GLaM: Efficient Scaling of Language Models with Mixture-of-Experts

Paper*: GLaM: Efficient Scaling of Language Models with Mixture-of-Experts

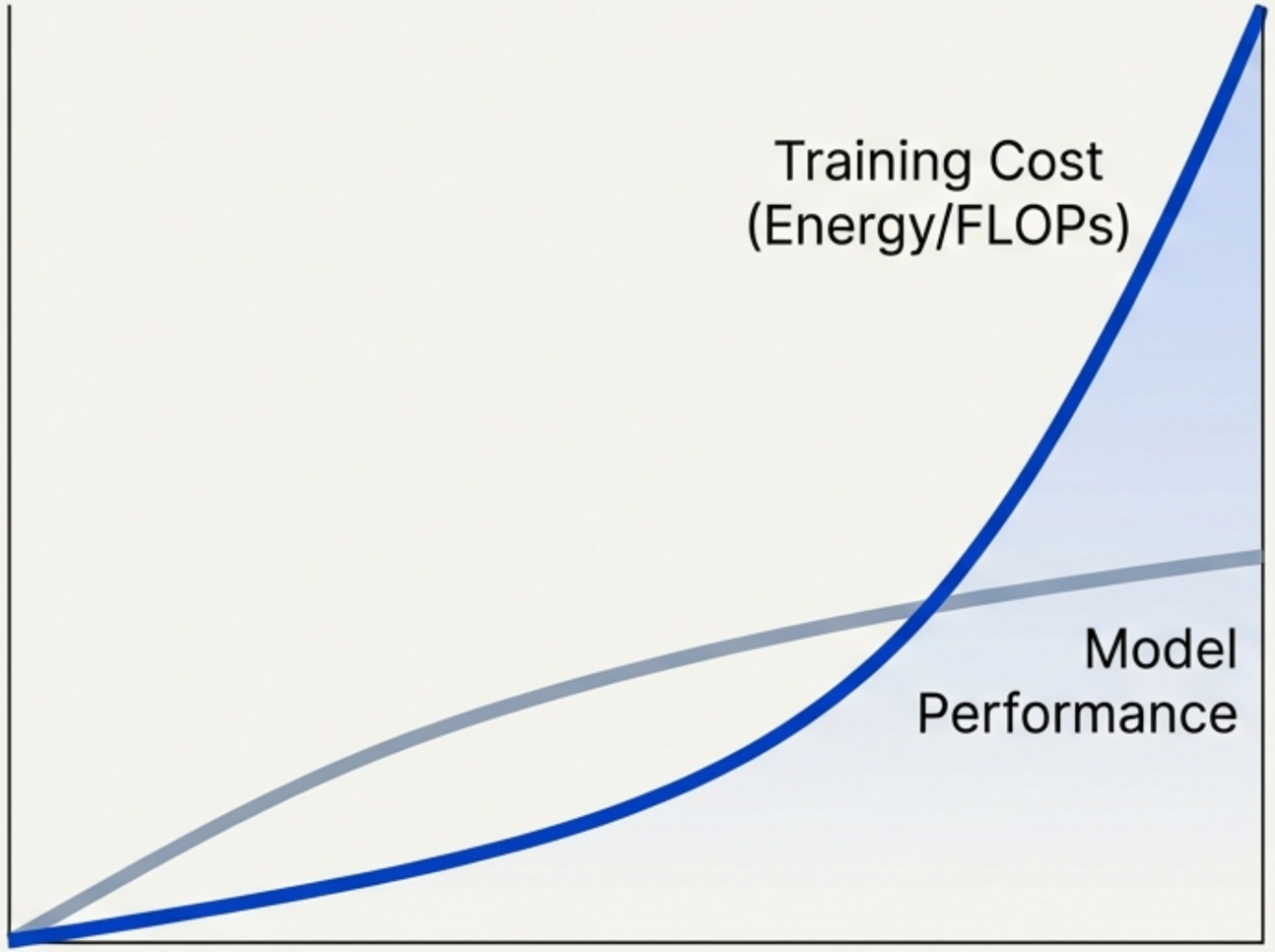
Authors: Nan Du*, Yanping Huang*, Andrew M. Dai*

Authors: Nan Du*, Yanping Huang*, Andrew M. Dai*, et al. (*Equal contribution)

Venue: Proceedings of the 39th International Conference on Machine Learning (ICML 2022)

