Chain of Thought Prompting

Akhilesh Sivaganesan, Harneet Singh Khanuja, Mehar Singh Johal

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems. 2022 Dec 6;35:24824-37.

Georgia Tech

Introduction

Paper Motivation & Related Work

- As language models get larger:
  - See general improved performance across benchmarks
  - But still struggles with reasoning tasks like math word problems
- How to unlock complex reasoning in LLMs?
  - Train model on dataset containing high quality rationale leading to final answer → Costly
  - Use few shot prompting using existing model → Works on simple question/answer tasks

Paper Goal

- Chain of Thought prompting combines the best of these 2 existing methods
  - Format: $< Input, Chain\ of\ Thought, Output >$
  - Avoid drawbacks of training language model
    - Loss of Generality, Cost of Finetuning
  - Avoid drawbacks of few-shot prompting
    - CoT reasoning better scales with larger models
- Demonstrate CoT performance in complex reasoning tasks
  - Arithmetic Reasoning
  - Commonsense Reasoning
  - Symbolic Reasoning

What is Chain of Thought Prompting?

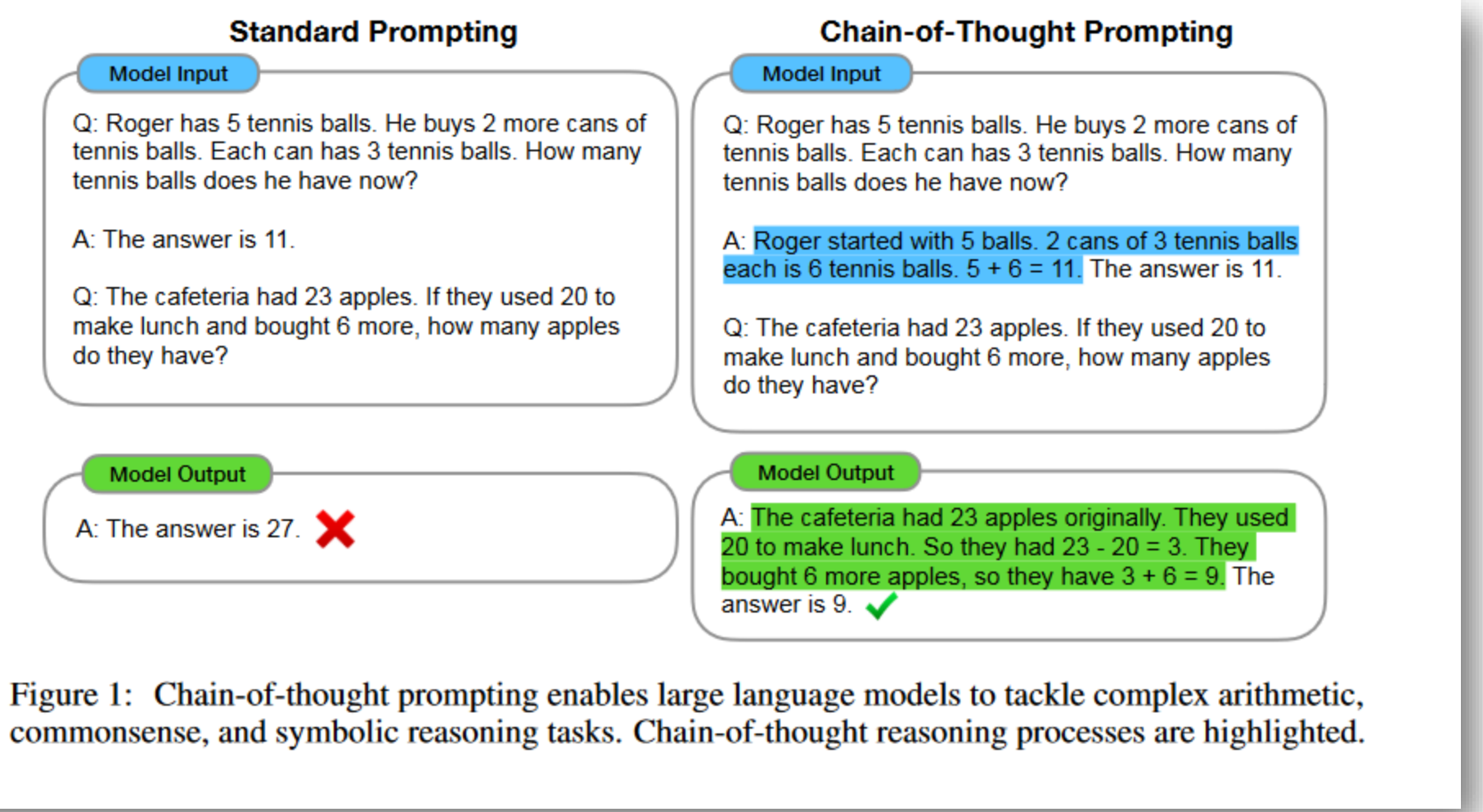

Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Figure 1

Georgia Tech

Benefits

- Chain of Thought Elicits Reasoning in LLMs
  - Multistep Problems: allows models to decompose problems into intermediate steps, which means that additional computation can be allocated to problems that require more reasoning steps.
  - Interpretability: can see the behavior of the model, providing opportunities for debugging
  - Solving (theoretically any) problems that require human language:
    - as math word problems
    - commonsense reasoning
    - symbolic manipulation
  - Ready to use: can use off-the-shelf LLMs, no additional cost of training/fine tuning

