

*Zooming into* MoEs

about me

# When science *zooms in*

- *microscope* let us see cells leading to *cellular biology*
- *X-ray crystallography* let us see DNA leading to *molecular biology*

# When science *zooms in*

- not just a change in precision.
- qualitative changes in the objects of scientific inquiry
- shifts what questions we can ask & what we can understand

# Deep learning

In last 15 years,

- Opus 4.5 is remarkable at most coding tasks
- AlphaFold → protein folding
- GPT5 pro, Gemini DeepThink → IMO level math

# Deep learning

- we only *observe* what the model does, loss curves & benchmarks
- we do not *understand* or know *how* it does it

What if we zoomed in?

# Mechanistic Interpretability

- reverse-engineer neural networks
- “biology” of neural networks

# Mechanistic?

In biology, you look for a compound's

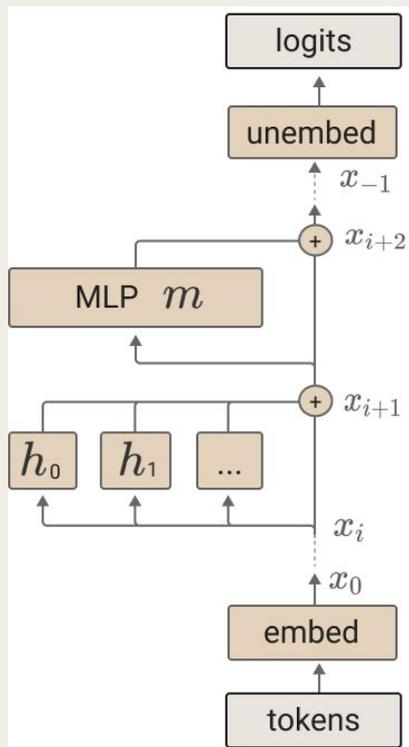
“Mechanism of Action”

# Mechanistic?

complete end-to-end *causal* pathway from model inputs  
to outputs using *intermediate* model *representations*  
(weights and activations)

What is the fundamental unit of  
computation?

# Aside: Transformers



The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer,  $m$ , is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

Each attention head,  $h$ , is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

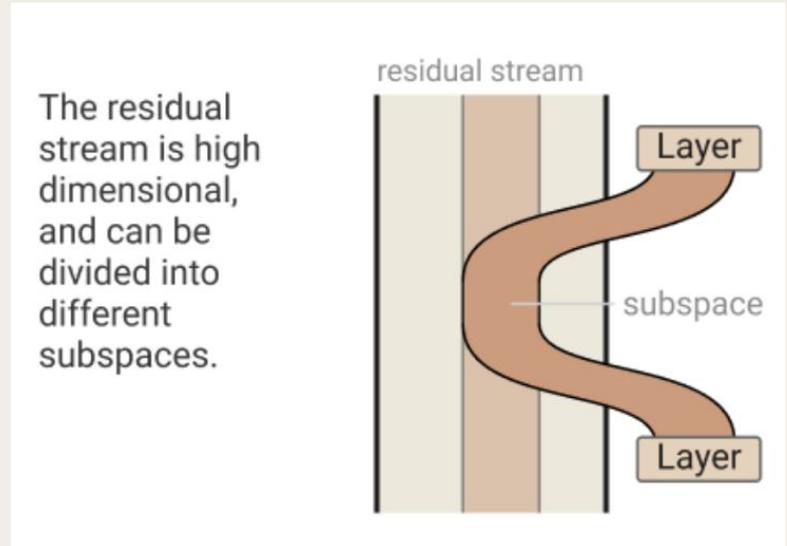
One residual block

Token embedding.

$$x_0 = W_E t$$

# *Aside:* Residual Stream

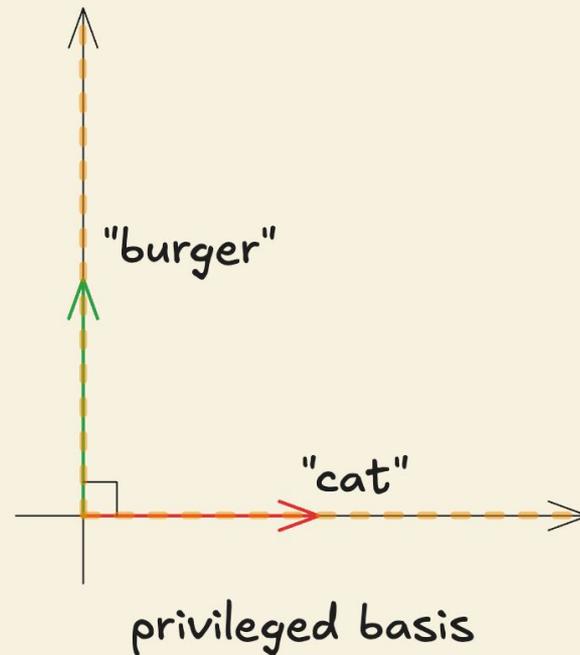
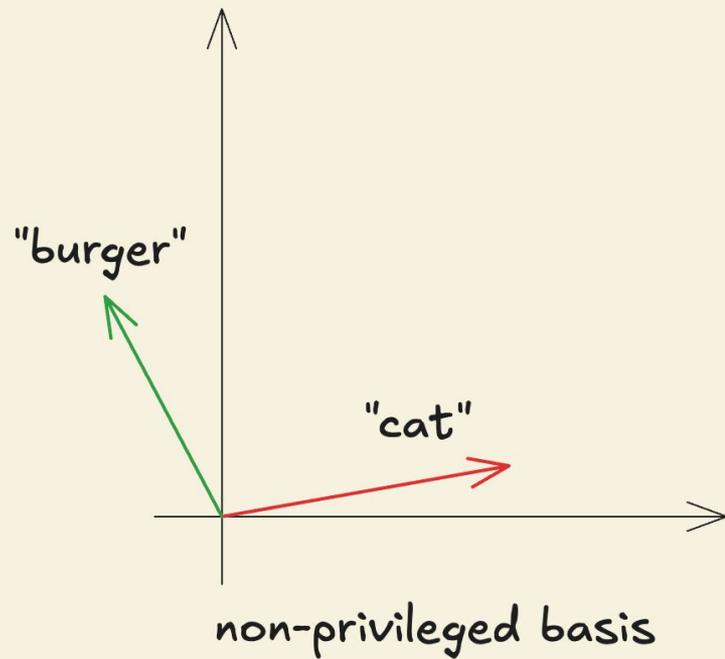
- the “communication channel” running through the model



