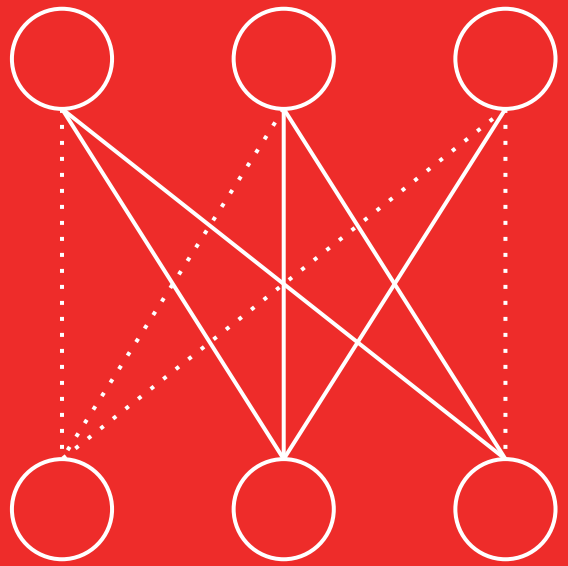
Democratizing AI:
Are we in the Vacuum Tube Era of AI?



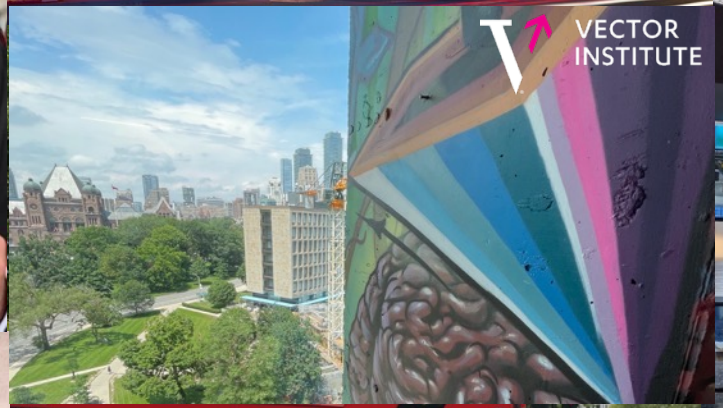
Yani Ioannou
Associate Professor, Schulich Research Chair
Dept. of Electrical & Software Engineering,
Schulich School of Engineering, University of Calgary

Sungkyunkwan University
July 15th, 2026

UNIVERSITY OF CALGARY

UNIVERSITY OF
CAMBRIDGE

Microsoft


VECTOR
INSTITUTE

NASA

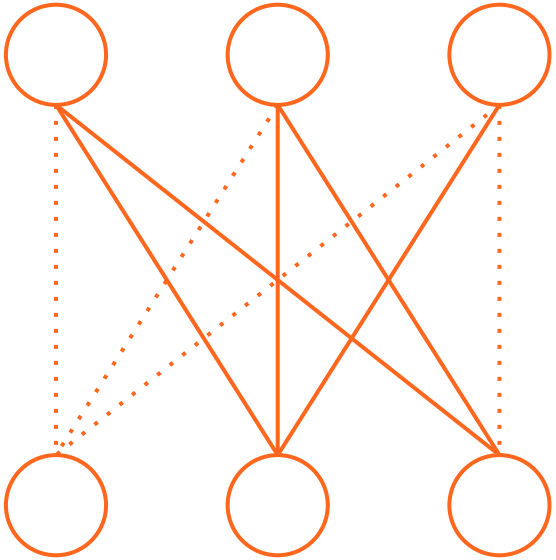
GY69 UES
WAYVE

Google

Calgary

Banff

Democratizing AI: Are we in the Vacuum Tube Era of AI?

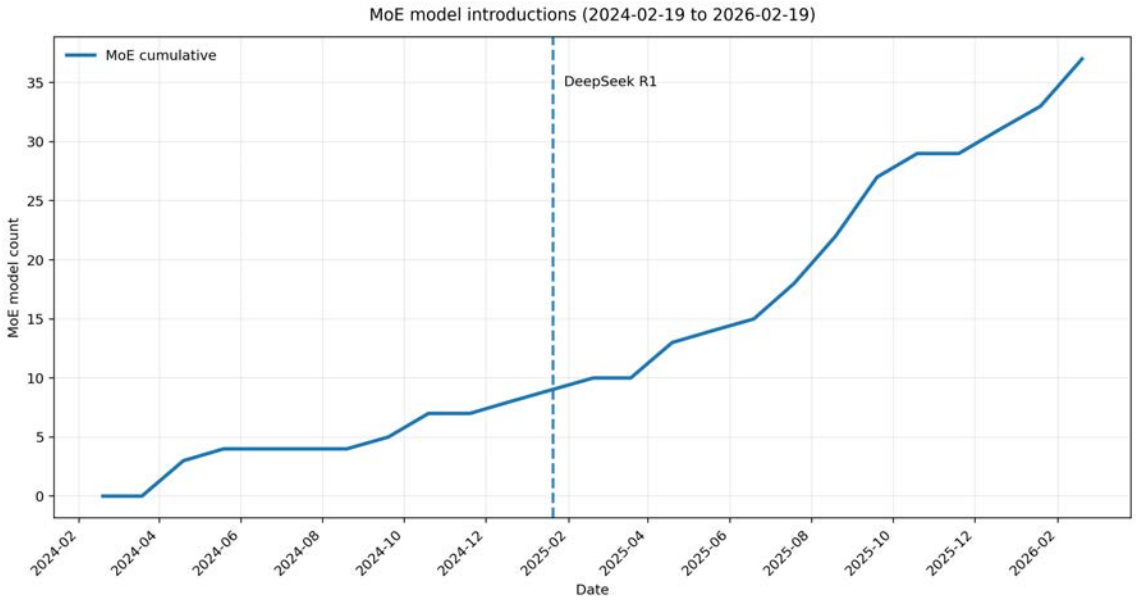


1. Mixture of Expert (MoE) Models

2. Expert Compression and REAP
3. The AI hobbyist community
4. An analogy in computing history
5. Overparameterization in NNs

Mixture of Expert (MoE) Models

- State-of-the-art LLMs and VLMs are increasingly **Sparse Mixture-of-Expert (MoE)** models
- Popular open-weight models include DeepSeek-R1,V2,V3, Qwen 3.5, GLM-5, Kimi K2.5, Qwen3-480B-Coder, ...



huggingface.co/blog/moe-transformers

Learning Factored Representations in a Deep Mixture of Experts. David Eigen, Marc'Aurelio Ranzato, Ilya Sutskever. ICLR Workshops, 2014.
Decision Forests, Convolutional Networks and the Models in-Between. Yani Ioannou, Duncan Robertson, Darko Zikic, Peter Kontschieder, Jamie Shotton, Matthew Brown, Antonio Criminisi. Microsoft Research Technical Report, 2016.

Mixture of Expert (MoE) Models

- A MoE replaces a single layer with a layer composed of a **router** and **multiple experts**
- In Transformers, every feed-forward network (FFN) layer is replaced by MoE layer

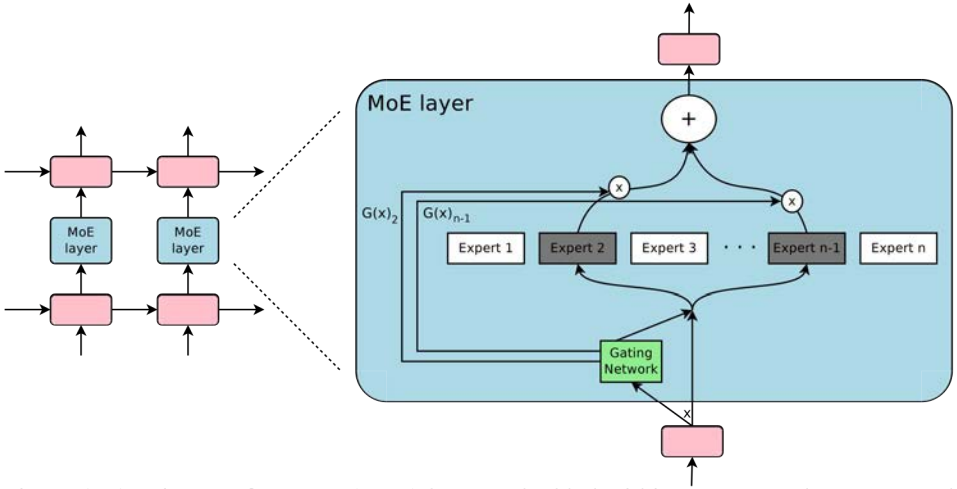
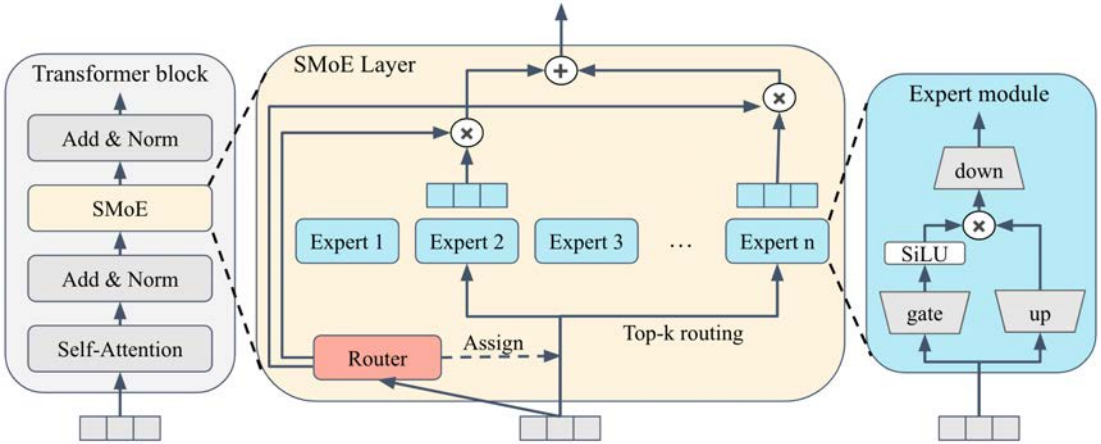


Figure 1: A Mixture of Experts (MoE) layer embedded within a recurrent language model. In this case, the sparse gating function selects two experts to perform computations. Their outputs are modulated by the outputs of the gating network.