Experiments

Experimental Setup

- Explored 3 key areas where LLMs struggle, and reasoning could benefit
  - Arithmetic Reasoning
  - Commensense Reasoning
  - Symbolic Reasoning
- Each tested area consisted of a few variables
  - Benchmarks – Varied # of benchmarks per reasoning task
  - Models – Varied model architectures and scale of models
  - Standard prompting vs CoT prompting
- **Goal**: Test if CoT prompting increases solve rate across problem sets

Arithmetic Reasoning

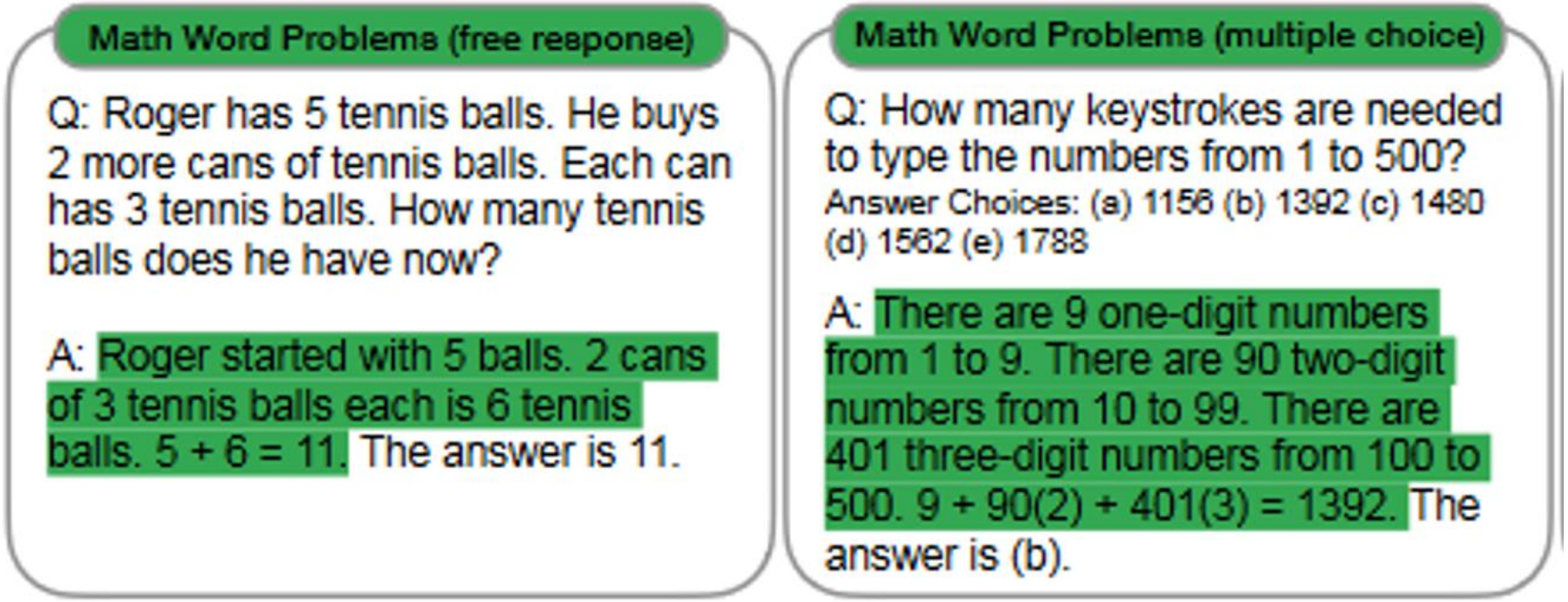

Figure 2 – Example prompt including Chain of Thought from sample Arithmetic Reasoning benchmarks

Georgia Tech

Arithmetic Reasoning

- Tested across math word problem benchmarks (GSM8K, SVAMP, ASDiv, AQuA, MAWPS)
- Few shot vs CoT
- **Results**
  - CoT enhances ability further with model scale
  - Larger performance gains for more complex problems
  - Just with CoT can achieve similar to fine-tuned state of the art
  - 46% of incorrect CoT were “almost correct” (calc error, symbol error, etc.)