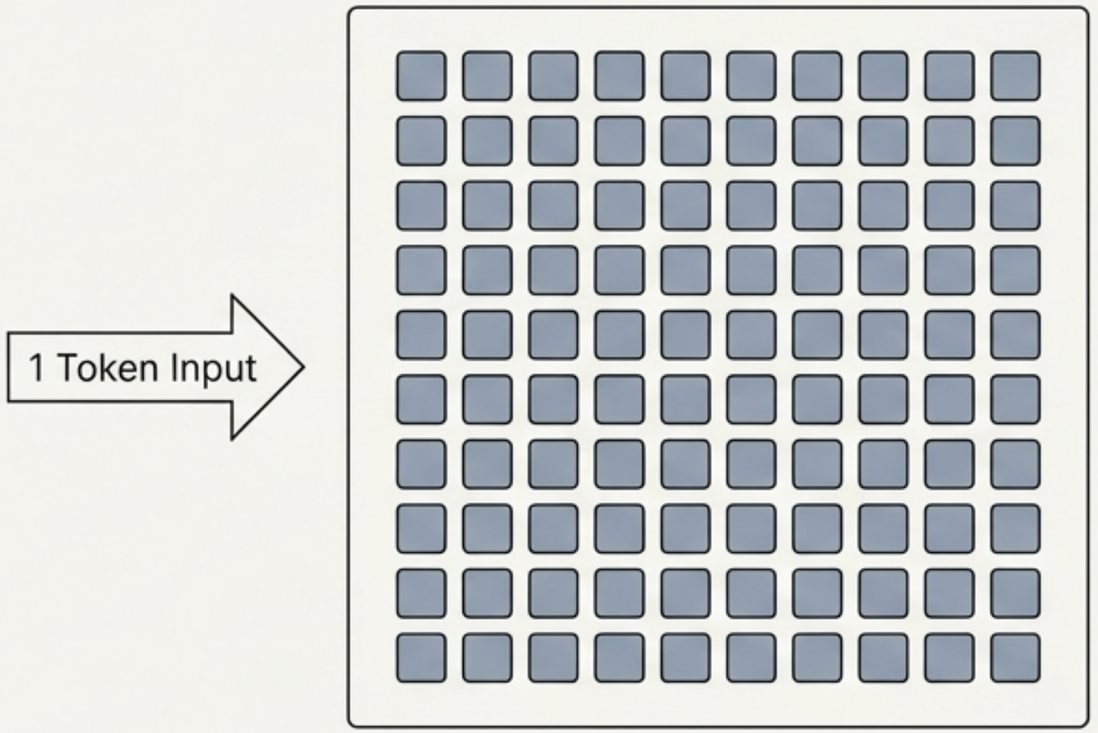
The Unrelenting Cost of Scaling Dense Models

- Scaling models like GPT-3 has driven incredible progress in **few-shot**, in-context learning.
- However, this progress comes at a prohibitively expensive cost in computation and energy.
- **The Dilemma:** Further scaling is becoming infeasible, requiring a more efficient approach to achieve greater capability.