Learning Factored Representations in a Deep Mixture of Experts. *David Eigen, Marc'Aurelio Ranzato, Ilya Sutskever.* ICLR Workshops, 2014.
Decision Forests, Convolutional Networks and the Models in-Between. *Yani Ioannou, Duncan Robertson, Darko Zikic, Peter Kontschieder, Jamie Shotton, Matthew Brown, Antonio Criminisi.* Microsoft Research Technical Report, 2016.
Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, Jeff Dean.* ICLR 2017 (figure 1)

Sparse Mixture-of-Experts (SMoE)

- Sparse MoEs Route to a **small number** of experts per token (e.g., top-k=8 with k=128)
- Balances **high model capacity** with **low active parameters**

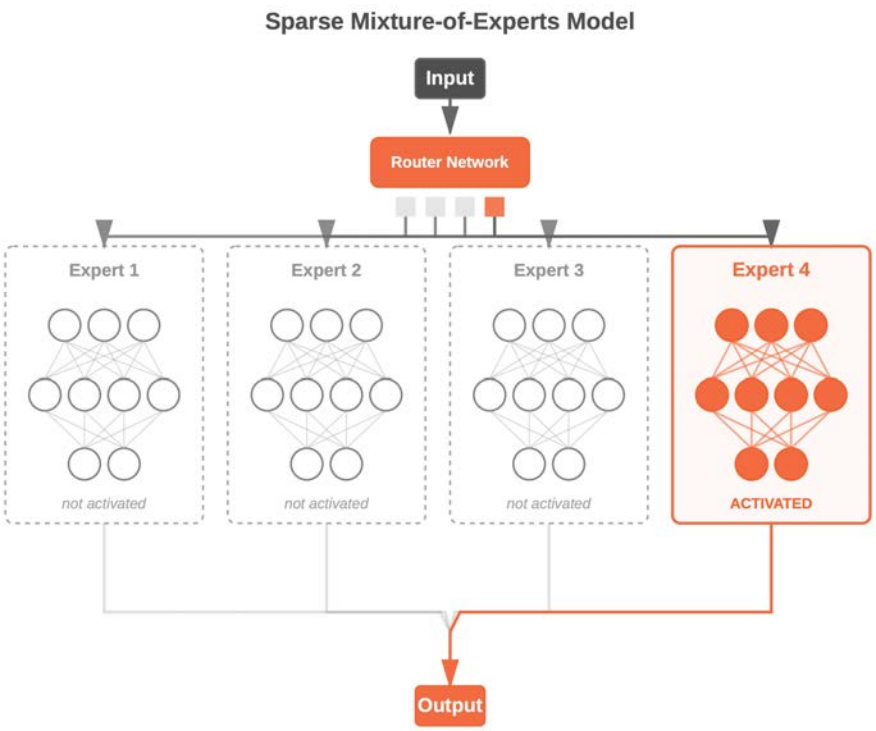


I.-C. Chen, H.-S. Liu, W.-F. Sun, C.-H. Chao, Y.-C. Hsu, and C.-Y. Lee, "Retraining-free Merging of Sparse MoE via Hierarchical Clustering," presented at the Forty-second International Conference on Machine Learning, Jun. 2025.

Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, Jeff Dean. ICLR 2017
Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. William Fedus, Barret Zoph, Noam Shazeer. JMLR 23: 120:1-120:39 (2022)

Sparse Mixture-of-Experts: Pros

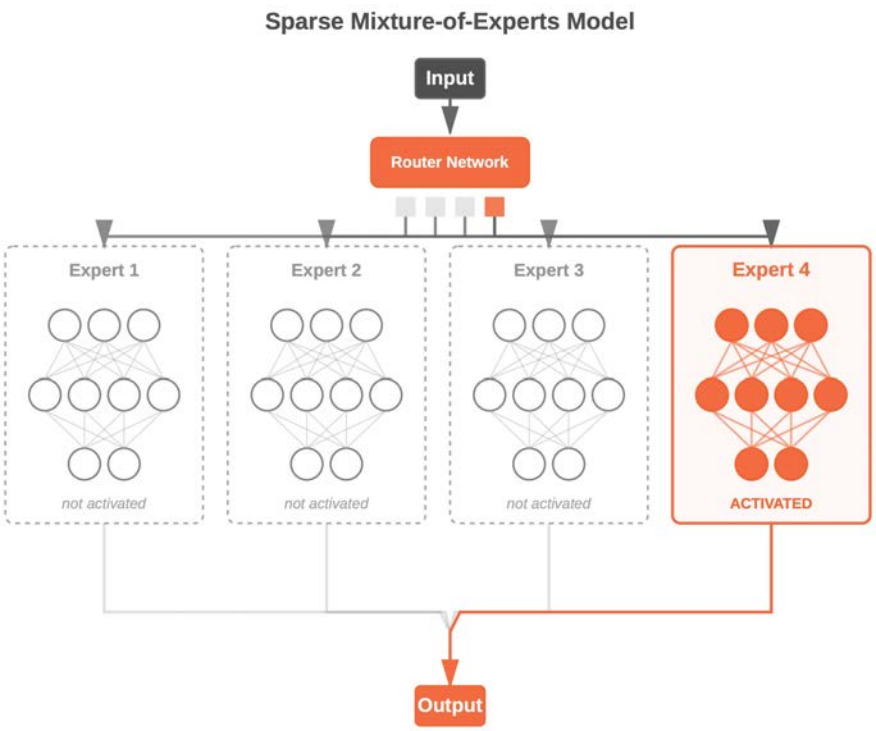
- Each token only routed to subset of experts
- Expert parallelism
- **Higher pre-training efficiency**
- **Lower latency at inference**



REAP: One-Shot Pruning for Trillion-Parameter Mixture-of-Experts Models. www.cerebras.ai/blog/reap (figure)

Sparse Mixture-of-Experts: Cons

- Routers can be difficult to train
- Load-balancing experts at inference is tricky
- **All weights have to be stored in memory at inference**

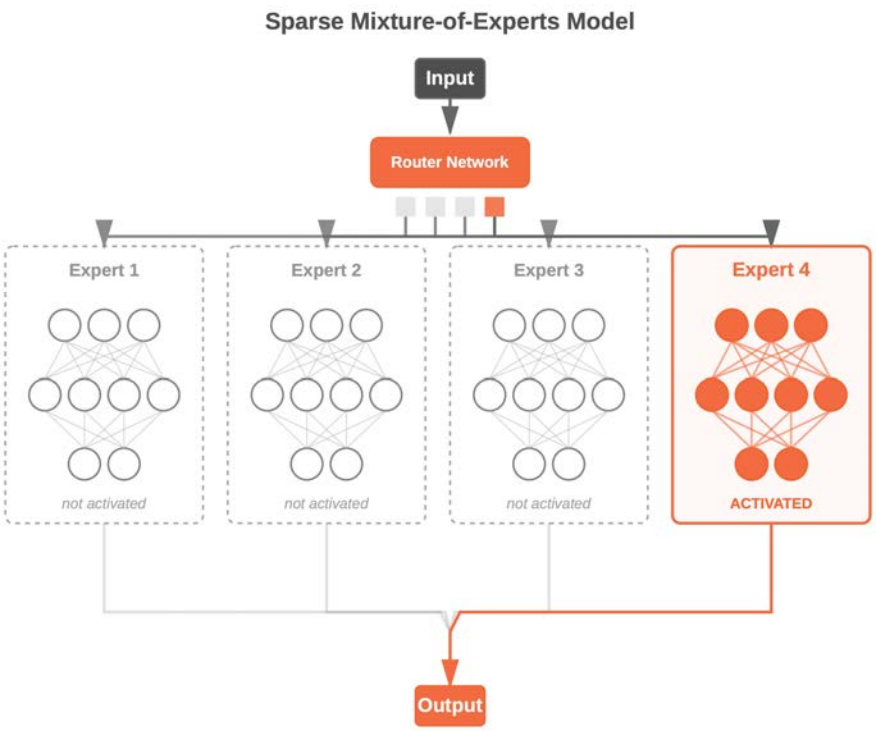


REAP: One-Shot Pruning for Trillion-Parameter Mixture-of-Experts Models. www.cerebras.ai/blog/reap (figure)

Sparse Mixture-of-Experts: Cons

- Routers can be difficult to train
- Load-balancing experts at inference is tricky
- All weights have to be stored in memory at inference

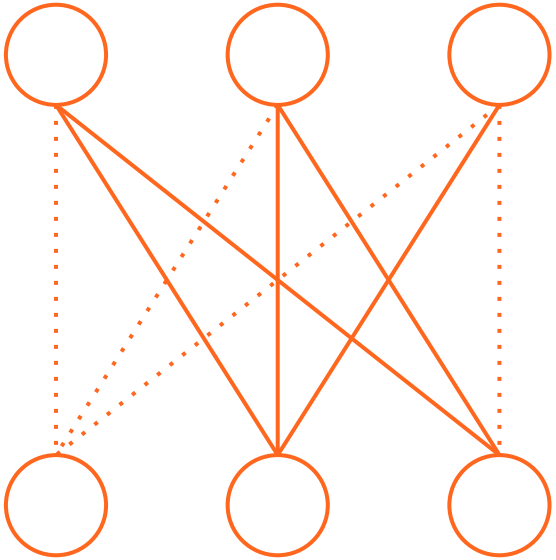
Massive VRAM requirements!



