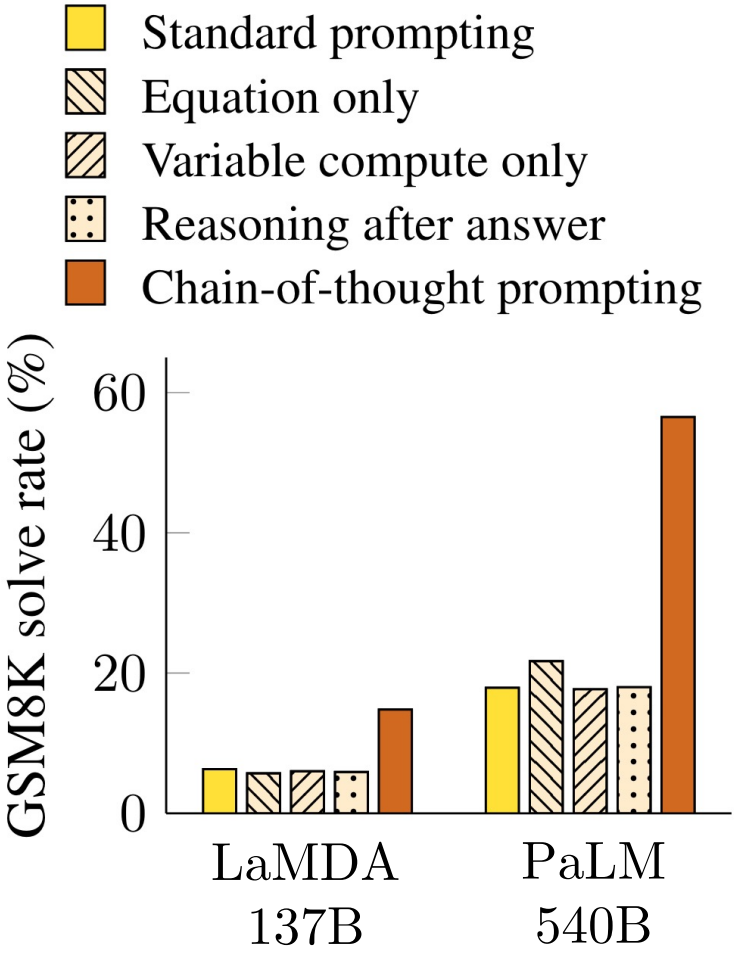
..... The answer is $8.

CoT after answer

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

The answer is $8.

Five bagels each costing $3 is total 5 * 3 = 15. She started with $23 so 23 - 15 = 8.

