Self-Consistency Improves Chain of Thought Reasoning in Language Models

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sahran Narag, Aakanksha Chowdhery, Denny Zhou (Google Research, Brain Team)

Published at ICLR 2023

Presenters: Andrew Liu and Kyung Jae Lee

Background on Reasoning in Language Models

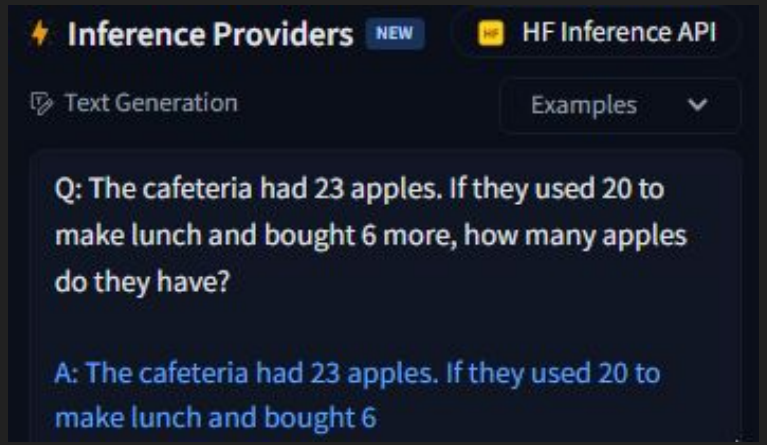
Works on transformer based architectures have led to foundation models that, when fine-tuned, can lead to SOTA performance on numerous NLP benchmarks once thought very difficult.

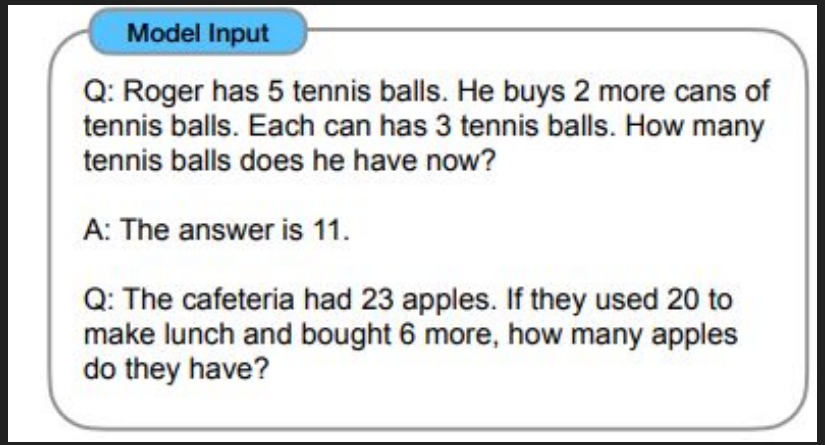
- BERT
- GPT2

However, fine-tuning still demands meticulously crafted task-specific datasets, as well as additional compute to perform the fine-tune updates.

Devlin et al. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Radford et al. (2019) Language Models are Unsupervised Multitask Learners