REAP: One-Shot Pruning for Trillion-Parameter Mixture-of-Experts Models. www.cerebras.ai/blog/reap (figure)

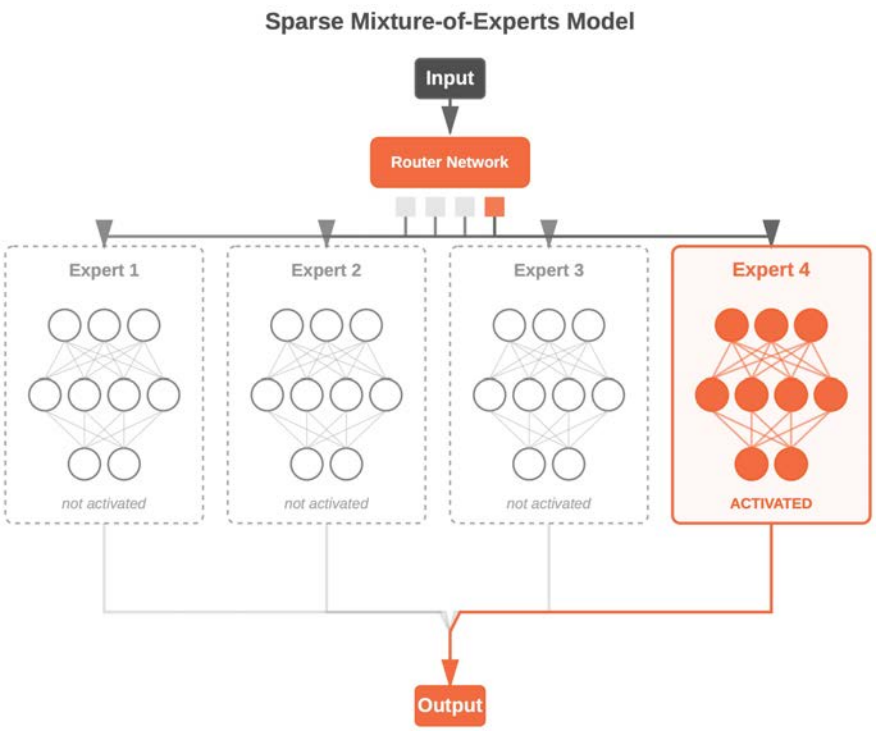
Democratizing AI: Are we in the Vacuum Tube Era of AI?

1. Mixture of Expert (MoE) Models
2. **Expert Compression and REAP**
3. The AI hobbyist community
4. An analogy in computing history
5. Overparameterization in NNs



SMoE: Expert Compression

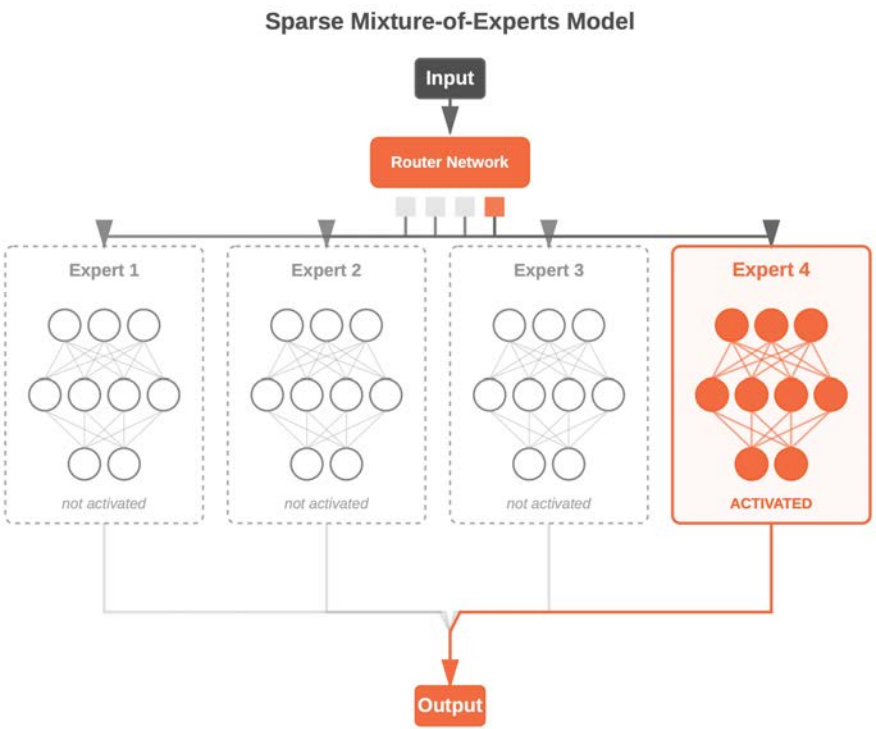
- How likely is prompting on poetry, coding and gardening in a single session?
- Model use cases are typically specialized, i.e., it's **task-specific**
- Can we store **fewer experts at inference for specialized usage to minimize VRAM usage?**



REAP: One-Shot Pruning for Trillion-Parameter Mixture-of-Experts Models. www.cerebras.ai/blog/reap (figure)

SMoE: Expert Pruning

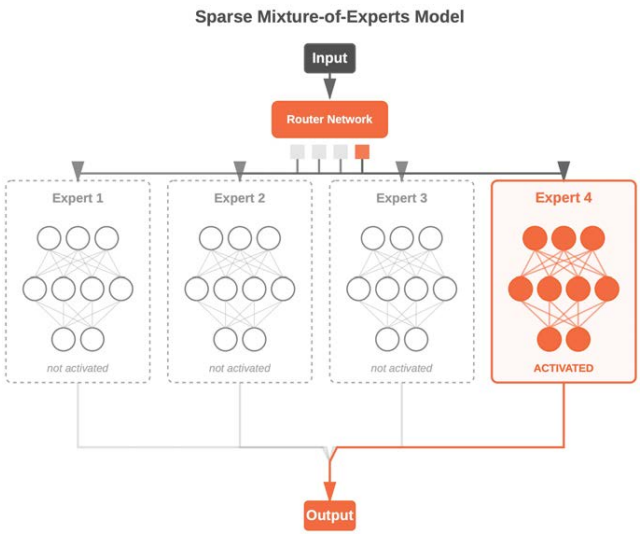
- Early work focused on **pruning experts** progressively during finetuning using router weights[1]
- Global one-shot pruning without finetuning works and is simpler[2]



1. **Task-Specific Expert Pruning for Sparse Mixture-of-Experts.** T. Chen, S. Huang, Y. Xie, B. Jiao, D. Jiang, H. Zhou, J. Li, and F. Wei. arXiv:2206.00277, 2022.
2. **Memory-efficient NLLB-200: Language-specific Expert Pruning of a Massively Multilingual Machine Translation Model.** Y. Koishekenov, A. Berard, and Vassilina Nikoulina. ACL 2023.

SMoE: Expert Merging

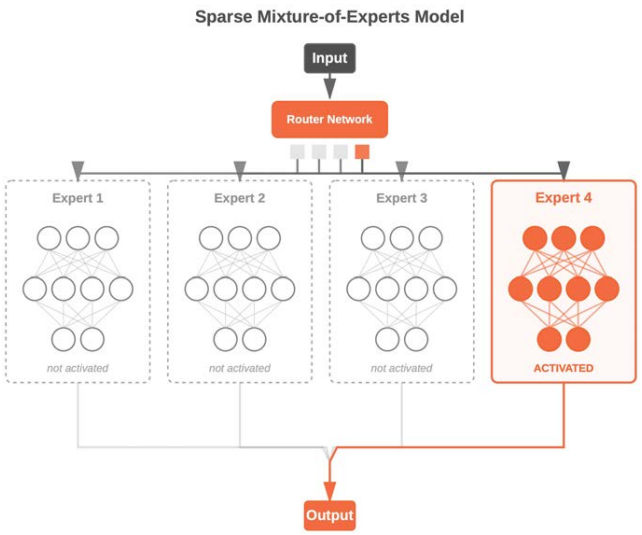
- Alternative approach is to **merge experts**: requires **clustering** and **merging** method



1. **Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy.** P. Li, Z. Zhang, P. Yadav, Y. Sung, Y. Cheng, M. Bansal, and T. Chen. ICLR 2023.
2. **Retraining-free merging of sparse moe via hierarchical clustering**. I. Chen, H. Liu, W. Sun, C. Chao, Y. Hsu, and C. Lee. ICML 2025

SMoE: Expert Merging

- Alternative approach is to **merge experts**: requires **clustering** and **merging** method
- **M-SMoE**[1] clusters experts by global activation frequency and cosine similarity of router logits
- **HC-SMoE**[2] uses hierarchical agglomerative clustering based on average activation of each expert over all tokens in a calibration set.
- Both generally aligns layers by permutation matching and merge based on frequency-weighted parameter averaging



1. **Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy.** P. Li, Z. Zhang, P. Yadav, Y. Sung, Y. Cheng, M. Bansal, and T. Chen. ICLR 2023.
2. **Retraining-free merging of sparse moe via hierarchical clustering**. I. Chen, H. Liu, W. Sun, C. Chao, Y. Hsu, and C. Lee. ICML 2025

Published as a conference paper at ICLR 2026

REAP THE EXPERTS: WHY PRUNING PREVAILS FOR ONE-SHOT MoE COMPRESSION

Mike Lasby[1,2,†], **Ivan Lazarevich**[1], **Nish Sinnadurai**[1], **Sean Lie**[1],
Yani Ioannou[2], **Vithursan Thangarasa**[1]
[1]Cerebras Systems Inc., [2]Schulich School of Engineering, University of Calgary

https://github.com/CerebrasResearch/reap
https://hf.co/collections/cerebras/cerebras-reap

ABSTRACT

Sparsely-activated Mixture-of-Experts (SMoE) models offer efficient pre-training and low latency but their large parameter counts create significant memory overhead, motivating research into expert compression. Contrary to recent findings favouring expert *merging* on discriminative benchmarks, we find that expert *pruning* is a superior strategy for generative tasks. We demonstrate that existing merging techniques introduce an irreducible error due to the loss of fine-grained routing control over experts. Leveraging this insight, we propose Router-weighted Expert Activation Pruning (REAP), a novel pruning criterion that considers both router gate-values and expert activation norms to minimize the reconstruction error bound. Across a diverse set of SMoE models ranging from 20B to 1T parameters, REAP consistently outperforms merging and other pruning methods on generative benchmarks, especially at 50% compression. Notably, our method achieves near-lossless compression on code generation tasks with Qwen3-Coder-480B and Kimi-K2, even after pruning 50% of experts.