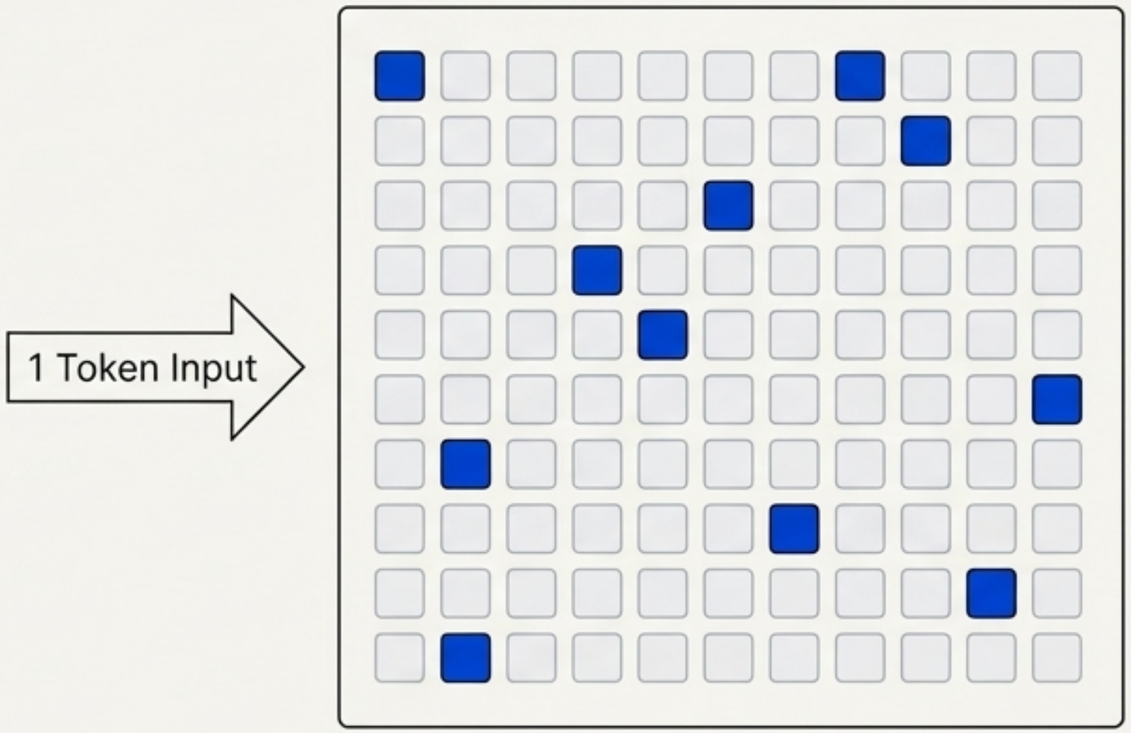
The Inefficiency of Dense Architectures

Dense Model (e.g., GPT-3)



All 175B parameters are activated for every token.

Sparse Activation



Only a fraction of parameters are activated per token.

The Gap: We need to increase a model's total parameter count (its capacity) without proportionally increasing the computational cost per token.

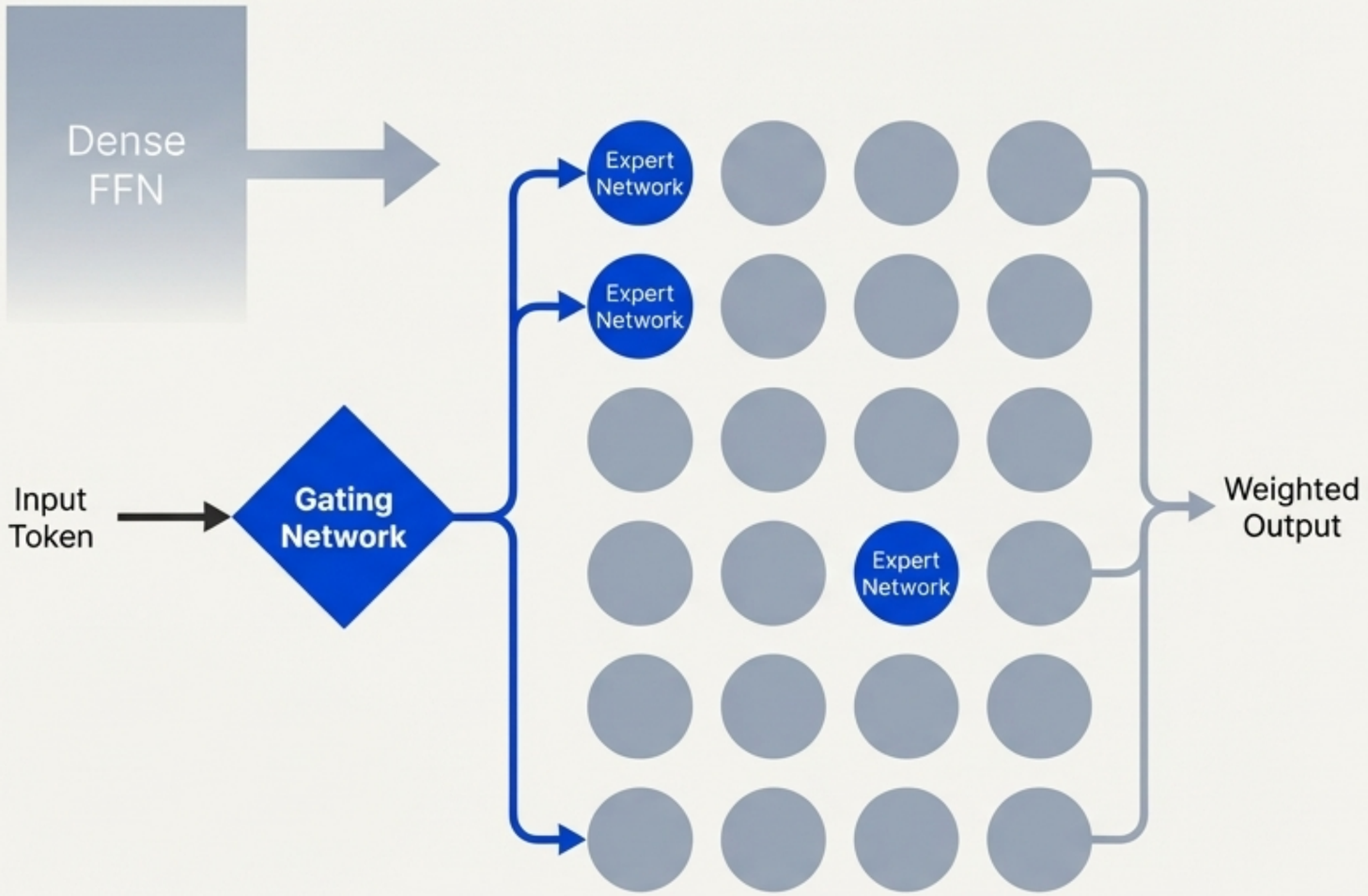
The Solution: Sparsity via Mixture-of-Experts