Prompting to Extract Reasoning



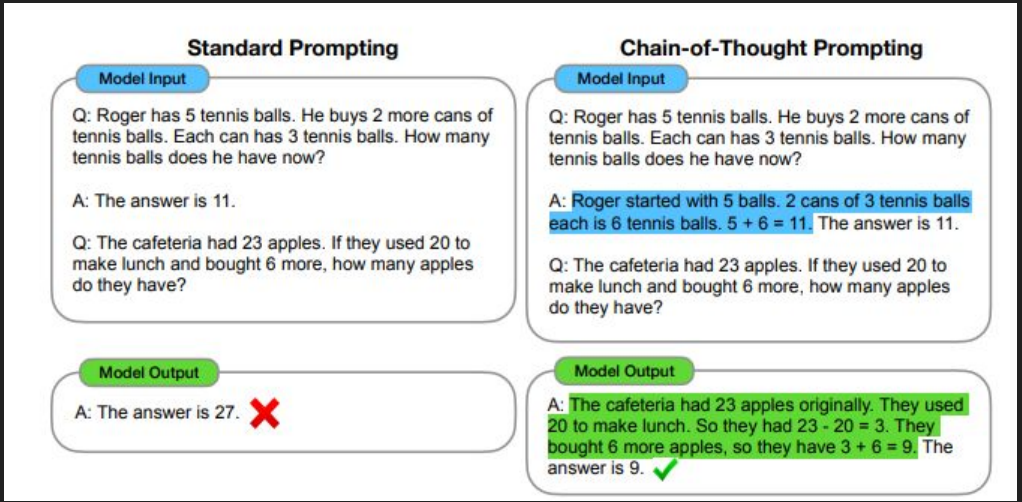


Language models in their base form are simply not trained to answer questions. Prompting can fix this without any additional tuning of the model

Wei et al. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models provided the figure

CoT Prompting

Helpful to not only give the model examples of answers, but to include in the examples how they arrived at the answer. At a high level, this encourages the model itself to ‘think things through.’



Wei et al. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models provided the figure

Decoding Choices

For autoregressive models like GPT3, given a sequence of prior tokens z_1, z_2, z_3, …, z_{n-1} , it produces a distribution across the vocabulary size that tells us the probability of the next tokens. This gives us some degrees of freedom in terms of how we decode

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1})$$

Decoding Choices

Greedy Decoding

Pros: Always taking the most likely token, less likely for inexplicable outputs to appear (hallucinations)

Cons: Boring, uncreative outputs. Heuristic that ‘things that appeared already are more likely to appear again’ leads to the model repeating over and over again

Decoding Choices

Sampling Based Decoding

Temperature:

$$P(x_i|x_{1:i-1}) = \frac{\exp(u_i/t)}{\sum_j \exp(u_j/t)}$$

Top-k: Take only the k highest likelihood tokens, then sample from those

Top-p (Nucleus): Take only the tokens in the top p% of tokens, then sample from those

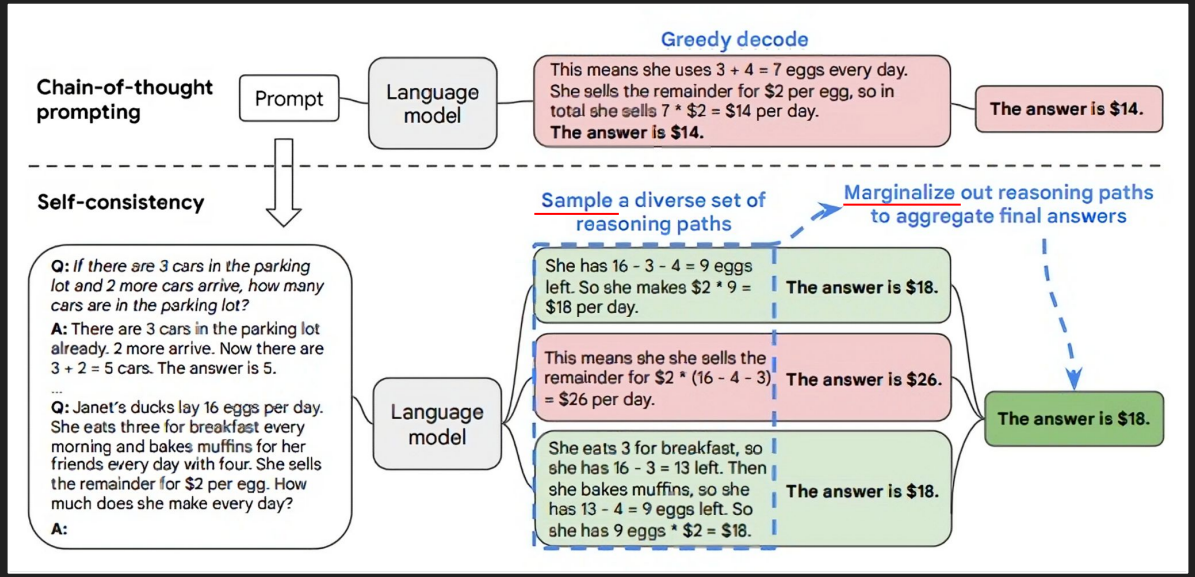
Concerns with Sampling Based?