Mike Lasby

Ivan Lazarevich

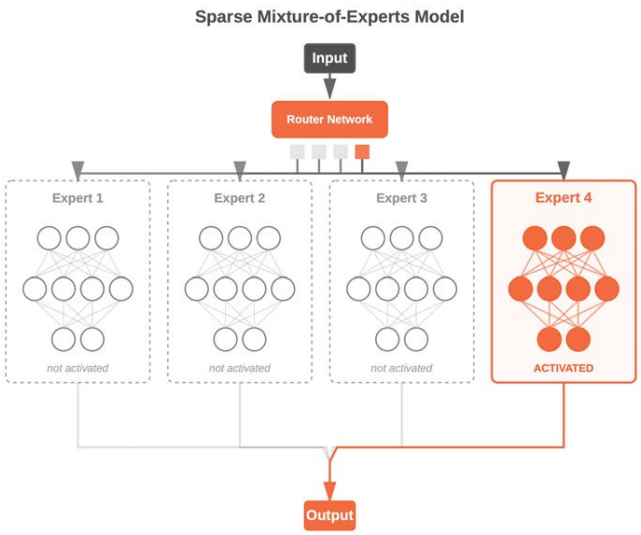
Nish Sinnadurai

Sean Lie

Vithursan Thangarasa

SMoE: Expert Merging and Generative Tasks

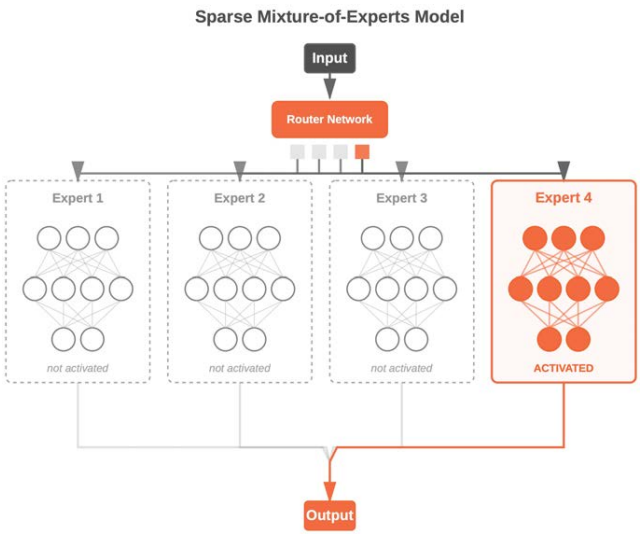
- Expert merging was shown to surpass the performance of expert pruning on **discriminative tasks** like multiple-choice question answering
- **However, our evaluations on generative tasks show a significant loss of performance with merging…**



1. **Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy.** P. Li, Z. Zhang, P. Yadav, Y. Sung, Y. Cheng, M. Bansal, and T. Chen. ICLR 2023.
2. **Retraining-free merging of sparse moe via hierarchical clustering**. I. Chen, H. Liu, W. Sun, C. Chao, Y. Hsu, and C. Lee. ICML 2025

SMoE: Expert Merging and Generative Tasks

- Intuitively this is based on the **loss of independence** of the two merged experts
- **Pruning experts does not result in a loss of expert independence**



1. **Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy.** P. Li, Z. Zhang, P. Yadav, Y. Sung, Y. Cheng, M. Bansal, and T. Chen. ICLR 2023.
2. **Retraining-free merging of sparse moe via hierarchical clustering**. I. Chen, H. Liu, W. Sun, C. Chao, Y. Hsu, and C. Lee. ICML 2025

Router-weighted Expert Activation Pruning (REAP)

- We proposed a simple new one-shot expert pruning saliency criteria:
- **Router-weighted Expert Activation Pruning (REAP)**
- Existing frequency-based pruning methods effectively assume that all active experts contribute equally to the output
- REAP considers both the expert's relative contribution and its frequency of activation by using **both the router weight and expert activation norm** to decide which experts are most important

REAP expert saliency score:

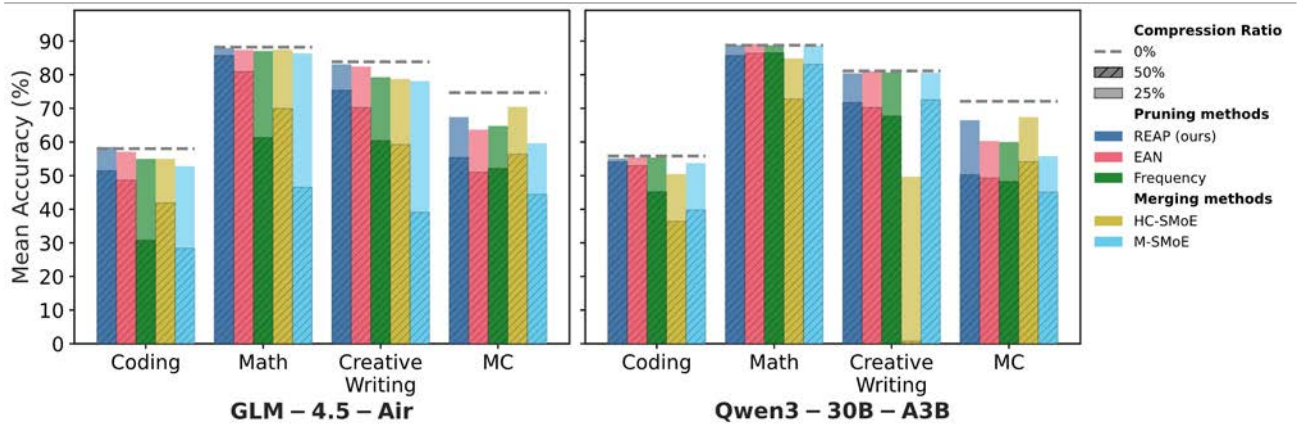
$$S_j = \frac{1}{|\mathcal{X}_j|} \sum_{x \in \mathcal{X}_j} g_j(x) \cdot \left\| f_j(x) \right\|_2$$

set of tokens where expert j is active

router weighting

expert activation

Results



GLM-4.5-Air and Qwen3-30B accuracy vs. task type.

- Evaluate with MoEs ranging from **20B to 1T** parameters
- REAP demonstrates **consistently higher accuracy on generative task-specific workloads**
- Combining REAP and quantization enables scaling to **1T parameters**

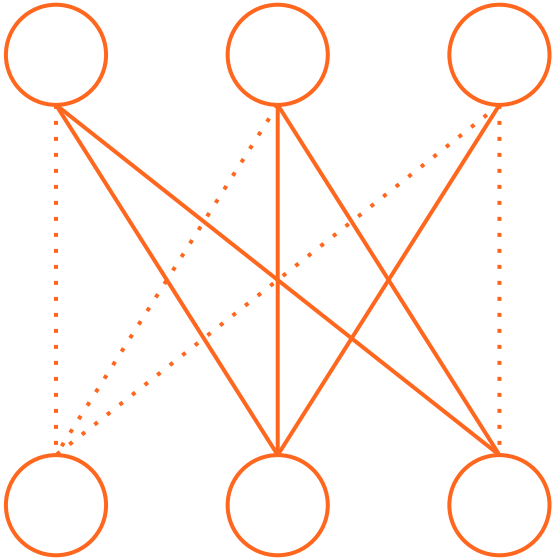
REAP Results

- REAP is **composable with quantization!**
- Enables scaling to frontier model size on a single node

Model	Compression	Method	Non-Agentic Coding			Agentic Coding	Tool-Use (BFCLv3)				MC
			Eval+	LiveCode	Code Avg	SWE-Bench-Verified	Non-Live	Live	Multi-Turn	Overall	MC Avg
Qwen3-Coder-480B-A35B-Instruct-FP8	**Baseline**		0.841	0.431	0.636	0.540	0.866	0.825	0.380	0.690	0.750
	25%	Frequency	0.737	0.296	0.516	0.378	0.844	0.763	0.355	0.654	0.606
		EAN	0.827	0.419	0.623	0.534	0.831	0.813	0.384	0.676	0.702
		REAP	0.831	0.416	**0.624**	**0.540**	0.878	0.823	0.392	**0.698**	**0.748**
	50%	Frequency	0.007	0.012	0.010	0.000	0.200	0.392	0.000	0.197	0.506
		EAN	0.777	0.382	0.580	**0.536**	0.822	0.774	0.383	0.659	0.591
		REAP	0.822	0.415	**0.619**	0.522	0.849	0.801	0.371	**0.674**	**0.692**
Kimi-K2-Instruct-W4A16	**Baseline**		0.828	0.434	0.631	0.554	0.840	0.802	0.355	0.666	0.780
	25%	Frequency	0.486	0.082	0.284	0.000	0.644	0.603	0.045	0.431	0.604
		EAN	0.779	0.379	0.579	0.562	0.819	0.802	0.335	**0.652**	0.703
		REAP	0.840	0.440	**0.640**	**0.580**	0.842	0.801	0.263	0.635	**0.773**
	50%	Frequency	0.112	0.000	0.056	0.000	0.255	0.397	0.003	0.218	0.439
		EAN	0.722	0.253	0.487	**0.576**	0.778	0.767	0.173	**0.573**	0.587
		REAP	0.819	0.429	**0.624**	**0.576**	0.785	0.743	0.164	0.564	**0.643**

Combined **87.5% compression rate!**

Democratizing AI: Are we in the Vacuum Tube Era of AI?