Core Idea: Replace the dense Feed-Forward Network (FFN) in a Transformer with a Mixture-of-Experts (MoE) layer.

An MoE layer contains many independent “expert” FFNs.

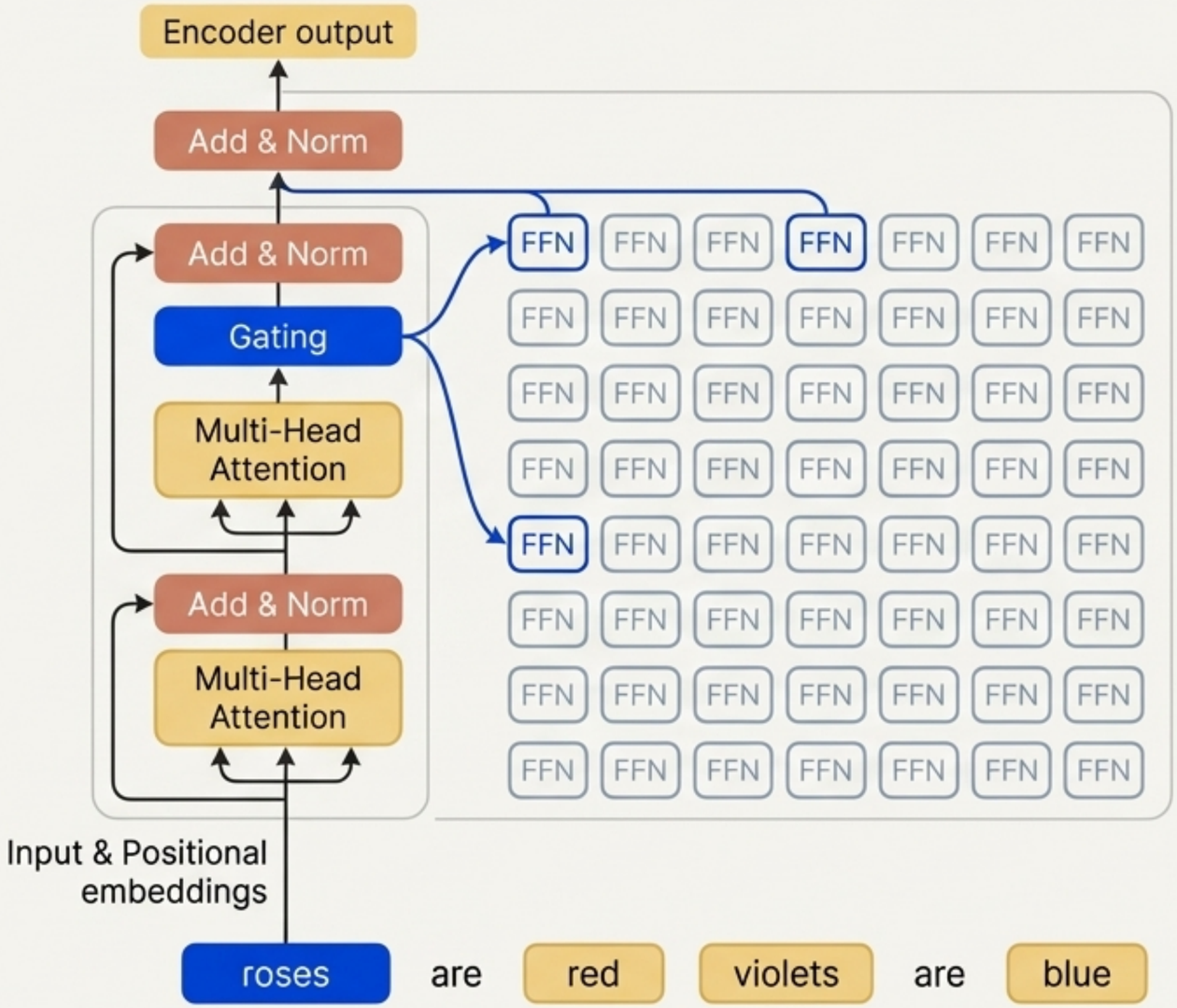
A lightweight “gating network” learns to route each token to the two most relevant experts.