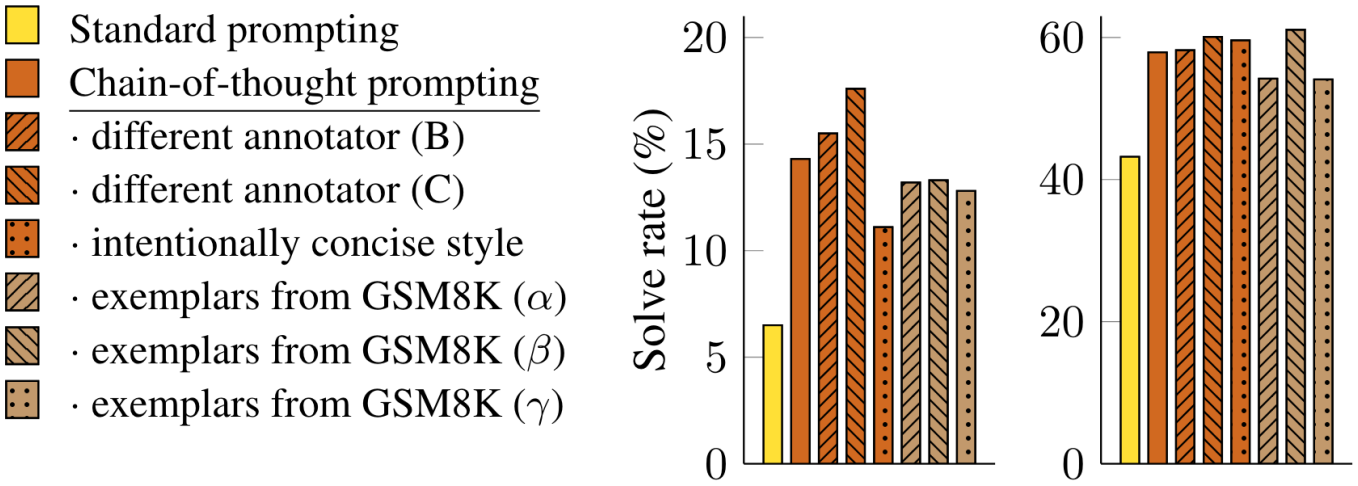


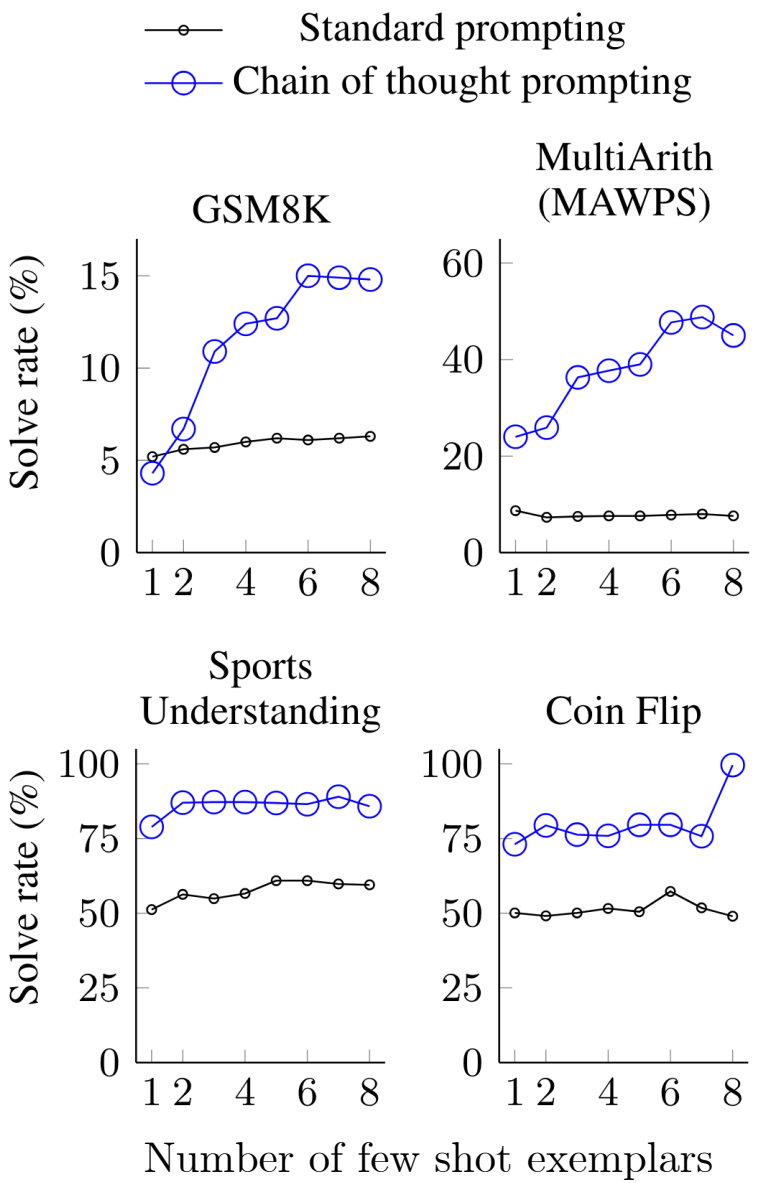
(Wei et al., 2022)