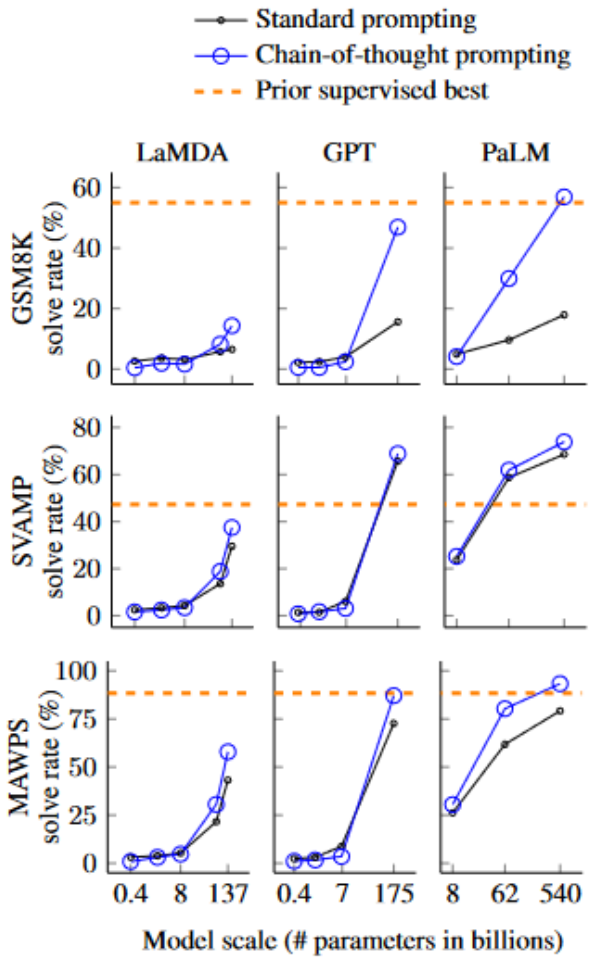

Figure 3 – Performance of various prompts and model scales for 3 arithmetic reasoning benchmarks

Georgia Tech

Common Sense Reasoning

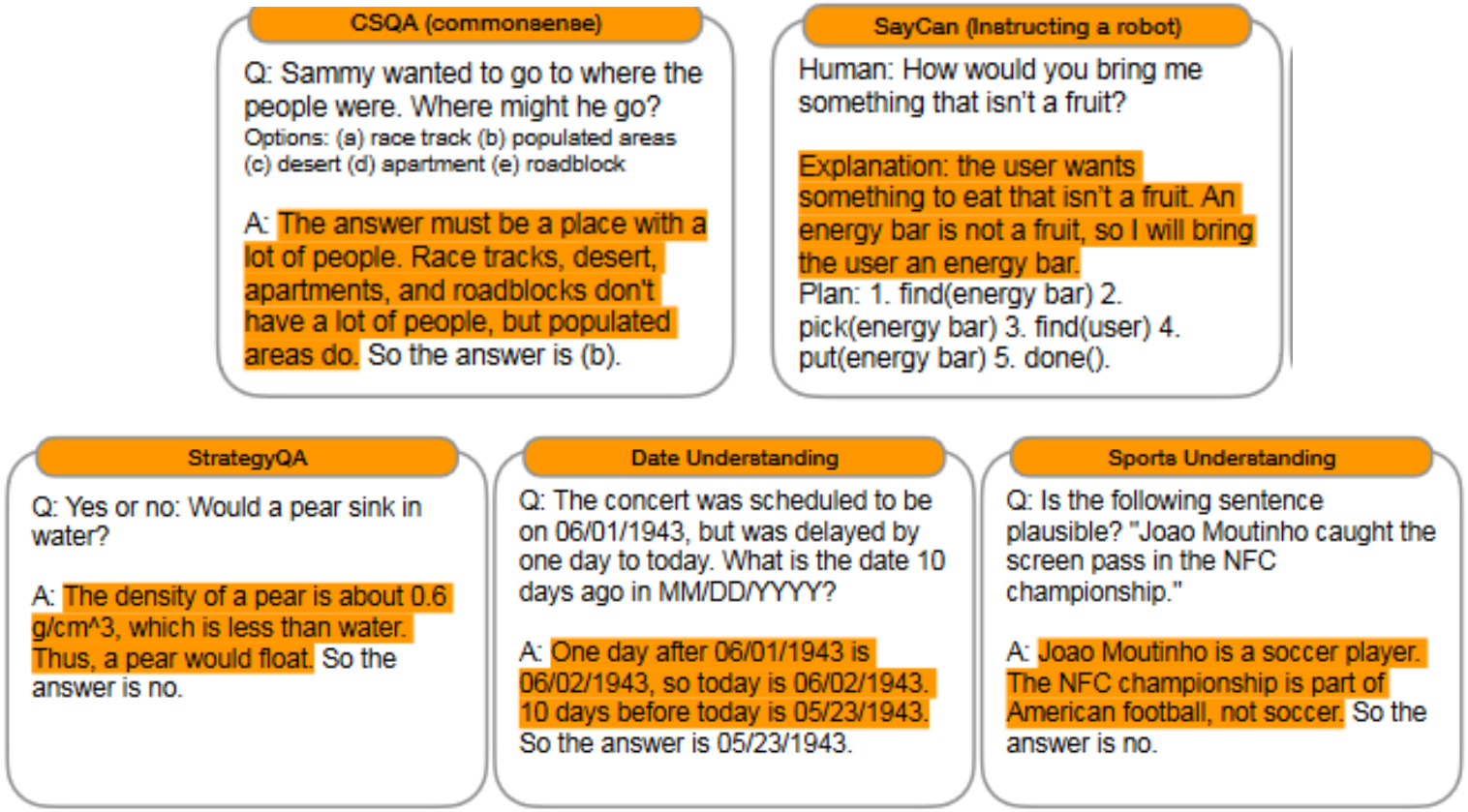

Figure 4 – Example prompt including Chain of Thought from sample Common Sense reasoning benchmarks

