Mixture of Parrots 🦜🦜🦜: Experts improve memorization more than reasoning

Samy Jelassi[(1)] Clara Mohri[(1)] David Brandfonbrener[(1,2)] Alex Gu[(3)] Nikhil Vyas[(1)] Nikhil Anand[(2)]
David Alvarez-Melis[(1,2)] Yuanzhi Li[(4)] Sham Kakade[(1,2)] Eran Malach[(1,2)]

[(1)]Harvard University [(2)]Kempner Institute [(3)]MIT [(4)]Microsoft Research

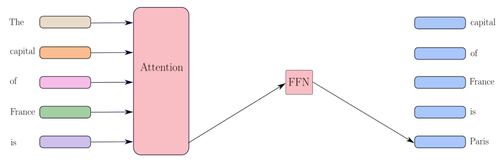
1 - Summary

Mixture of Expert (MoEs) were recently introduced as an alternative to Transformer-based language models since they considerably increase the total parameter count while maintaining a low inference cost. We show that no matter the number of experts, a critical width is needed for solving reasoning tasks. In contrast, increasing the number of experts improves the performance on memorization tasks.

2 - Motivation

Scaling Laws for Language Models [1]: **More parameters** in the model ⇒ **lower perplexity**.

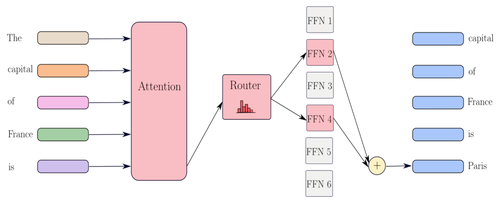
Transformers scale poorly



Parameter count & Inference **increase with width**

- Parameters: $O(d^2)$
- Inference: $O(d^2)$

Mixture of Experts (MoEs): scaling parameter count while maintaining cheap inference cost



– Models that replicate the FFN block E times (experts).

– Router assigns each token to two experts.

– Inference **independent of number of experts**

- Parameters: $O(Ed^2)$
- Inference: $O(d^2)$

When to choose MoEs over Transformers?

– Computational constraints at inference

– High training compute/data budget

– Many accelerators (GPUs/TPUs) available

– **Our work**: at **equal parameter count**, do Transformers outperform MoEs?

3 - Setting

"Reasoning" vs "Memorization" downstream tasks

Problem. If $n \equiv 2 \pmod 7$, then find the remainder when $(n+2)(n+4)(n+6)$ is divided by 7.

Solution. Since $n \equiv 2 \pmod 7$, we can write $n = 7k+2$ for some integer k. Substituting, we get $(n+2)(n+4)(n+6) = (7k+4)(7k+6)(7k+8) \equiv 4 \cdot 6 \cdot 1 \pmod 7$. Lastly, $4 \cdot 6 \cdot 1 \equiv 24 \pmod 7 \equiv 3 \pmod 7$.

Math reasoning (MATH)

Question. Sammy wanted to go to where the people were. Where might he go?

Choices. A) race track B) populated areas C) desert D) apartment

Answer. B) populated aread

NLP reasoning (CommonsenseQA)

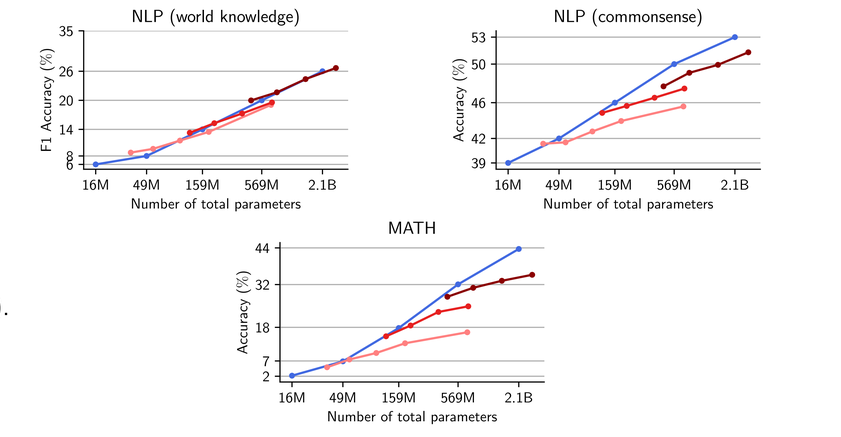
Question. Which Lloyd Webber musical premiered in the US on 10th December 1993?

Answer. Sunset Boulevard

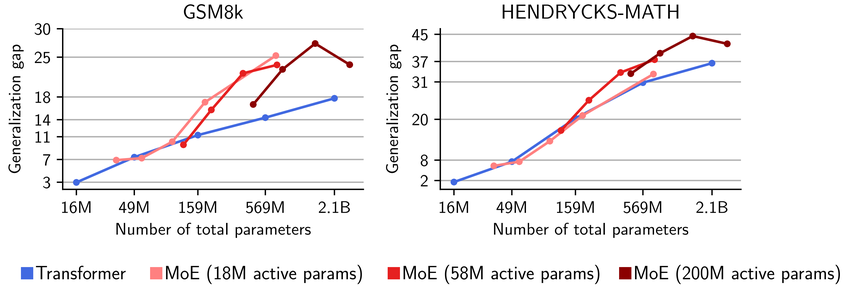
World-knowledge (TriviaQA)

4 - Empirical results

Transformers excel at reasoning, MoEs better at memorization

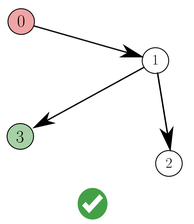


MoEs overfit on reasoning tasks, suggestive of memorization



5 - Main theorems

Synthetic reasoning and memorization tasks



Sequence	Label
aaaa	✗
aaab	✓
aaac	✗
aaad	✓

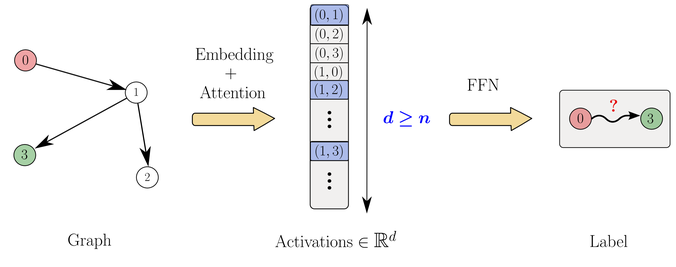
Graph Reasoning: is there a length-2 path between two nodes in a graph with n edges?

Memorization: memorize arbitrary labels for n random Gaussian inputs.

	Graph Reasoning	Memorization
Transformer (width m)	Params: $\Theta(n^2)$ Compute: $\Theta(n^2)$	Params: $\Theta(n)$ Compute: $\Theta(n)$
MoE (K experts, width m)	Params: $\Theta(\boldsymbol{Kn^2})$ Compute: $\Theta(n^2)$	Params: $\widetilde{\Theta}(\boldsymbol{n + Km})$ Compute: $\widetilde{\Theta}(n/K + Km)$

6 - Proof sketch

A critical width is needed for reasoning



MoEs are more efficient for memorization