Arithmetic Reasoning

Robustness

- CoT performance on LaMDA 137B for
  - different annotators
  - different exemplars
  - varying number of exemplars





(Wei et al., 2022)

Commonsense Reasoning

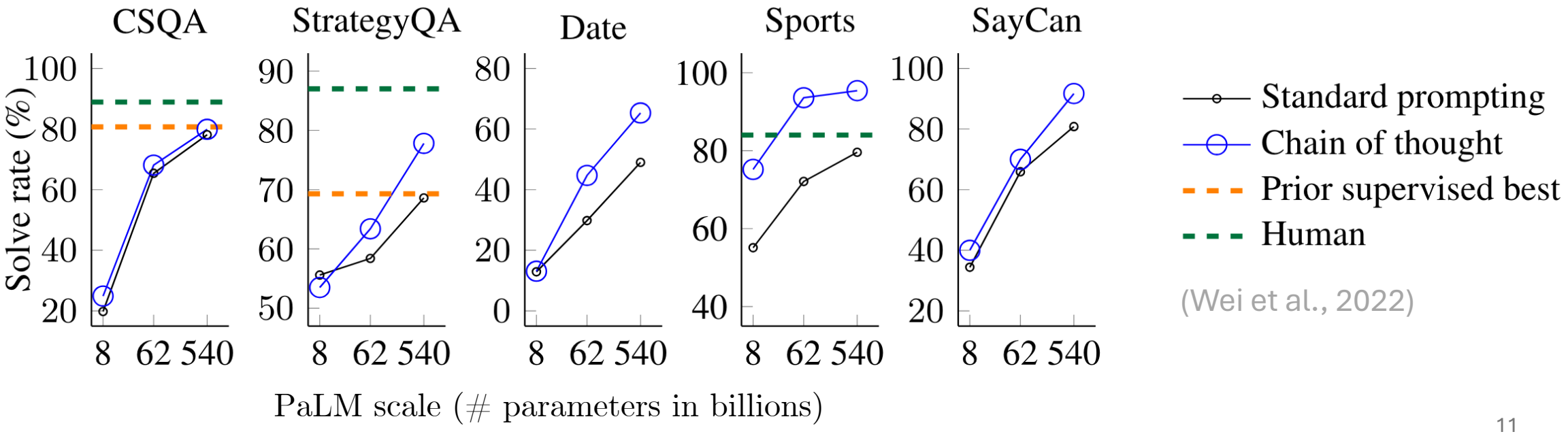
Setup & Results

- same experimental setup as before
- **benchmarks**
  - CSQA, StrategyQA, Big-Bench, SayCan

StrategyQA

Yes or no: Would an apple sink in water?

The density of an apple is about 0.8 g/cm^3, which is less than water. So, the answer is no.



(Wei et al., 2022)