This creates a sparsely activated model: massive total capacity, but low computation per token.



The GLaM Architecture: An MoE Transformer

- GLaM replaces the FFN of every other Transformer layer with an MoE layer.
- For each input token (e.g., 'roses'), the gating network dynamically selects the best 2 out of 64 available experts.
- This provides over 2000 potential FFN combinations per MoE layer (`O(E^2)` flexibility).
- The final token representation is the weighted combination of the two expert outputs.



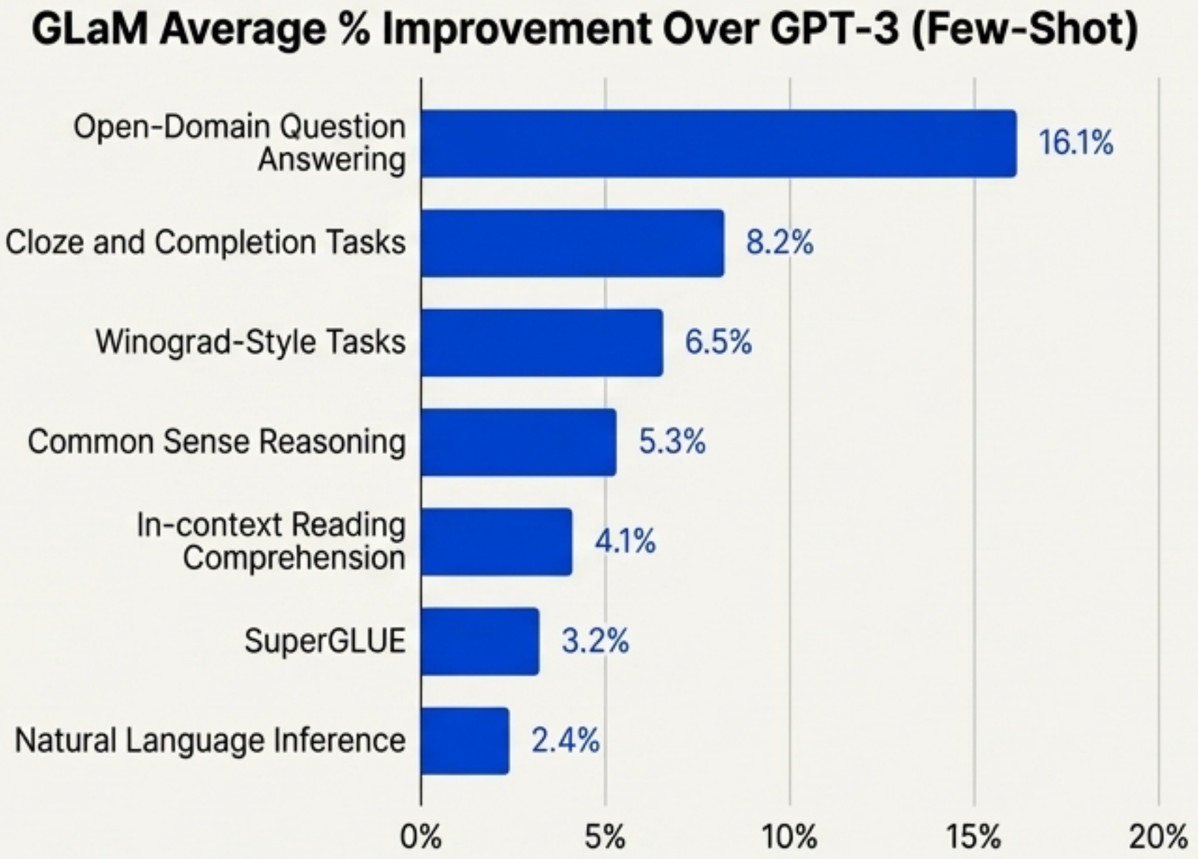
GLaM Outperforms GPT-3 at a Fraction of the Cost

	Metric	GLaM (1.2T)	GPT-3 (175B)	Improvement
	**Training Energy**	**456 MWh**	1287 MWh	**-65%**
	**Inference FLOPs**	**180 G**	350 G	**-49%**
	**Few-Shot Accuracy**	**68.1**	65.2	**+4.4%**

GLaM activates only 96.6B parameters per token—far more efficient despite its 7x larger total size.

