Iteration Heads:
A Mechanistic Study of Chain-of-Thought

Vivien Cabannes, Charles Arnal, Wassim Bouaziz, Alice Yang,
Francois Charton, Julia Kempe

Meta AI, FAIR, Core Learning & Reasoning

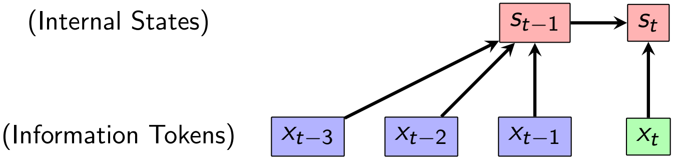
NeurIPS 2024

Why and how does Chain-of-Thought (CoT) improve transformers' reasoning capabilities ?

Reasoning and Iterative Algorithms

Reasoning involves updating some internal state, representing the current thought process, as new information is incorporated.

It can be framed as an iterative algorithm $s_t = F(s_{t-1}, x_t)$.



(Internal States)

s_{t-1} s_t

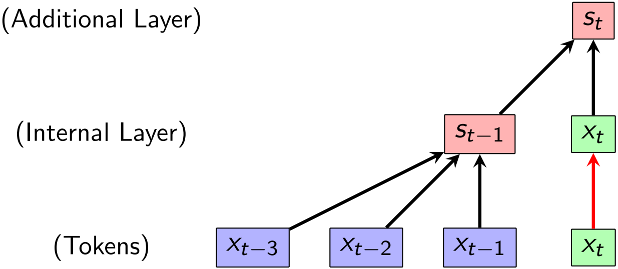
(Information Tokens)

x_{t-3} x_{t-2} x_{t-1} x_t

Transformers and Reasoning

How can transformers learn to reason iteratively?

Transformers and Reasoning - Naive Solution: Stack Layers



Problem: Number of reasoning hops limited by number of layers.

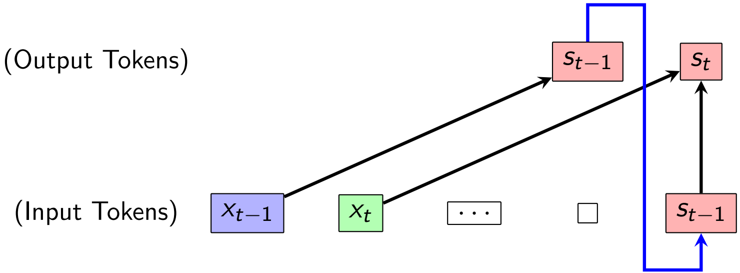
Transformers and Reasoning - Better Solution: CoT

Chain of Thought has been observed to drastically improve reasoning capabilities.

Can we find a mechanistic explanation?

Iteration Heads

A simple internal circuit can implement iterative reasoning with CoT by writing latent states into token space.



This "*Iteration Head*" circuit can be implemented with only two transformer blocks.

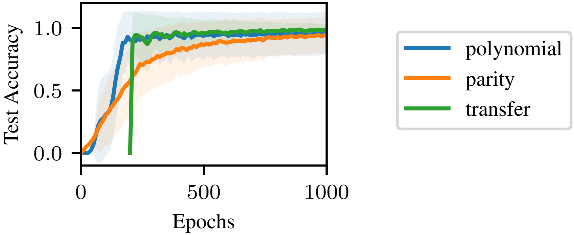
Iteration Heads - Experimental Results

Experiments in a maths-inspired controlled setting show that such circuits do emerge in practice.

Iteration Heads - Transfer Learning

Reasoning capabilities yielded by iteration heads transfer well to other iterative tasks.

In fact, we observe that training on a highly structured iterative task first can help learn other iterative tasks faster.



This might help explain the importance of data mixing (e.g. adding code to training data).