Symbolic Reasoning

Setup & Results

- same experimental setup as before
- evaluated UL2, LaMDA, and PaLM

Last Letter

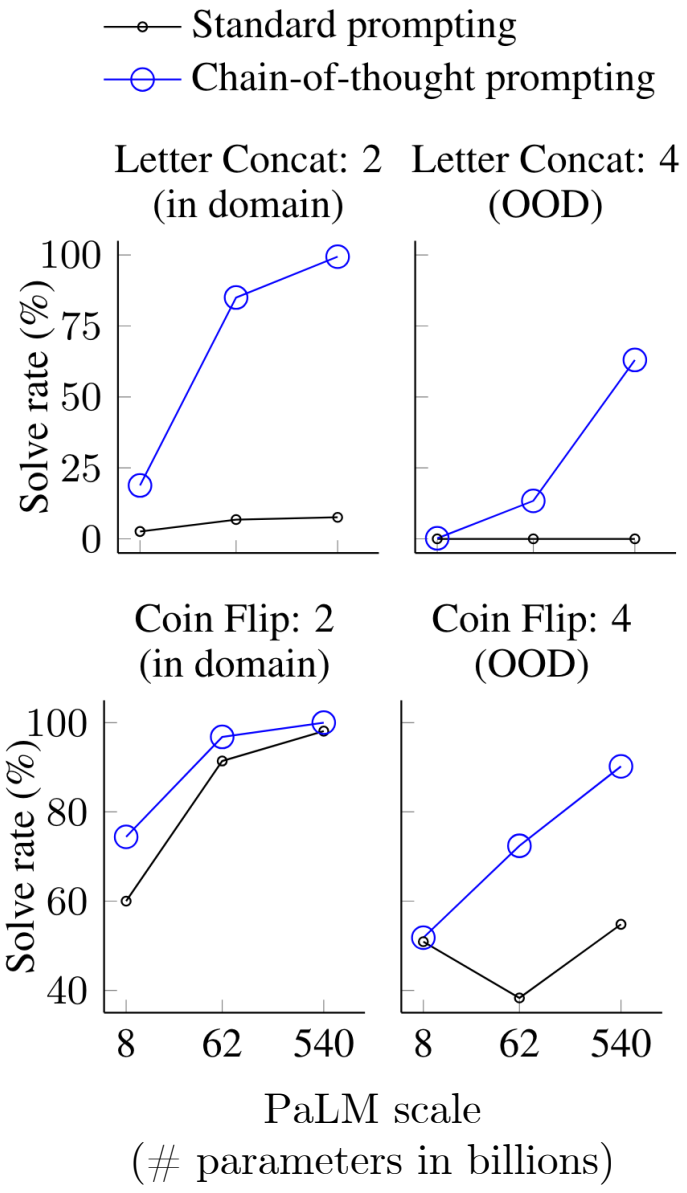
Take the last letters of the words in “Amy Brown” and concatenate them.

The last letter of “Amy” is “y”. The last letter of “Brown” is “n”. Concatenating them is “yn”. So, the answer is yn.

Coin Flip

A coin is heads up. Amy flips the coin. Ben doesn’t flip the coin. Is the coin still heads up?

The coin was flipped by Amy 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.



(Wei et al., 2022)

Chain-of-Thought Performance

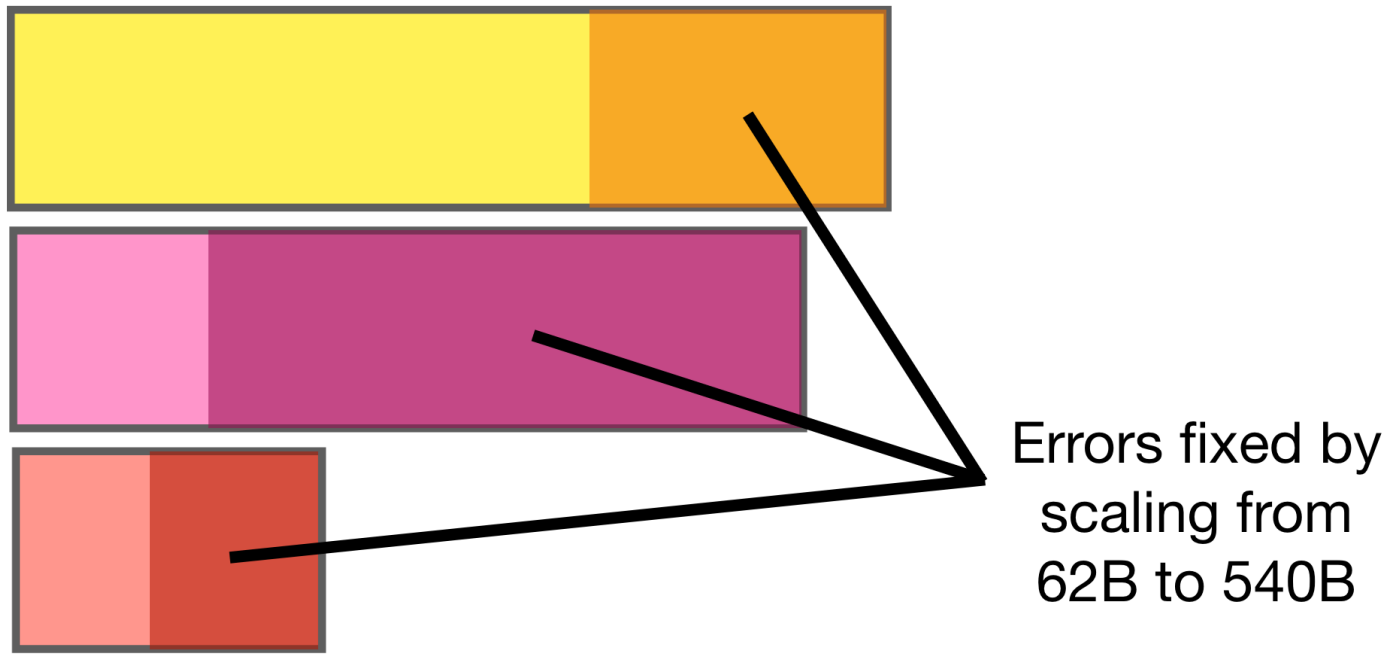
Why does increasing the model size improve CoT results?