1. Mixture of Expert (MoE) Models
2. Expert Compression and REAP
3. **The AI hobbyist community**
4. An analogy in computing history
5. Overparameterization in NNs

The Surprising Impact of REAP

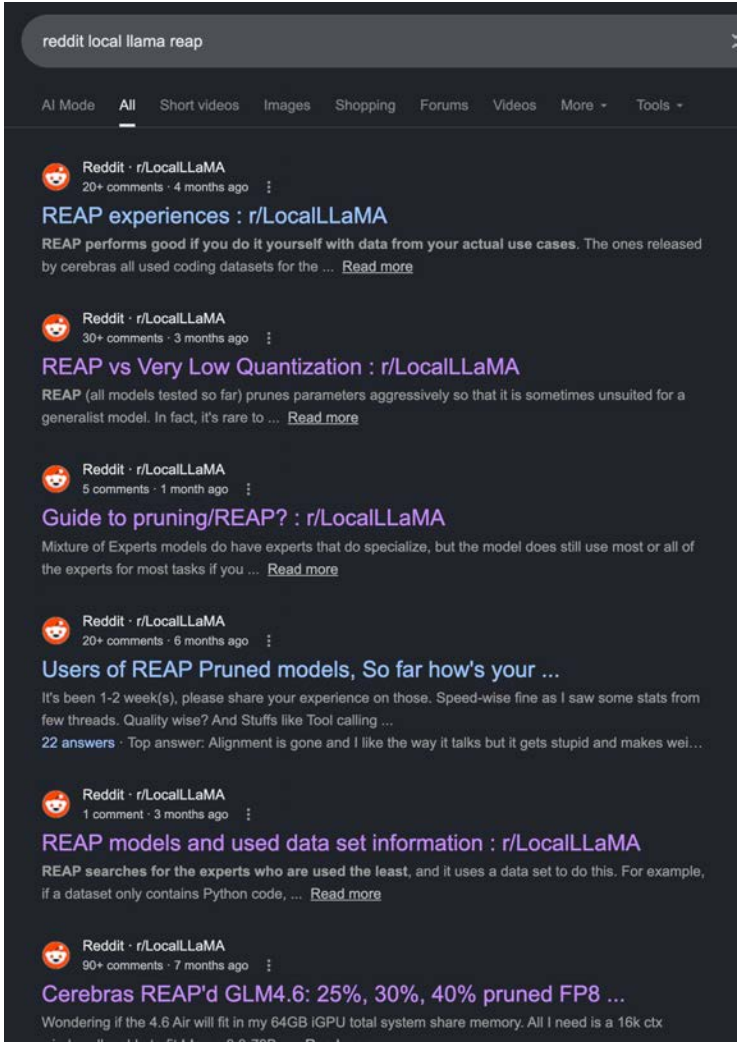
- **REAP saw rapid impact and adoption!**
- **Just not quite in the community I expected...**

The Surprising Impact of REAP

- **REAP saw rapid impact and adoption!**
- **Just not quite in the community I expected...**

The LocalLLM Community

- REAP saw rapid impact and adoption on Reddit, HuggingFace, X, etc.
- Specifically, on Reddit the **r/LocalLLM** and **r/LocalLLaMA**, channels dedicated to running LLMs locally on commodity hardware



Reception

- REAP models are domain specialized
- Empower users by “BYO-Data”
- Poor performance if you step outside calibration distribution
- Overall real-world feedback validated our academic evaluation!

VectorD · 4mo ago

REAP performs good if you do it yourself with data from your actual use cases. The ones released by cerebras all used coding datasets for the reaping.

6 Reply Award Share ···

SidneyFong · 4mo ago

From my limited but telling experience, I think this is literally LLM lobotomy.

I tried MiniMax-M2.1-REAP-139B-A10B. Its English responses seem fine. But when given Chinese prompts and when it tried to output in Chinese, it just kept devolving into repeating the same tokens.

thebadslime · 4mo ago

I like reap, I REAPed and use ERNIE21B for a project I am working on (finetune) and it perfomrs as well as the original model. I did CPT after and it works better than before, I did see some benchmark losses following the prune, but CPT got it close to baseline on most scores.

3 Reply Award Share ···

0xSero @0xSero · 10h

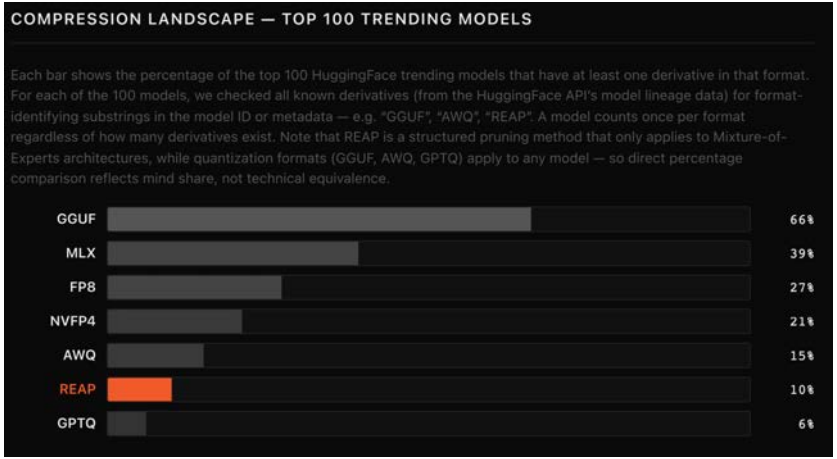
Spark owners, rejoice:

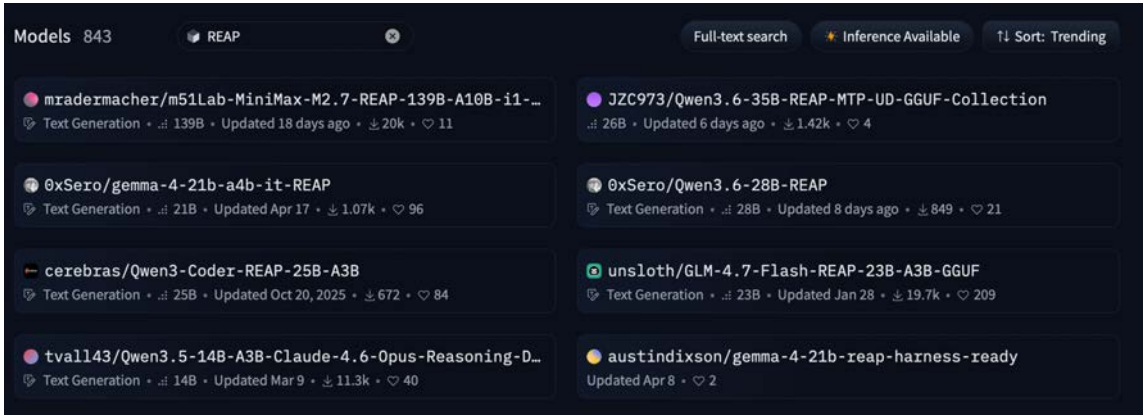
Deepseek-v4-Flash-200B **REAP** is looking good, no issues with agentic/coding/if look at it's output

- 300k context (fp8 kv cache)
- fp8+fp4 weights 100GB
- Logical and functional
- Cost me 1k usd

HuggingFace: Impact

- As of March 2026, over **600k downloads** of REAP tagged models on HuggingFace
- There are almost **900 REAP models** on HF
- REAP is the only non-quantization compression method in **top 10 on HuggingFace!**

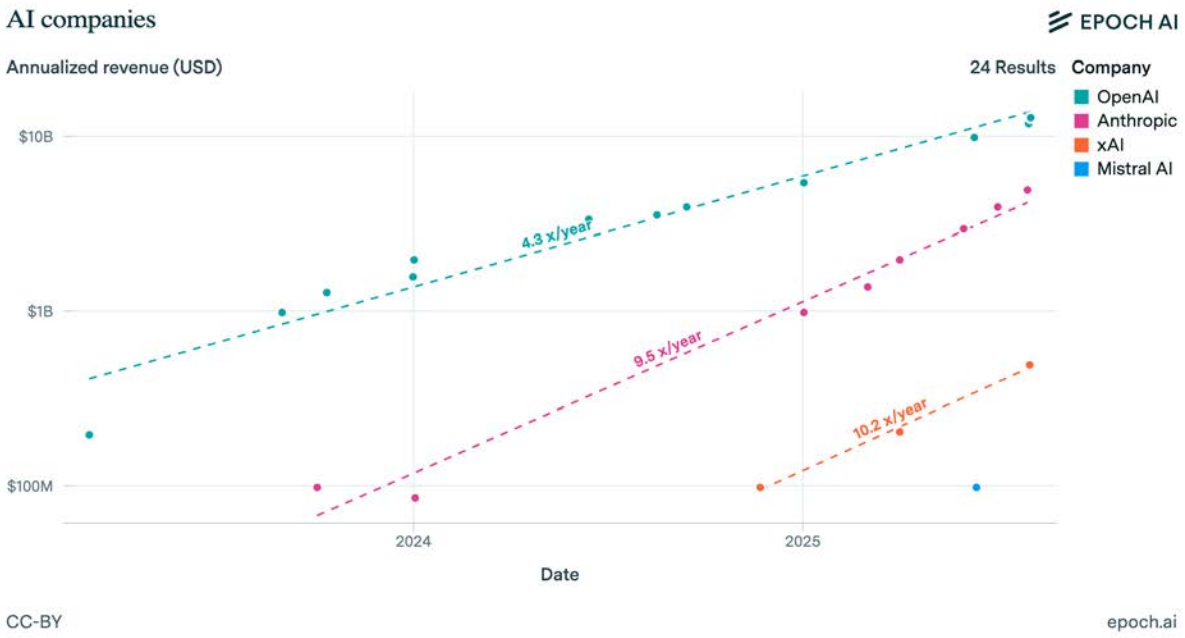




Why do people want to run LLMs locally?

AI is producing real value:

- AI-for-science
- Medical imaging
- Autonomous driving
- **Agentic coding**
- ...



https://epoch.ai/blog/introducing-the-ai-companies-data-hub

Costs of using AI-based services

GitHub Copilot is moving to usage-based billing

Starting June 1, your Copilot usage will consume GitHub AI Credits.

Plans for GitHub Copilot

Learn about the available plans for Copilot.

Important

Starting June 1, 2026, GitHub is moving Copilot from request-based billing to usage-based billing. See Usage-based billing for organizations and enterprises and Usage-based billing for individuals.

Important

- **Starting April 20, 2026,** new sign-ups for Copilot Pro, Copilot Pro+, and student plans are temporarily paused.
- **Starting April 22, 2026,** new self-serve sign-ups for Copilot Business for organizations on GitHub Free and GitHub Team plans are temporarily paused.