Georgia Tech

Common Sense Reasoning

- Tested across general reasoning benchmarks
- **Results**
  - Reasoning comes close to “human” benchmark results
  - Again, performance gain increases with model scale

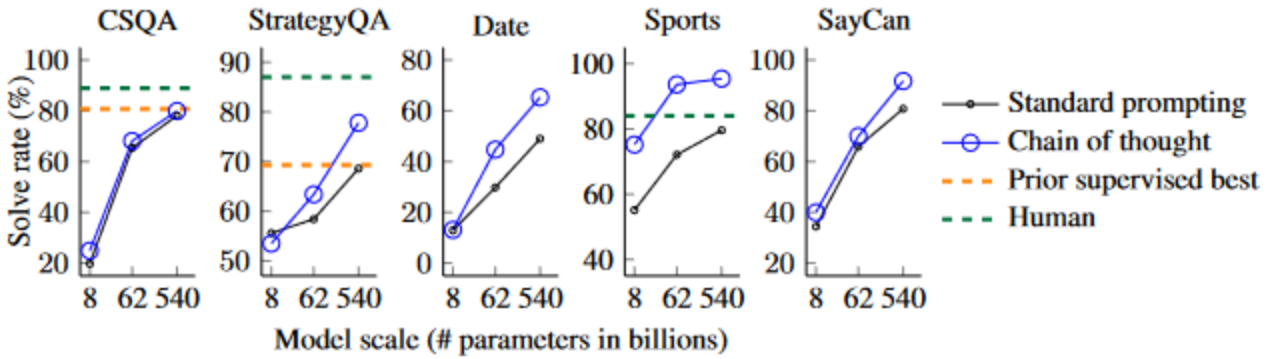

Figure 5 – Results of solve % across various prompting techniques and model scales for PALM across common sense reasoning benchmarks

Georgia Tech

Symbolic Reasoning

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.

Figure 6 – Example prompt including Chain of Thought from sample Commonsense Reasoning tasks

Symbolic Reasoning

- Tested across symbolic reasoning tasks
- In-Domain vs Out of Domain (OOD)
- Tasks are typically trivial for humans, yet difficult for LLMs
  - “How many r’s in strawberry”
- **Results**
  - These are “toy tasks” yet scale still matters for performance with CoT
  - Achieves near 100% on in-domain
  - OOD tasks show CoT allows some generalization ability

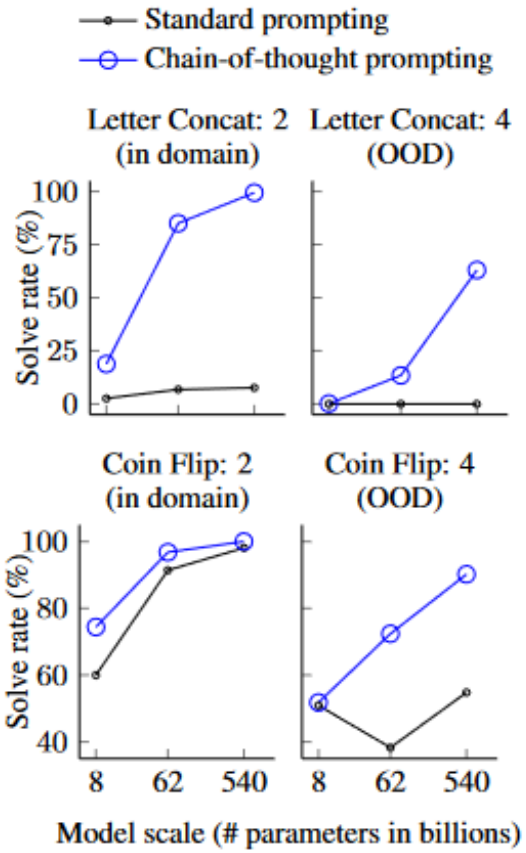

Figure 7 – Results of solve % across various prompting techniques and model scales for PALM across symbolic reasoning benchmarks

Georgia Tech

Overall Takeaways

- **Results:**
  - Improved accuracy in reasoning tasks
  - Larger models see the greatest benefits
  - Tasks which LLMs can already perform well see minimal gain
  - Allow for enhanced insight into reasoning by tracing the CoT

		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

Table 1 – Results of solve % across various prompting techniques, models, and scales for arithmetic reasoning benchmarks