Performance Gains are Consistent Across Task Categories

- GLaM outperforms GPT-3 in 6 out of 7 major benchmark categories on average.
- Gains are shown across zero, one, and few-shot settings.
- Performance is particularly strong on knowledge-intensive tasks like Open-Domain Question Answering.



Massive Sparse Capacity Unlocks New Knowledge

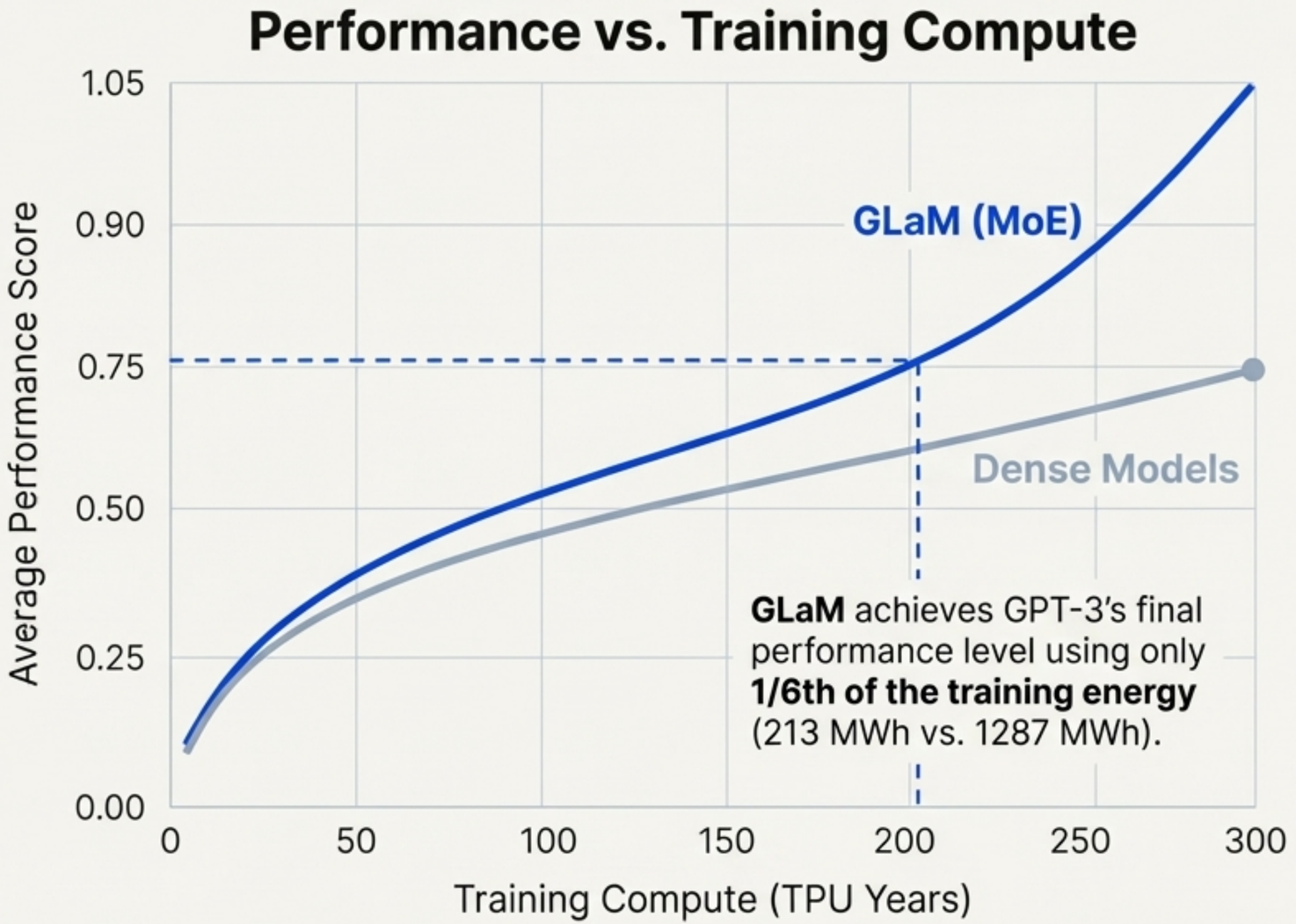
- On the challenging TriviaQA benchmark, model capacity is crucial for storing world knowledge.

TriviaQA (Open-Domain)
GLaM's **one-shot** result (75.8 dev) outperforms the previous **fine-tuned** state-of-the-art model (69.8 test).