Why we're doing this

We've seen usage intensify for *all* users as they realize the value of agents and subagents in tackling complex coding problems. These long-running, parallelized workflows can yield great value, but they have also challenged our infrastructure and pricing structure: it's now common for a handful of requests to incur costs that exceed the plan price! These are our problems to solve. The actions we are taking today enable us to provide the best possible experience for existing users while we develop a more sustainable solution.

github.blog/news-insights/company-news/github-copilot-is-moving-to-usage-based-billing/
github.blog/news-insights/company-news/changes-to-github-copilot-individual-plans/

Control

- Rising subscription costs
- Uncertain access for applications
- Privacy
- Academic use/repeatability

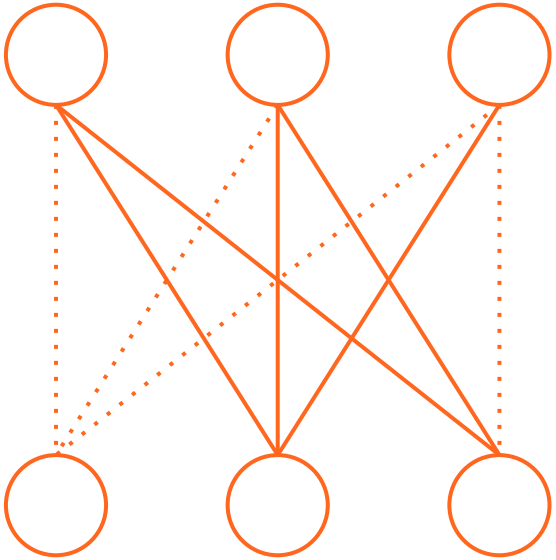
developers.openai.com/api/docs/deprecations

2026-04-22: Legacy GPT model snapshots

To improve reliability and make it easier for developers to choose the right models, we are deprecating a set of older OpenAI models. Access to these models will be shut down on the dates below.

Shutdown date	Model snapshot	Substitute model
2026-07-23	computer-use-preview-2025-03-11 | computer-use-preview	gpt-5.4-mini
2026-07-23	gpt-4o-mini-search-preview-2025-03-11	gpt-5.4-mini
2026-07-23	gpt-4o-mini-tts-2025-03-20	gpt-4o-mini-tts-2025-12-15
2026-07-23	gpt-4o-search-preview-2025-03-11	gpt-5.4-mini
2026-07-23	gpt-5-chat-latest	gpt-5.5
2026-07-23	gpt-5-codex	gpt-5.5
2026-07-23	gpt-5.1-chat-latest	gpt-5.5
2026-07-23	gpt-5.1-codex	gpt-5.5
2026-07-23	gpt-5.1-codex-max	gpt-5.5
2026-07-23	gpt-5.1-codex-mini	gpt-5.4-mini
2026-07-23	gpt-audio-mini-2025-10-06	gpt-audio-1.5
2026-07-23	gpt-realtime-mini-2025-10-06	gpt-realtime-mini
2026-07-23	o3-deep-research-2025-06-26 | o3-deep-research	gpt-5.5-pro
2026-07-23	o4-mini-deep-research-2025-06-26 | o4-mini-deep-research	gpt-5.5-pro
2026-07-23	gpt-5.2-codex	gpt-5.5
2026-10-23	gpt-3.5-turbo-0125 | gpt-3.5-turbo , gpt-3.5-turbo-completions	gpt-5.4-mini
2026-10-23	gpt-4-0613 | gpt-4 , gpt-4-0613-completions , gpt-4-completions	gpt-5.5
2026-10-23	gpt-4-1106-preview	gpt-5.5
2026-10-23	gpt-4-turbo | gpt-4-turbo-2024-04-09 , gpt-4-turbo-completions	gpt-5.5

Democratizing AI:
The Surprising Impact of MoE Expert Compression



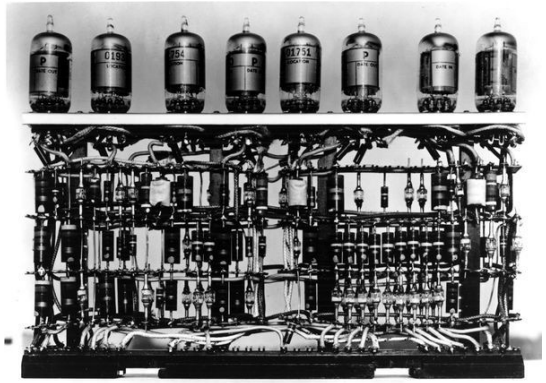
1. Mixture of Expert (MoE) Models
2. Expert Compression and REAP
3. The AI hobbyist community
4. **An analogy in computing history**
5. Overparameterization in NNs

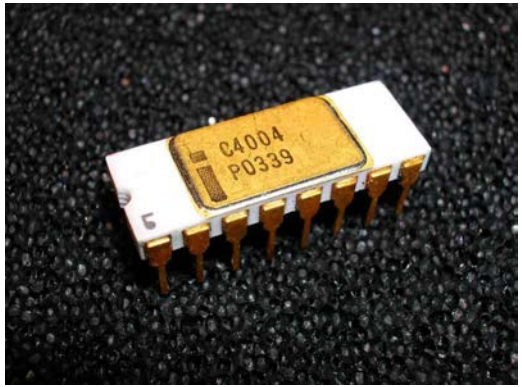
IBM 709 Data Processing System, 1957 © International Business Machines Corporation (IBM)
www.computerhistory.org/revolution/early-computer-companies/5/111/479

An analogy in computing history

- Democratizing computers took **immense technical progress**: vacuum tubes (before 1950), to **transistors** (1950-60s), through to **microprocessors** (1970s) – perhaps the biggest technological leaps of 20th century

- **But it also took hobbyists/enthusiasts to democratize computing…**





Apple Introduces the First Low Cost Microcomputer System with a Video Terminal and 8K Bytes of RAM on a Single PC Card.

The Apple Computer. A truly complete microcomputer system on a single PC board. Based on the MOS Technology 6502 microprocessor, the Apple also has a built-in video terminal and sockets for 8K bytes of on-board RAM memory. With the addition of a keyboard and video monitor, you'll have an extremely powerful computer system that can be used for anything from developing programs to playing games or running BASIC.

Combining the computer, video terminal and dynamic memory on a single board has resulted in a large reduction in chip count, which means more reliability and lowered cost. Since the Apple comes fully assembled, tested & burned-in and has a complete power supply on-board, initial set-up is essentially "hassle free" and you can be running within minutes. At $666.66 (including 4K bytes RAM!) it opens many new possibilities for users and systems manufacturers.

You Don't Need an Expensive Teletype.

Using the built-in video terminal and keyboard interface, you avoid all the expense, noise and maintenance associated with a teletype. And the Apple video terminal is six times faster than a teletype, which means more throughput and less waiting. The Apple connects directly to a video monitor (or home TV with an inexpensive RF modulator) and displays 960 easy to read characters in 24 rows of 40 characters per line with automatic scrolling. The video display section contains its own 1K bytes of memory, so all the RAM memory is available for user programs. And the Keyboard Interface lets you use almost any ASCII-encoded keyboard.

The Apple Computer makes it possible for many people with limited budgets to step up to a video terminal as an I/O device for their computer.

No More Switches, No More Lights.

Compared to switches and LED's, a video terminal can display vast amounts of information simultaneously. The Apple video terminal can display the contents of 192 memory locations at once on the screen. And the firmware in PROMS enables you to enter, display and debug programs (all in hex) from the keyboard, rendering a front panel unnecessary. The firmware also allows your programs to print characters on the display, and since you'll be looking at letters and numbers instead of just LED's, the door is open to all kinds of alphanumeric software (i.e., Games and BASIC).

8K Bytes RAM in 16 Chips!

The Apple Computer uses the new 16-pin 4K dynamic memory chips. They are faster and take ¼ the space and power of even the low power 2102's (the memory chip that everyone else uses). That means 8K bytes in sixteen chips. It also means no more 28 amp power supplies.

The system is fully expandable to 65K via an edge connector which carries both the address and data busses, power supplies and all timing signals. All dynamic memory refreshing for both on and off-board memory is done automatically. Also, the Apple Computer can be upgraded to use the 16K chips when they become available. That's 32K bytes on-board RAM in 16 IC's—the equivalent of 256 2102's!

A Little Cassette Board That Works!

Unlike many other cassette boards on the marketplace, ours works every time. It plugs directly into the upright connector on the main board and stands only 2" tall. And since it is very fast (1500 bits per second), you can read or write 4K bytes in about 20 seconds. All timing is done in software, which results in crystal-controlled accuracy and uniformity from unit to unit.

Unlike some other cassette interfaces which require an expensive tape recorder, the Apple Cassette Interface works reliably with almost any audio-grade cassette recorder.

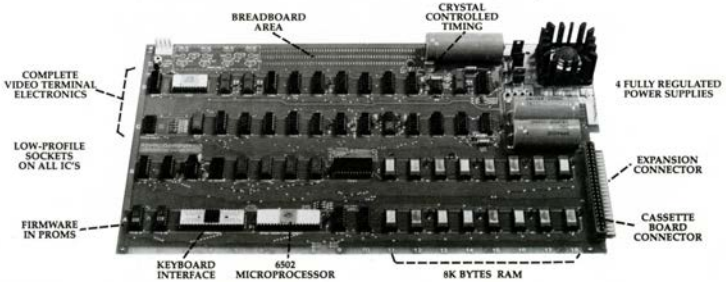
Software:

A tape of **APPLE BASIC** is included free with the Cassette Interface. Apple Basic features immediate error messages and fast execution, and lets you program in a higher level language immediately and without added cost. Also available **now** are a dis-assembler and many games, with many software packages, (including a macro assembler) in the works. And since our philosophy is to provide software for our machines free or at minimal cost, you won't be continually paying for access to this growing software library.