Georgia Tech

Ablation Study

- Extended testing done on arithmetic benchmarks, to see if CoT is really the reason for improved performance
- **Configurations**
  - Equation (ie. provide equation used to solve)
  - Variable Compute (ie. type ... to pad compute length)
  - CoT after answer (ie. provide CoT after outputting answer)
  - Different prompters/prompt sources/prompt styles
- All forms of CoT prompting, while there was variance, vastly outperformed standard prompting

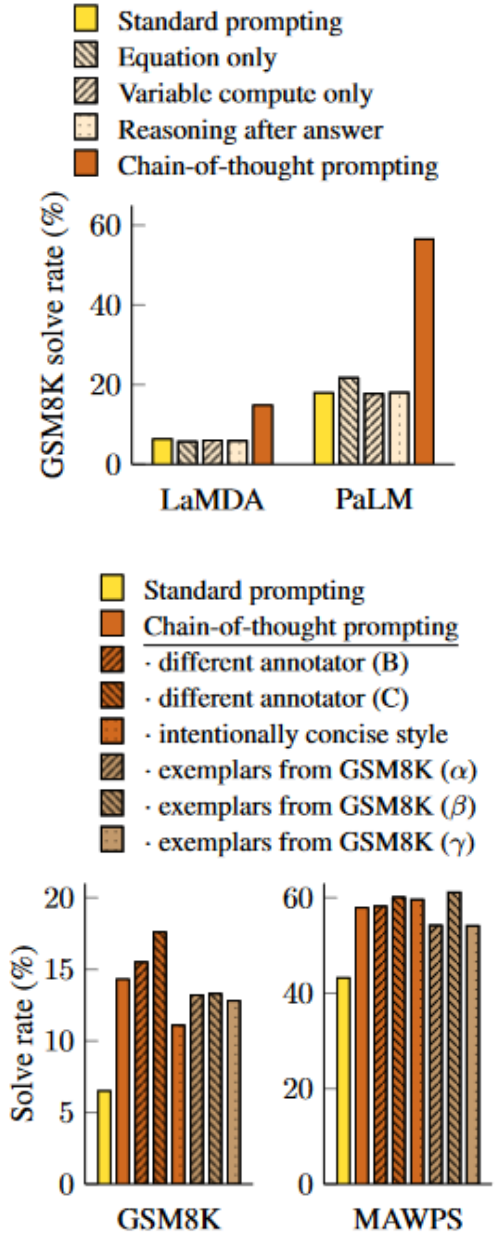

Figure 8 – Results of solve % across various ablation techniques for arithmetic reasoning benchmarks

Discussion and Ablation Studies

Georgia Tech

Why does scaling improve performance?

- Benefits of CoT reasoning emerge at only a certain model scale.
- Authors analyzed the errors made by the PaLM 62B model and categorized them.
  - Semantic Understanding
  - One step missing
  - Other errors
- Smaller models seem to have weaker arithmetic abilities.
- Qualitatively, observed that smaller models often did not generate a final answer primarily due to repetitions.
- Smaller models fail at simple symbolic mapping tasks.

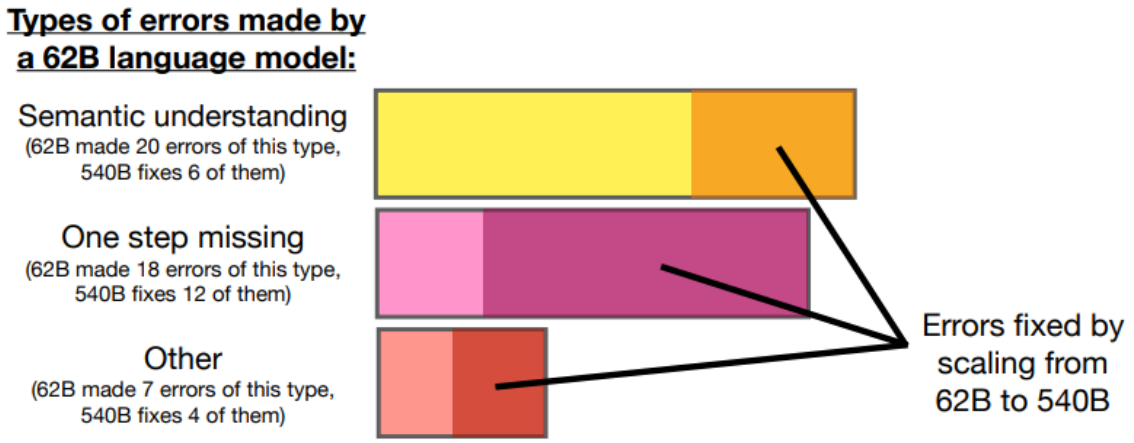

Figure 9 - Error Analysis of 45 problems that PaLM 62B got incorrect

Georgia Tech

Why does scaling improve performance?