- This suggests GLaM's 1.2T parameters, though sparsely activated, provide a significant advantage in knowledge storage and retrieval over smaller dense models.

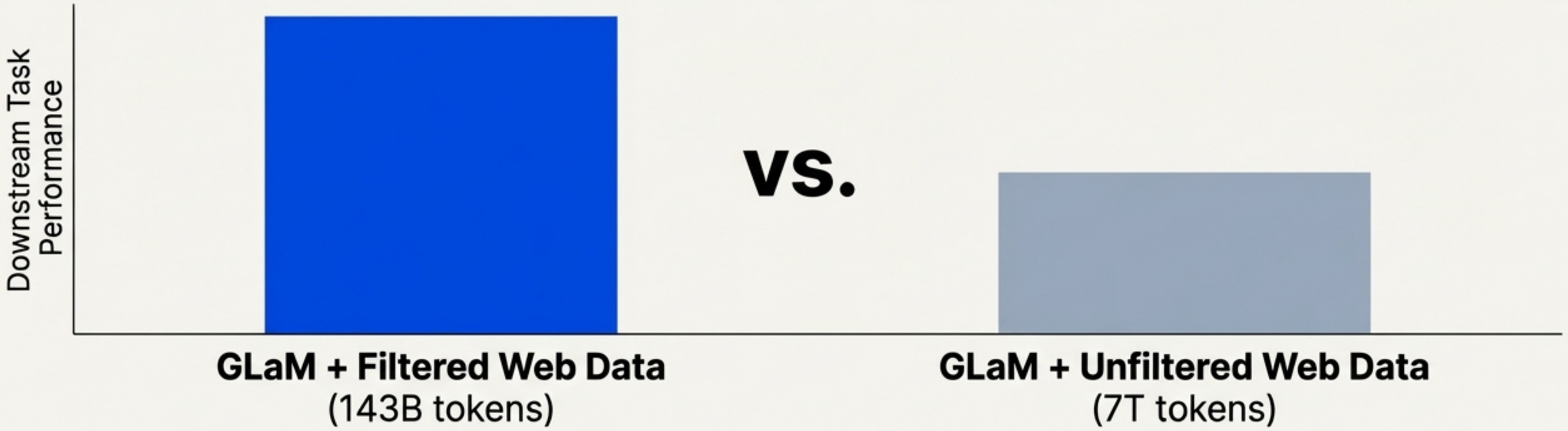
Sparse Models Learn More Efficiently

- For the same amount of training compute, MoE models consistently achieve higher performance.
- The performance gap widens at larger scales, suggesting sparsity is the superior path for future scaling.



Insight: Data Quality is Paramount, Even at Scale

- The authors curated a high-quality 1.6T token dataset, a key component of which was filtering low-quality web pages.
- **Finding**: A model trained on the smaller, high-quality filtered dataset consistently outperformed one trained on a much larger, unfiltered dataset.



Takeaway: Data quality should not be sacrificed for quantity, as it is a critical driver of performance, especially for generation.

A Promising Step Towards Fairer Models

Stereotypical Examples
Accuracy: **71.7%**

Anti-stereotypical Examples
Accuracy: **71.7%**

- GLaM's performance was evaluated on the WinoGender benchmark, which measures gender bias.
- It achieves a new state-of-the-art accuracy of 71.7%.
- **Crucially:** GLaM **is the first model to close the performance gap, achieving identical accuracy on both stereotypical and anti-stereotypical ('gotcha') examples.**
- This suggests large, sparse models may rely less on superficial statistical correlations.

The Road Ahead: Challenges and Open Questions

- **Serving Cost**

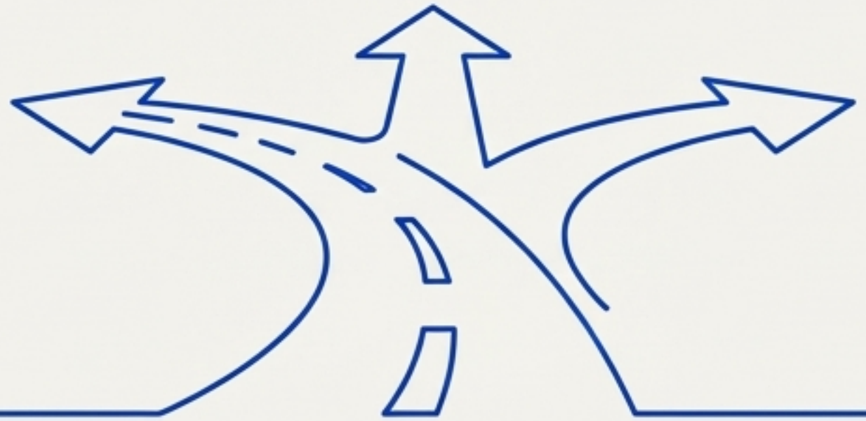
While inference is compute-efficient, the large total parameter count (1.2T) requires significant memory, making deployment complex and potentially costly.