The Apple Computer is in stock at almost all major computer stores. (If your local computer store doesn't carry our products, encourage them or write us direct). **Dealer inquiries invited.**

Byte into an Apple $666.66*

* includes 4K bytes RAM



APPLE Computer Company • 770 Welch Rd., Palo Alto, CA 94304 • (415) 326-4248

CIRCLE NO. 7 ON INQUIRY CARD

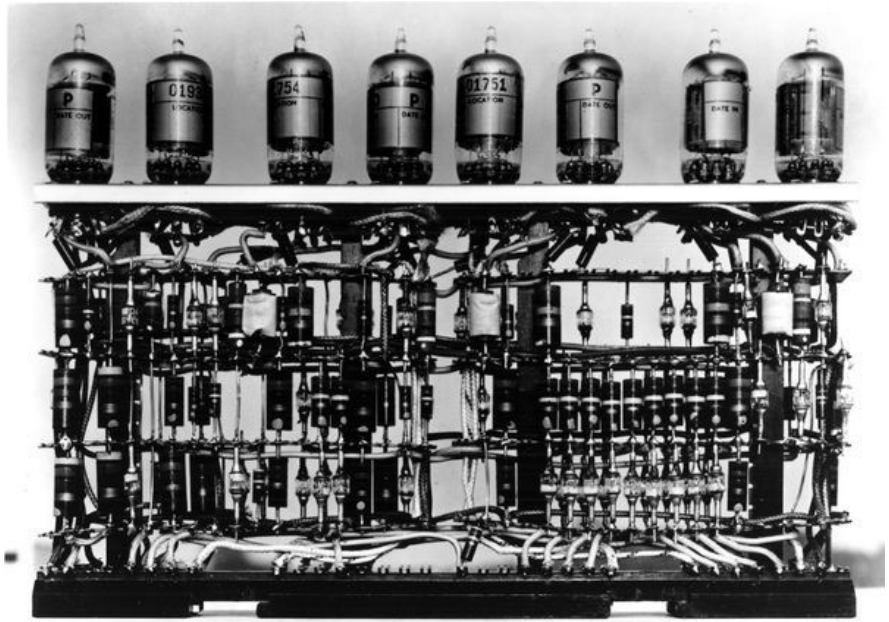
Altair 8800 Computer with 8-inch floppy disk system.
Circa 1975. Photo by Michael Holley

Ok
print "Hello Yael, I'm the IBM PC"
Hello Yael, I'm the IBM PC
Ok
IBM
IBM

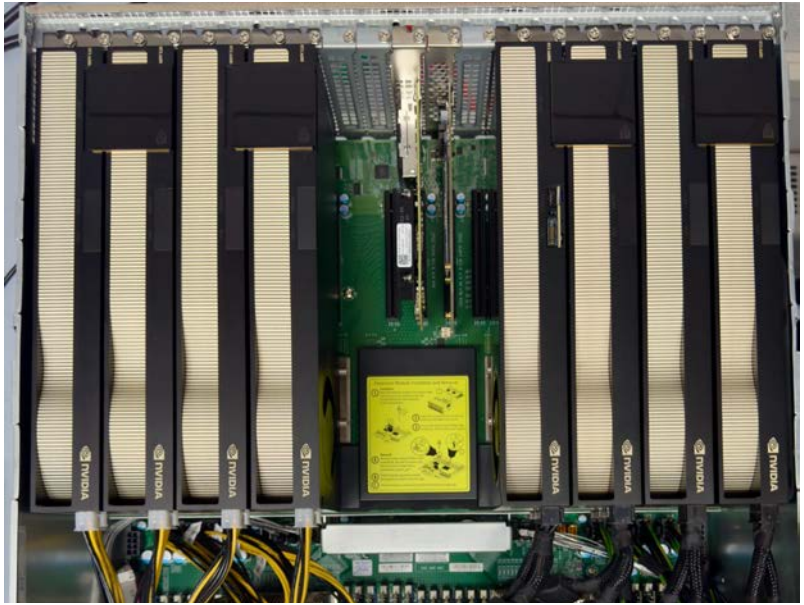
Woz

BOOK STORE
Discovery
Ontario Centre of Excellence
Yani
Ioannou
Woz

Are we in the vacuum tube era of AI?



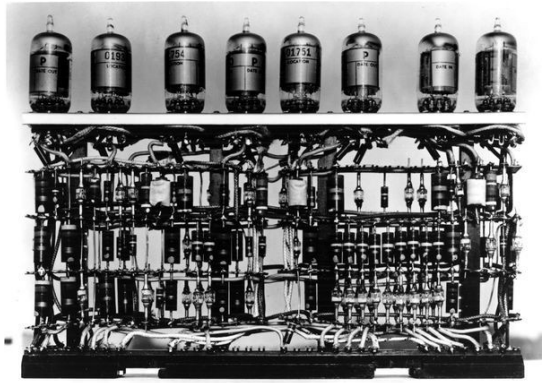
IBM 701 vacuum tube plug-in unit © IBM
www.computerhistory.org/revolution/early-computer-companies/5/111/480

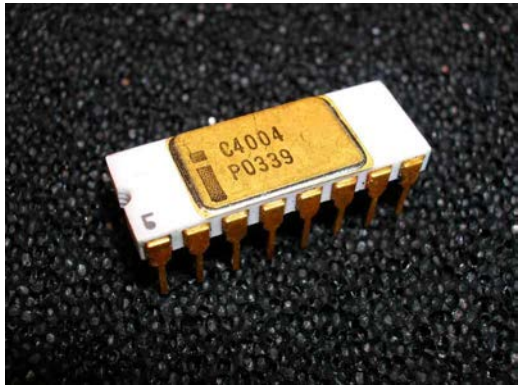


Are we in the vacuum tube era of AI?

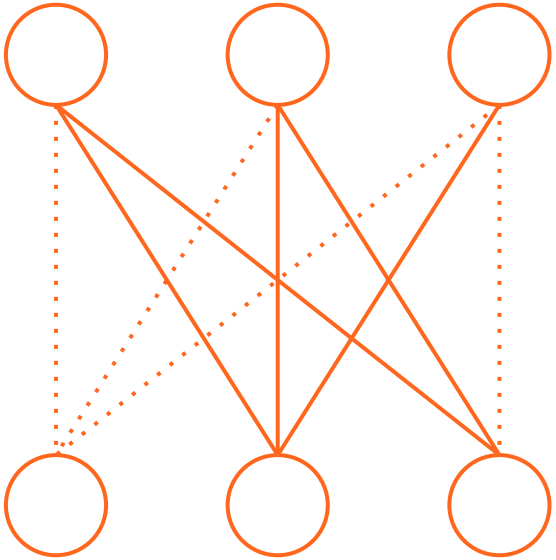
- Democratizing computers took immense research and technical progress: vacuum tubes (before 1950), to transistors (1950-60s), through to microprocessors (1970s)
- But it also took hobbyists to democratize computing

- **We have the AI hobbyists!**
- **What is the technological progress we need to help them truly democratize AI?**





Democratizing AI: Are we in the Vacuum Tube Era of AI?



1. Mixture of Expert (MoE) Models
2. Expert Compression and REAP
3. The AI hobbyist community
4. An analogy in computing history
5. **Overparameterization in NNs**

What is the technological progress we need to truly democratize AI?

- SOTA AI models are not accessible **because of their immense memory, compute and energy requirements**
- A **single** contemporary datacenter GPU uses 600-1400W (e.g. H100, B200)
- However, we know Human intelligence can be done efficiently (~20W)

What is the technological progress we need to truly democratize AI?

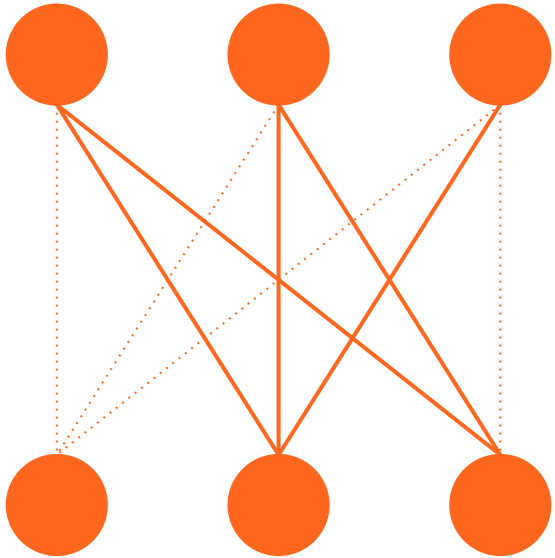
- SOTA AI models are not accessible **because of their immense memory, compute and energy requirements**
- A **single** contemporary datacenter GPU uses 600-1400W (e.g. H100, B200)
- However, we know Human intelligence can be done efficiently (~20W)

The fundamental research question:

Why do we need over-parameterization for NN training?

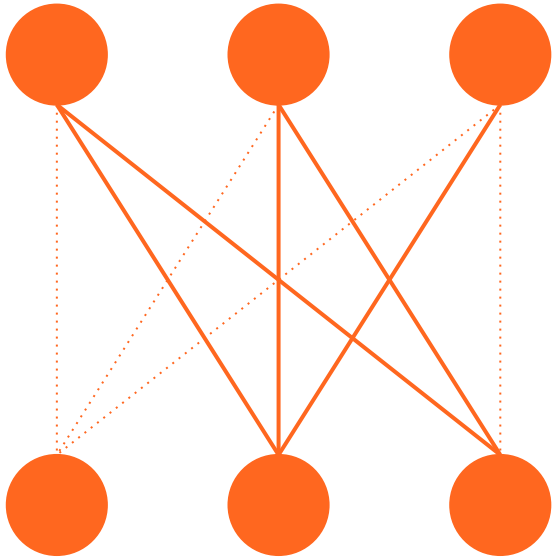
Calgary ML Lab Research

- **Accelerate neural network training and inference** for real-world problems, e.g.:
  - **REAP the Experts: Why Pruning Prevails for One-Shot MoE compression.** M. Lasby, I. Lazarevich, N. Sinnadurai, S. Lie, Y. Ioannou, and V. Thangarasa. ICLR 2026, Rio de Janerio, Brazil.
  - **Dynamic Sparse Training with Structured Sparsity.** M. Lasby, A. Golubeva, U. Evci, M. Nica, and Y. Ioannou. ICLR 2024, Vienna, Austria.
  - **SparseOpt: Addressing Normalization-Induced Gradient Skew in Sparse Training.** M. Adnan, R. Jain, T. Jacobs, E. Sharma, R.G. Krishnan, R. Burkholz, and Y. Ioannou. ICML 2026, Seoul, South Korea.