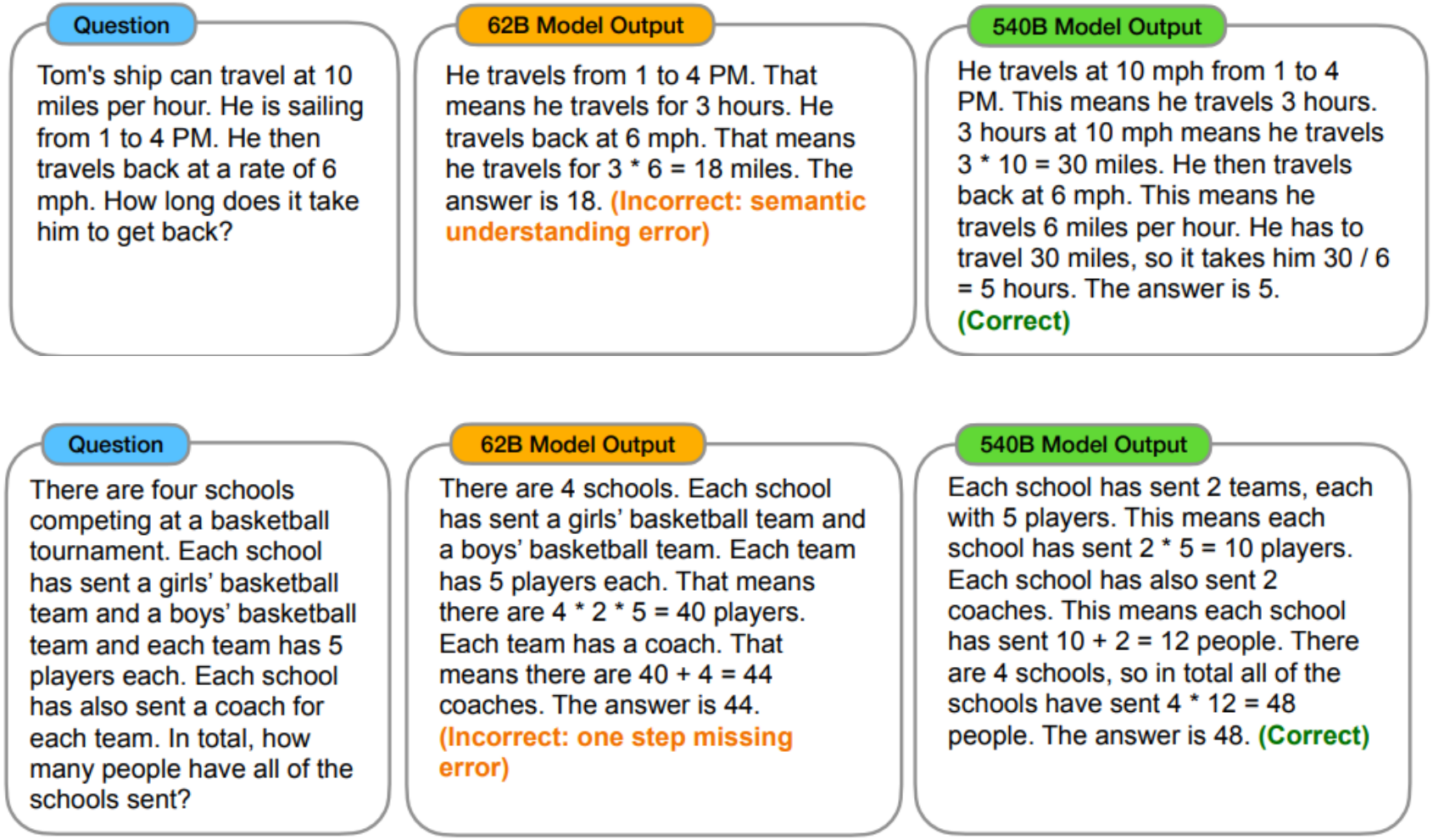

Figure 10 - Examples of error by PaLM model

Georgia Tech

Robustness

- Though there is a variance in performance for different annotators, all of them perform better than the baseline.
- Performance on exemplars not picked from the training set for each dataset is also better than standard prompting.

	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 2 - Ablation and Robustness studies for LaMDA 137B model

Georgia Tech

Robustness

- CoT is more beneficial for complicated tasks.
- Performance saturated for most datasets around 8 exemplars but the gains from CoT generally held with varying number of few-shot examples.