The freedom we afford the model to make when choosing the next token gives it creativity, but also can lead to instability.

The task of QA has always been more aligned with greedy decoding because of the desire for accuracy, self-consistency instead uses a sampling based method and leverages that creativity for better QA or problem solving performance.

Self-Consistency

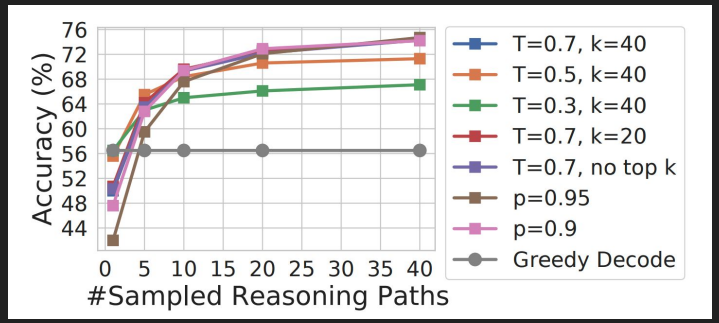
Intuition: Reasoning tasks often have multiple valid paths to arrive at the correct answer



Self-Consistency Improves Chain of Thought Reasoning in Language Models. Wang et al. ICLR 2023

Self-Consistency Steps

1. **Sample** multiple reasoning paths
   a. Response format: “{Reasoning path}. The answer is X.”
   b. Self-consistency robustly improves performance across all sampling strategies



Self-Consistency Improves Chain of Thought Reasoning in Language Models. Wang et al. ICLR 2023

2. **Marginalize** out final answer with a majority vote

Evaluation - Types of Tasks

1. Arithmetic reasoning
    - There are 64 students trying out for the school's trivia teams. If 36 of them didn't get picked for the team and the rest were put into 4 groups, how many students would be in each group?
2. Commonsense reasoning
    - George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?
    - A) "dry palms" B) "wet palms" C) "palms covered with oil" D) "palms covered with lotion"
3. Symbolic Reasoning
    - Concatenate the last characters of each word in the input
    - E.g., “self consistency is cool” -> “fysl”

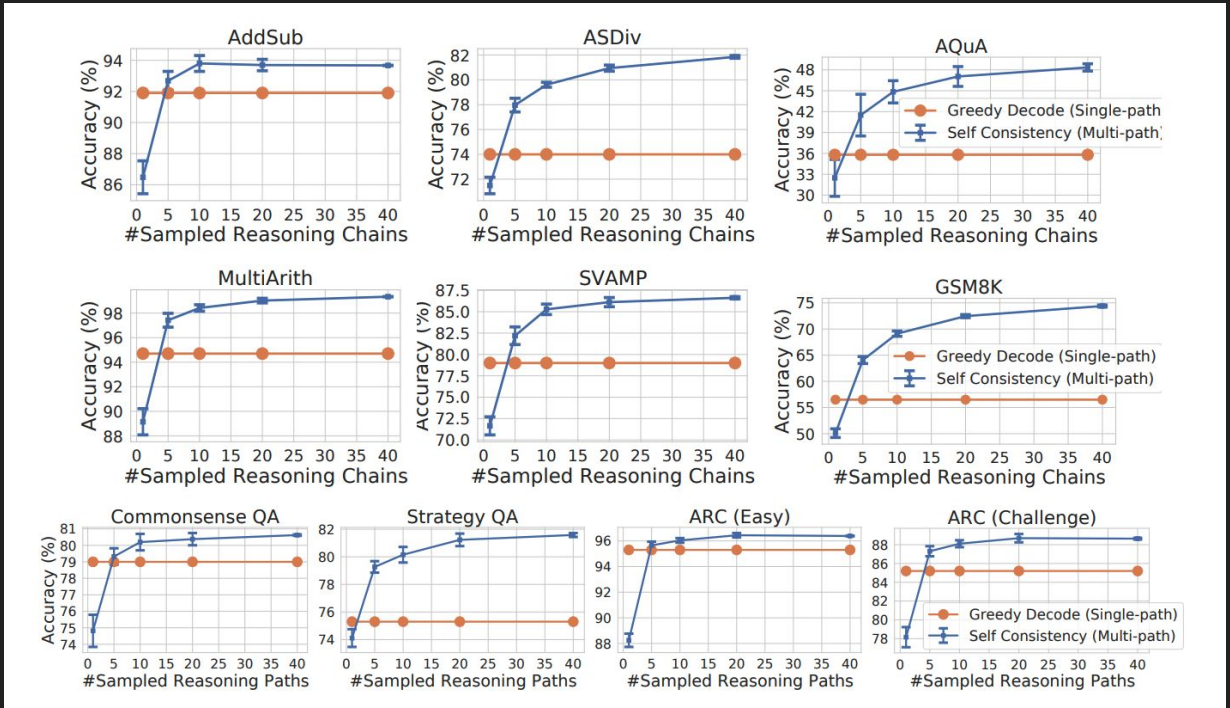
- (Mostly) Similar datasets/prompts as original CoT paper