Types of errors made by a 62B language model:

Semantic understanding
(62B made 20 errors of this type, 540B fixes 6 of them)

One step missing
(62B made 18 errors of this type, 540B fixes 12 of them)

Other
(62B made 7 errors of this type, 540B fixes 4 of them)



(Wei et al., 2022)

Conclusion

✓ Strengths

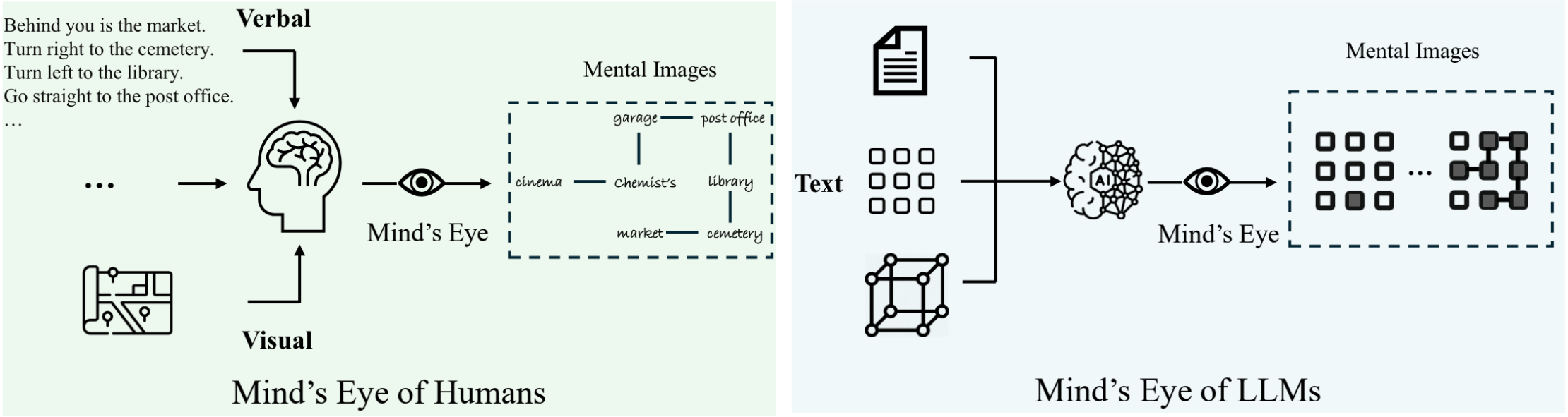
- impressive results
- added open-source model(s)
- comprehensive appendix
- useful for repeatable tasks