- **Expert Specialization**

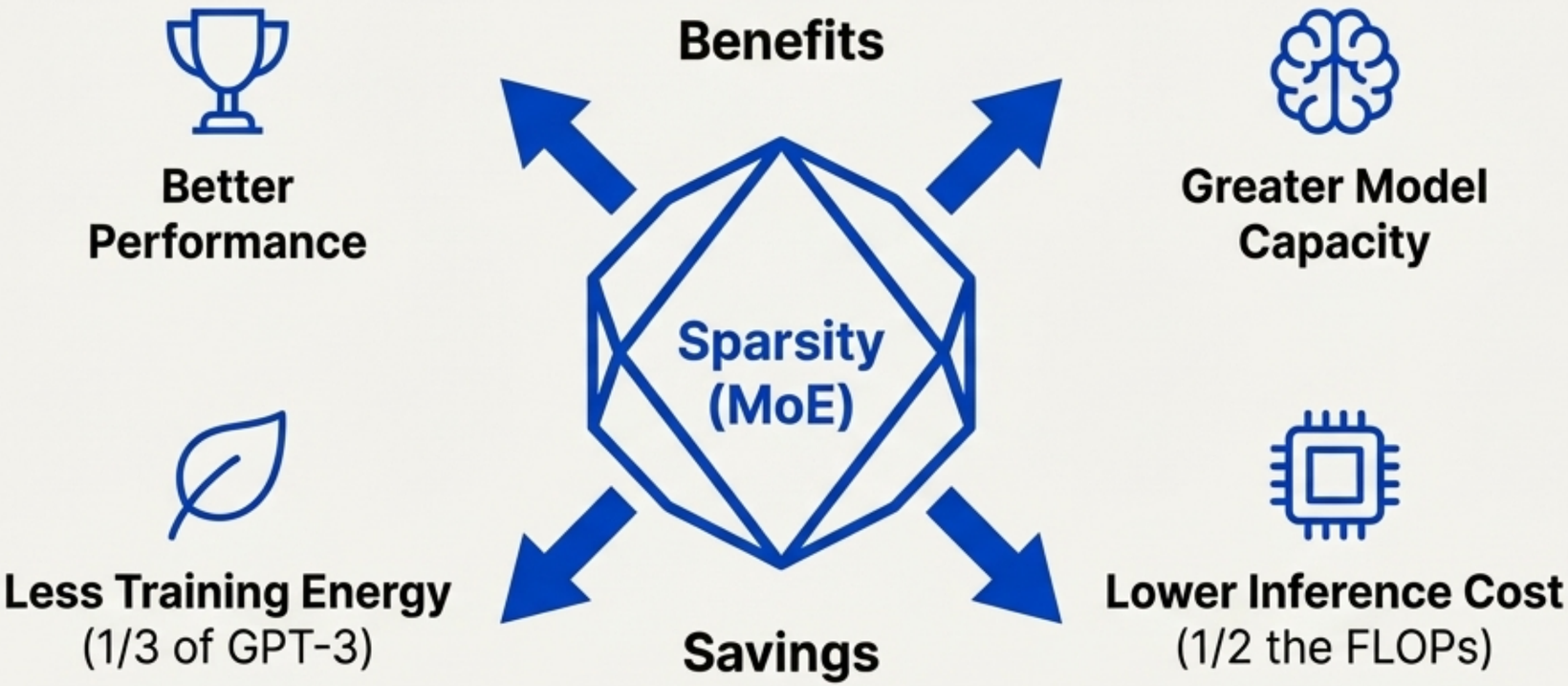
How do experts specialize? Understanding and controlling this process is a key area for future research.

- **Optimal Routing**

Can the gating network be improved? Exploring more sophisticated routing strategies is an open question.



A New Paradigm: Scaling Smarter, Not Just Bigger



- GLaM demonstrates that sparsely activated models achieve **better performance** than state-of-the-art dense models.
- ...while being **dramatically more efficient** in both training and inference.
- Sparsity is a proven, promising direction for building the next generation of powerful and sustainable AI.

Thank You

Read the Paper: https://arxiv.org/abs/2112.06905

Contact: dunan@google.com, huangyp@google.com, adai@google.com

Questions?