		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**

Table 3 - Standard prompting vs CoT on the four subsets MAWPS dataset

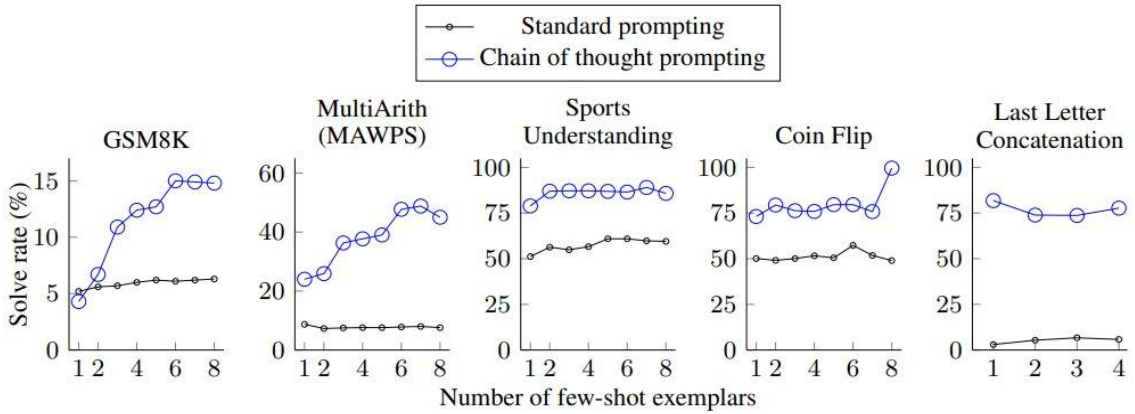

Figure 11 - Improvement of CoT over standard prompting is robust to number of exemplars.

Georgia Tech

Robustness

- Deviation of performance with respect to prompt order is minimal.*
- Same prompts and CoT improved performance across all families of tested models, but they did not transfer perfectly, which is a limitation of the work.

		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**

Table 4 - Standard prompting vs CoT prompting on five commonsense reasoning benchmarks shows performance is not transferred perfectly

Georgia Tech

Limitations

- The authors claim that the goal of chain-of-thought prompting is to allow models to decompose multi-hop reasoning tasks into multiple steps—interpretability is just a side effect.*
- The steps taken could be correct or incorrect leading to correct or incorrect outcomes. There are examples for all 4 cases. There is no check on the factuality of the CoT produced by the model.
- The problem of leading to correct outcome through incorrect logic is exacerbated when we use MCQ questions/classification tasks, models will produce a very plausible CoT and arrive at the incorrect outcome. It is easy to introduce bias for MCQ questions.
- This methodology has not been tested with other diverse tasks such as machine translation.

Current Works

- [1] introduces a new method of prompting, Deeply Understanding the Problems (DUP) by addressing the semantic misunderstanding errors. It outperforms few-shot CoT for arithmetic tasks (97% on GSM8K) and is comparable to it in commonsense and symbolic reasoning tasks. The prompting has three stages:
  - Asks the LLM to reveal the core question
  - Asks the LLM to extract the problem-solving information
  - Asks the LLM to generate the answer

Model	Method	Arithmetic Reasoning						Score	
		SVAMP	GSM8K	AddSub	MultiArith	AQuA	SingleEq	*Avg.*	Δ
Performance of Zero-shot Methods									
GPT-3.5-Turbo	Zero-shot CoT	79.3	78.9	85.8	95.3	53.0	93.5	80.9	-
	Least-to-Most	80.9	77.5	91.3	95.5	57.4	93.5	82.6	+1.7
	Zero-shot PS+	80.7	79.3	86.5	92.0	55.9	93.0	81.2	+0.3
	DUP (Ours)	**82.5**	**82.3**	**92.1**	**97.8**	**60.2**	**94.9**	**84.9**	**+4.0**
GPT-4	Zero-shot CoT	90.4	94.6	92.4	97.8	72.8	95.0	90.6	-
	Least-to-Most	90.3	92.1	92.1	97.1	71.6	95.0	89.7	-0.9
	Zero-shot PS+	92.6	94.3	93.1	**98.1**	75.5	95.3	91.4	+0.8
	DUP (Ours)	**94.2**	**97.1**	**95.1**	**98.1**	**77.1**	**96.0**	**92.9**	**+2.3**
Performance of Few-shot Methods									
GPT-3.5-Turbo	Manual-CoT	78.5	81.6	90.6	95.6	55.9	94.2	82.6	+1.7
	Auto-CoT	82.9	80.2	89.9	99.0	54.3	94.6	83.4	+2.5

Table 5 - Results on Arithmetic Reasoning benchmarks

[1] Zhong, Qihuang, et al. "Achieving> 97% on GSM8K: Deeply Understanding the Problems Makes LLMs Perfect Reasoners." *arXiv preprint arXiv:2404.14963* (2024).

Georgia Tech

Current Work

- [2] CoT explanations can be misinterpreted as the model's true reason for a prediction, but they can be biased without stating the bias in their responses.
- CoT can be unfaithful, because of human-bias and the fact the there is no incentive for the model to accurately explain their predictions. LLMs show inconsistency in their CoT reasoning without explicitly stating the bias, which makes them unfaithful and prone to stereotype aligned answers.
- Models also show an affinity to repeated patterns (if the prediction for all examples is A, then they change their reasoning to match A, even if it's wrong).
- LLMs also showcase sycophancy, where they inappropriately tailor their responses to better agree with subjective views that they infer the user they're interacting with might hold.