Evaluation - Models Used

1. UL2 (20B)
2. GPT-3 (175B)
3. LaMDA-137B
4. PaLM-540B

All models are used as provided, without additional training or fine-tuning

Evaluation - Results



Self-Consistency Improves Chain of Thought Reasoning in Language Models. Wang et al. ICLR 2023

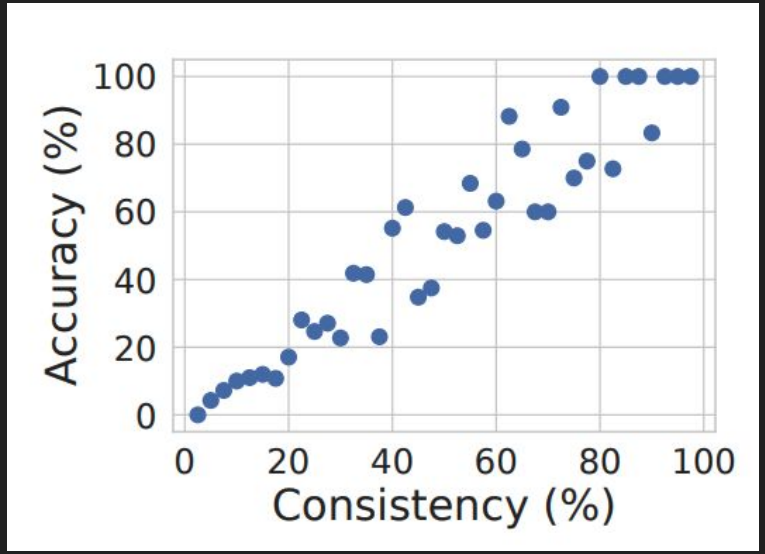
Advantages of Self-Consistency

1. Off-the-shelf method
    a. No need for additional training / human annotation / auxiliary models
    b. Self-ensemble
2. Effective
    a. Achieves state-of-the-art accuracy across many reasoning tasks
    b. Higher accuracy as more reasoning paths are sampled, but most gains are from first 5-10
3. Robust
    a. Sampling strategy and parameters
    b. Type of prompt - even improves performance on zero-shot / imperfect prompts
    c. Type of (reasoning) task
    d. Type/scale of model

Advantages of Self-Consistency

4. Consistency as a measure of confidence

- Consistency = Percentage of responses that agree on final answer



Self-Consistency Improves Chain of Thought Reasoning in Language Models. Wang et al. ICLR 2023

Scope and Limitations

Increased computation cost

-> Batch requests, prompt caching

Difficult to apply for problems with a free-form answer

-> Universal self-consistency (Chen et al., 2023)

Thank you!