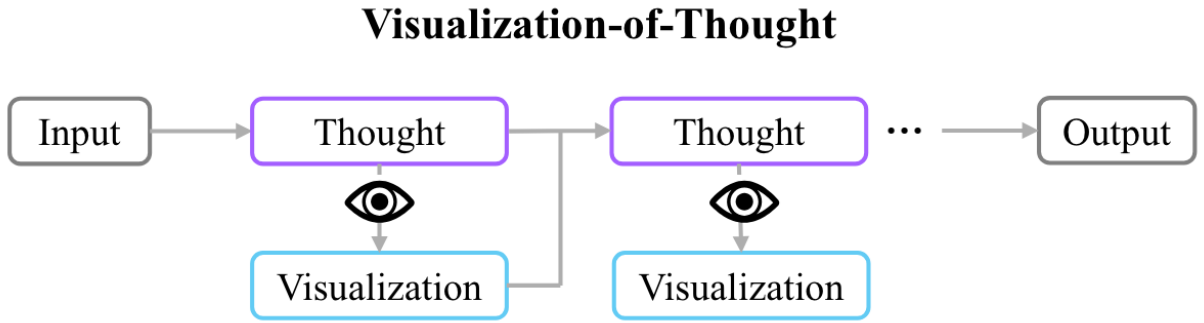
✗ Weaknesses

- only useful for large models
- reproducibility (PaLM, LaMDA)
- only empirical analysis
- writing examples requires work

New paper from 2 weeks ago

Visualization-of-Thought Elicits Spatial Reasoning in LLMs





(Wu et al., 2022)

References

- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
- Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... & Irving, G. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
- Wu, W., Mao, S., Zhang, Y., Xia, Y., Dong, L., Cui, L., & Wei, F. (2024). Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models. arXiv preprint arXiv:2404.03622.

Appendix: Arithmetic Reasoning

Results

		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**

(Wei et al., 2022)

Appendix: Arithmetic Reasoning

Results using external Python calculator

	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 (text-davinci-002)	Standard	15.6	65.7	70.3	24.8	72.7
	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 (code-davinci-002)	Standard	19.7	69.9	74.0	29.5	78.7
	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

(Wei et al., 2022)

Appendix: Arithmetic Reasoning

Four subsets of MAWPS

		SingleOp		SingleEq		AddSub		MultiArith	
Model		standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	24.9	**27.2**	18.0	**20.2**	18.5	18.2	5.0	**10.7**
LaMDA	420M	2.8	1.0	2.4	0.4	1.9	0.7	5.8	1.5
	2B	4.6	4.1	2.4	3.3	2.7	3.2	5.8	1.8
	8B	8.0	7.0	4.5	4.4	3.4	5.2	5.2	2.4
	68B	36.5	**40.8**	23.9	**26.0**	17.3	**23.2**	8.7	**32.4**
	137B	73.2	**76.2**	48.8	**58.7**	43.0	**51.9**	7.6	**44.9**
GPT	350M	3.2	1.8	2.0	0.2	2.0	1.5	2.3	0.8
	1.3B	5.3	3.0	2.4	1.6	2.3	1.5	2.2	0.5
	6.7B	13.5	3.9	8.7	4.9	8.6	2.5	4.5	2.8
	175B	90.9	88.8	82.7	**86.6**	83.3	81.3	33.8	**91.7**
Codex	-	93.1	91.8	86.8	**93.1**	90.9	89.1	44.0	**96.2**
PaLM	8B	41.8	**46.6**	29.5	28.2	29.4	**31.4**	4.2	**15.8**
	62B	87.9	85.6	77.2	**83.5**	74.7	**78.2**	7.3	**73.7**
	540B	94.1	94.1	86.5	**92.3**	93.9	91.9	42.2	**94.7**

(Wei et al., 2022)

Appendix: Arithmetic Reasoning

Ablation & Robustness (LaMDA 137B)

	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

(Wei et al., 2022)

Appendix: Commonsense Reasoning

Results

		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**

(Wei et al., 2022)

Appendix: Logical Reasoning

Results

		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**

(Wei et al., 2022)

Appendix: Commonsense & Logical Reasoning

Ablation & Robustness (LaMDA 137B)

	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

(Wei et al., 2022)