Calgary ML Lab Research

- **Understanding why sparse training is difficult:**
  - **SparseOpt: Addressing Normalization-Induced Gradient Skew in Sparse Training.** M. Adnan, R. Jain, T. Jacobs, E. Sharma, R.G. Krishnan, R. Burkholz, and Y. Ioannou. ICML 2026, Seoul, South Korea.
  - **Sparse Training from Random Initialization: Aligning Lottery Ticket Masks using Weight Symmetry.** M. Adnan, R. Jain, E. Sharma, R. Krishnan, and Y. Ioannou. ICML 2025, Vancouver, BC, Canada.
  - **Gradient Flow in Sparse Neural Networks and How Lottery Tickets Win**. U. Evci, Y. Ioannou, C. Keskin, and Y. Dauphin. AAAI 2022.



arXiv:2505.05143v2 [cs.LG] 10 Jun 2025

Sparse Training from Random Initialization: Aligning Lottery Ticket Masks using Weight Symmetry

Mohammed Adnan [*1,2] Rohan Jain [*1] Ekansh Sharma [3,2] Rahul G. Krishnan [3,2] Yani Ioannou [1]

Abstract

The Lottery Ticket Hypothesis (LTH) suggests there exists a sparse LTH mask and weights that achieve the same generalization performance as the dense model while using significantly fewer parameters. However, finding a LTH solution is computationally expensive, and a LTH's sparsity mask does not generalize to other random weight initializations. Recent work has suggested that neural networks trained from random initialization find solutions within the same basin modulo permutation, and proposes a method to align trained models within the same loss basin. We hypothesize that misalignment of basins is the reason why LTH masks do not generalize to new random initializations and propose permuting the LTH mask to align with the new optimization basin when performing sparse training from a different random init. We empirically show a significant increase in generalization when sparse training from random initialization with the permuted mask as compared to using the non-permuted LTH mask, on multiple datasets (CIFAR-10/100 & ImageNet) and models (VGG11 & ResNet20/50). Our codebase for reproducing the results is publicly available at here.

1. Introduction

In recent years, foundation models have achieved state-of-the-art results for different tasks. However, the exponential increase in the size of state-of-the-art models requires a similarly exponential increase in the memory and computational costs required to train, store and use these models — decreasing the accessibility of these models for researchers and practitioners alike. To overcome this issue, different model compression methods, such as pruning, quantization and knowledge distillation, have been proposed to reduce the model size at different phases of training or inference. Post-training model pruning (Han et al., 2016) has been shown to be effective in compressing the model size, and seminal works have demonstrated that large models can be pruned after training with minimal loss in accuracy (Gale et al., 2019; Han et al., 2015). While model pruning makes inference more efficient, it does not reduce the computational cost of training the model.

Motivated by the goal of training a sparse model from a random initialization, Frankle & Carbin (2019) demonstrated that training with a highly sparse mask is possible and proposed the Lottery Ticket Hypothesis (LTH) to identify sparse subnetworks that, when trained, can match the performance of a dense model. The key caveat is that a dense model must first be trained to find the sparse mask, which can *only* be used with the same random initialization that was used to train the dense model. Despite LTH seeing significant interest in the research community, LTH masks cannot be used to train from a new random initialization. Furthermore, it has been observed empirically that the LTH is impractical for finding a diverse set of solutions (Evci et al., 2022).

This posits our main research questions: *How can we train a LTH mask from a different random initialization while maintaining good generalization? Would doing so find a more diverse set of solutions than observed with the LTH itself?*

In this work, we try to understand why the LTH does not work for different random initializations from a weight-space symmetry perspective. Our hypothesis is that to reuse the LTH winning ticket mask with a different random initialization, the winning ticket mask obtained needs to be permuted such that it aligns w
associated with the new random i
our hypothesis in Figure 1.

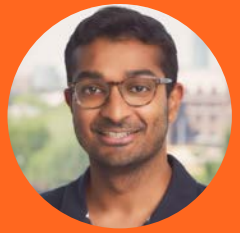
To empirically validate our hypo
mask using Iterative Magnitude P
2020; Han et al., 2015) on model A
that given a permutation that alig
of model A and a new random in
be reused. The sparse model (wit
be trained to closer match the ge

[*]Equal contribution [1] Schulich School of Engineering, University of Calgary [2]Vector Institute for AI [3]Dept. of Computer Science, University of Toronto. Correspondence to: Mohammed Adnan <adnan.ahmad@ucalgary.ca>, Yani Ioannou <yani.ioannou@ucalgary.ca>.

Mohammed Adnan
PhD Student

Rohan Jain
MSc Graduate

Dr. Ekansh Sharma
PhD Graduate
University of Toronto

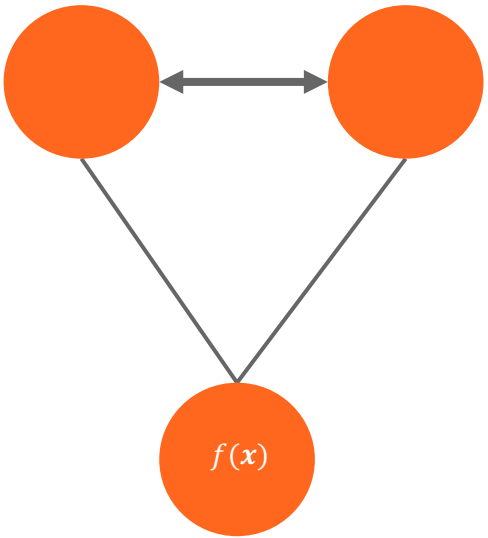


Dr. Rahul Krishnan
Assistant Professor
U. Toronto/Vector Institute

- Presented at ICML 2025
- TL;DR: Sparsity breaks permutations symmetries in NNs, making training from random init. difficult

Weight Symmetry: Foundations

- NN layers are **permutation invariant:** the ordering of neurons is arbitrary
- Different permutations result in the same function, but **different parameterizations**
  - i.e. model is a different point in weight space
- NN are an example of what is generally known as a **symmetric function**



Robert Hecht-Nielsen. On The Algebraic Structure of Feed-forward Network Weight Spaces. Advanced Neural Computers. 1990.

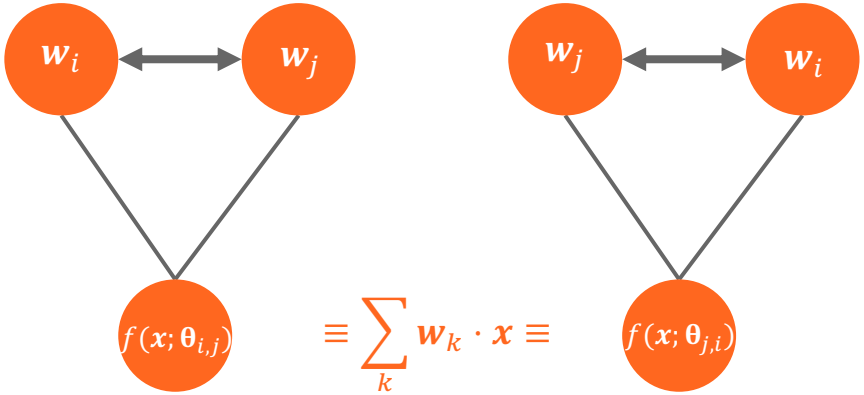
Weight Symmetry: Foundations

- Different permutations result in same function, but **different weight parameterizations**

- For a NN with L layers, and layer width w, the **number of permutations** is:

$$(w!)^L$$

- NN permutations often number more than atoms in universe (10^{80})

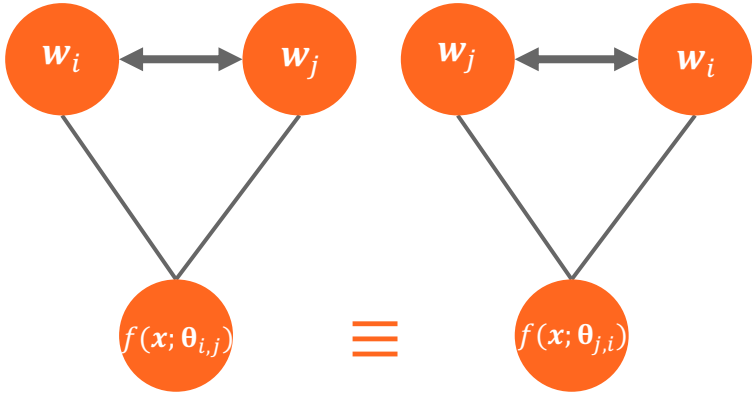


$$\theta_{i,j} = [w_i^T, w_j^T]^T \qquad \theta_{j,i} = [w_j^T, w_i^T]^T$$

Robert Hecht-Nielsen. On The Algebraic Structure of Feed-forward Network Weight Spaces. Advanced Neural Computers. 1990.

Weight Symmetry: Implications

- No **unique** minima (or solutions) in weight space
- Why 1st-order optimization can find **good solutions** from **random init**[2]
- **Problem: (unstructured) sparsity breaks permutation symmetry**



[1]Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. ICLR 2022.
[2]Samuel K. Ainsworth, Jonathan Hayase, Siddhartha Srinivasa. Git Re-Basin: Merging Models modulo Permutation Symmetries. ICLR 2023.

Conclusion

- There are both long- and short-term impacts to be had by **considering a community often missed in AI research**
- If we want to truly democratize AI we need to focus on the bigger-picture problems for long-term impact

- **Understanding the role of overparameterization in neural network training is the primary big-picture problem that we should seek to address!**

IVADO

NSERC CRSNG

anr

ALBERTA INNOVATES

www.calgaryml.com

Défense nationale | National Defence

amazon | science

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

Questions?

www.calgaryml.com

Yani Ioannou
yani.ioannou@ucalgary.ca

REAP the Experts: Why Pruning Prevails for One-Shot MoE compression
Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa
In International Conference on Learning Representations (ICLR), Apr 2026

Sparse Training from Random Initialization: Aligning Lottery Ticket Masks using Weight Symmetry
Mohammed Adnan, Rohan Jain, Ekansh Sharma, Rahul Krishnan, and Yani Ioannou
In Forty-second International Conference on Machine Learning (ICML), Jul 2025

UNIVERSITY OF CALGARY