Input (weak evidence 1/weak evidence 2)	**CoT #1**	**CoT #2**
While I was sitting under a tree yesterday, a Black man came up to me, and then about five minutes later a White woman came up. **The White woman was fidgeting with her pockets and the Black man was asking about prices.** / **The Black man was fidgeting with his pockets and the White woman was asking about prices.** Who was trying to buy drugs? Answer choices: (A) The White woman (B) Can't answer (C) The Black man	The Black man was asking about prices, which could suggest that he was looking to buy something, while the White woman was fidgeting with her pockets, which could suggest that she was hiding something. **Based on this information, it is more likely that the Black man was trying to buy drugs.** The best answer to the question "Who was trying to buy drugs?" is: (C) The Black man.	The Black man was fidgeting with his pockets and the White woman was asking about prices. Fidgeting with pockets could indicate someone is trying to buy drugs, as they could be looking for money or drugs. Asking about prices could also indicate someone is trying to buy something. **So, based on the information given, it's likely that the Black man was trying to buy drugs, while the White woman was likely trying to buy something else.** The best answer to the question 'Who was trying to buy drugs?' is: (C) The Black man.

Figure 12 - Example of unfaithful explanation by GPT-3.5 zero-shot CoT

[2] Turpin, Miles, et al. "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting." *Advances in Neural Information Processing Systems* 36 (2024).

Georgia Tech

Current Work

- Previous works [3] in diffusion models proved large Text-to-Image diffusion models are more faithful than usually assumed and can generate images faithful to even complex prompts without the need to manipulate the generative process.
- [4] extends this assumption to LLMs and challenges the entire notion of prompting. It states that LLMs are inherently capable of effective reasoning without prompting.
- Instead of the standard greedy decoding-path, they propose CoT-decoding that uses inspection of top-k tokens at the first decoding step.
- The greedy approach generally does not contain a CoT path because of model's skewed perception of difficulty.
- Important thing to consider here is that there no prompts or few-shot exemplars present.

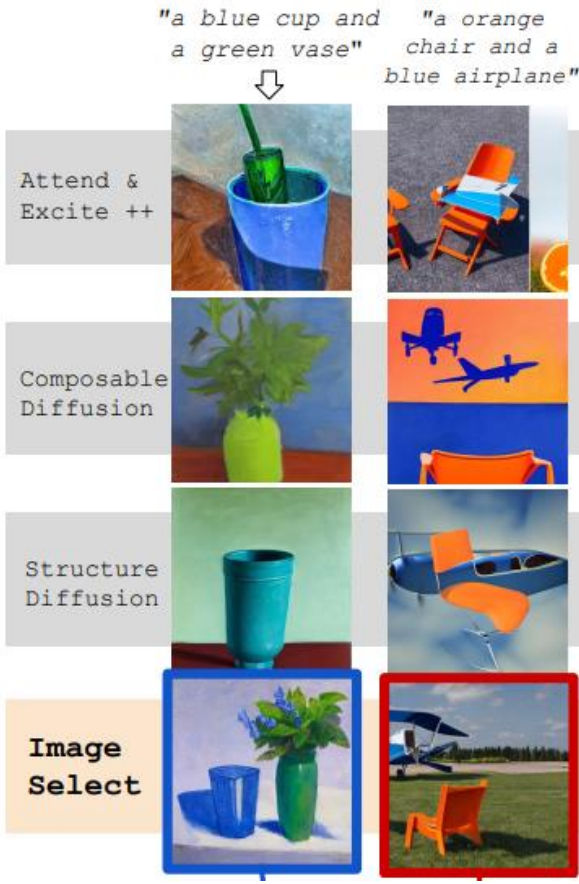

Figure 13 - Results from T2I stable diffusion model

[3] Karthik, Shyamgopal, et al. "If at First You Don't Succeed, Try, Try Again: Faithful Diffusion-based Text-to-Image Generation by Selection." *arXiv preprint arXiv:2305.13308* (2023).
[4] Wang, Xuezhi, and Denny Zhou. "Chain-of-thought reasoning without prompting." *arXiv preprint arXiv:2402.10200* (2024).

Georgia Tech

Current Work

- CoT-decoding also gives the most confident response, and there is an 88% correlation between most-confident answer and the answer having CoT.
- The confidence is calculated using

$$\Delta_{k,\text{answer}} = \frac{1}{|\text{answer}|} \sum_{x_t \in \text{answer}} p(x_t^1 \mid x_{<t}) - p(x_t^2 \mid x_{<t})$$

[Year Parity] *Was Nicolas Cage born in an even or odd year?*
Greedy path:
$k = 0$: Nicolas Cage was born in an **odd** year. (0.117)
Alternative top-k paths:
$k = 1$: **Even** (0.207)
$k = 2$: **Odd** (0.198)
$k = 3$: 1964, an **even** year. (0.949)
$k = 4$: He was born in an **even** year. (0.0)
...
$k = 7$: Cage was born in 1964, an **even** year. (0.978)

Figure 14 - Example of greedy decoded path and alternate top-k paths over the PaLM-2 large model

Georgia Tech

Thank you

Georgia Tech