10. Scaling Language Models

Guest lecture by Jason Wei

Scaling language models

Scaling measures: model size (# parameters), training data (# tokens), training compute (FLOPs)

Why is scaling challenging?

- Compute is expensive
- Model parallelism & hardware challenges
- Loss divergences & hardware failures
- Other reasons: inductive bias, different from human learning, incentive mismatch

Scaling laws: Language modeling performance improves smoothly as we scale up the model

What can’t language models learn from next-word prediction?

Current world knowledge, arbitrarily long arthmetic, many-step reasoning

Emergent abilities

Emergent abilities are abilities that are not present in small models, but are present in large models.

There are a lot of emergent abilities.

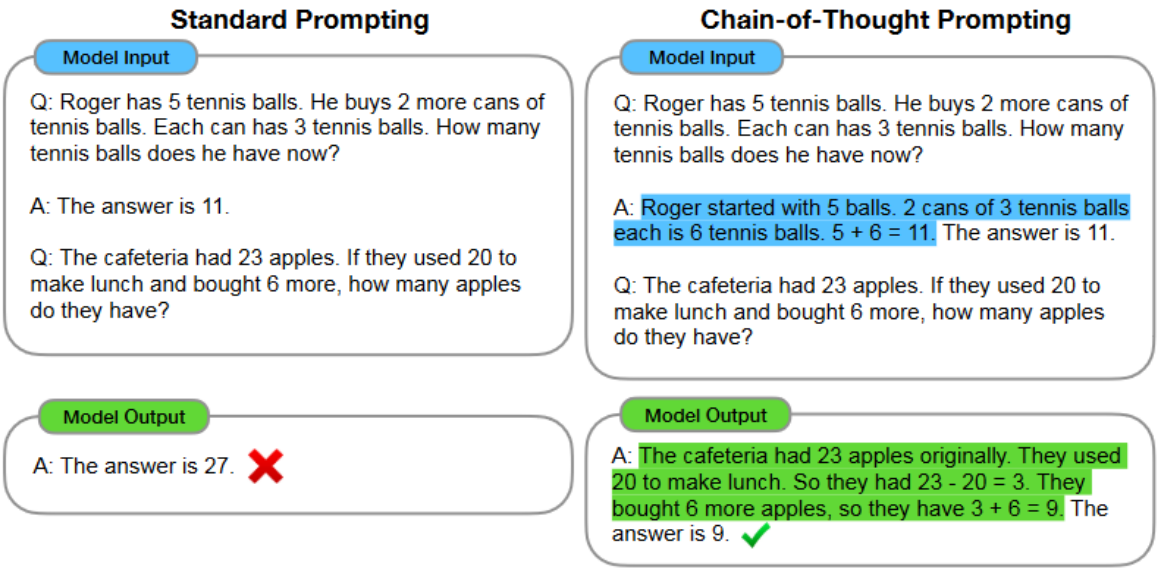
There are a lot of unknown unkowns. We don’t know the full range of emergent abilities.

One model, any task.

Prompt engineering

Multi-step reasoning is hard for language models.

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



Self-Consistency Improves Chain of Thought Reasoning in Language Models: a new decoding strategy, self-consistency, to replace the naive greedy decoding used in chain-of-thought prompting. It first samples a diverse set of reasoning paths instead of only taking the greedy one, and then selects the most consistent answer by marginalizing out the sampled reasoning paths. Self-consistency leverages the intuition that a complex reasoning problem typically admits multiple different ways of thinking leading to its unique correct answer.