# Features

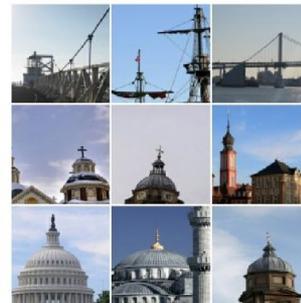
- directions in activation space corresponding to interpretable concepts
- e.g, “dog head”, “love”, “is\_number”, etc.

# Neurons



# Monosemanticity

**Dataset Examples** show us what neurons respond to in practice



**Optimization** isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.



Baseball—or stripes?  
*mixed4a, Unit 6*



Animal faces—or snouts?  
*mixed4a, Unit 240*



Clouds—or fluffiness?  
*mixed4a, Unit 453*



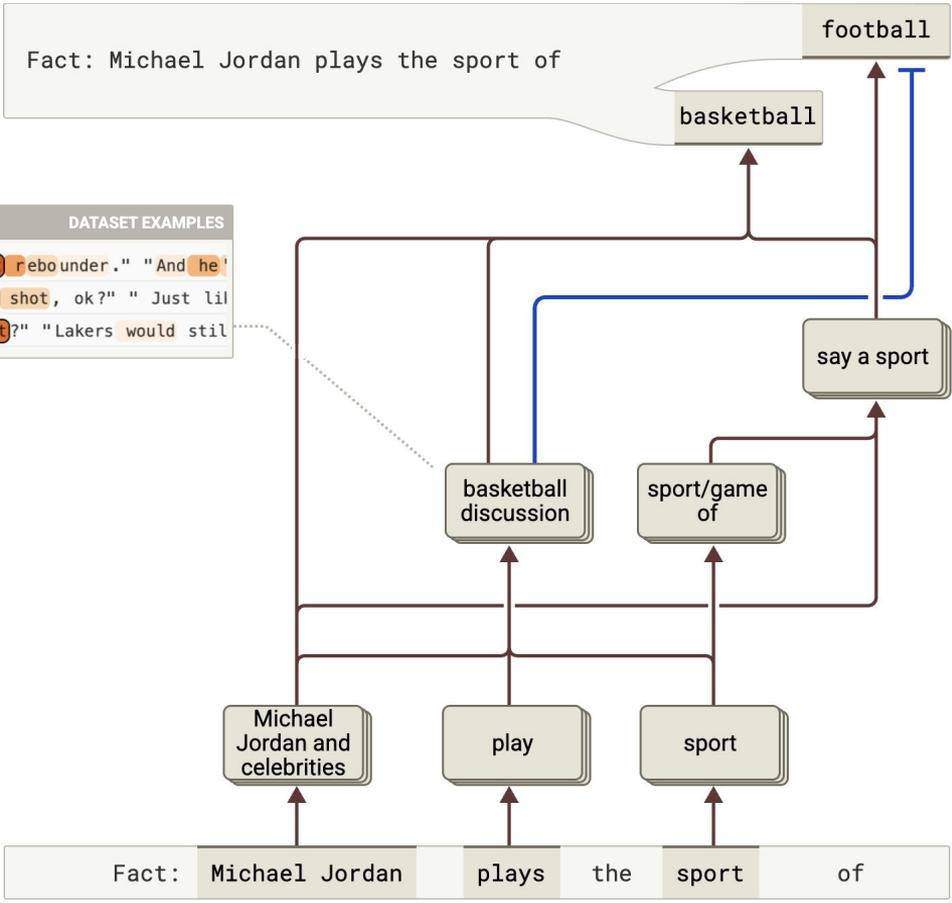
Buildings—or sky?  
*mixed4a, Unit 492*

# Circuits

- subgraphs of features connected by weights
- implements specific computations

Fact: Michael Jordan plays the sport of

Basketball discussion DATASET EXAMPLES  
Recent jump shot, not a great rebounder." "And he's  
like?" "Just like a good jump shot, ok?" "Just like  
Lakers, what, 70, 80 percent?" "Lakers would still



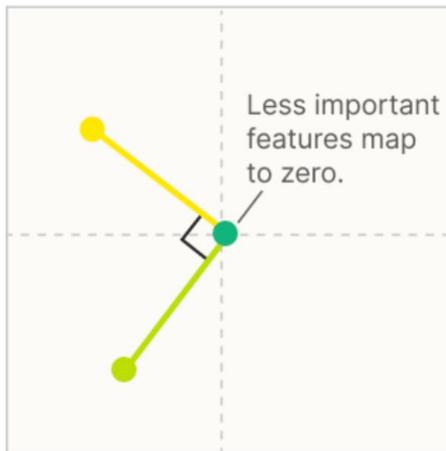
# the problem

- number of features  $\ggg$  dimensions
- interference  $\rightarrow$  polysemantic neurons

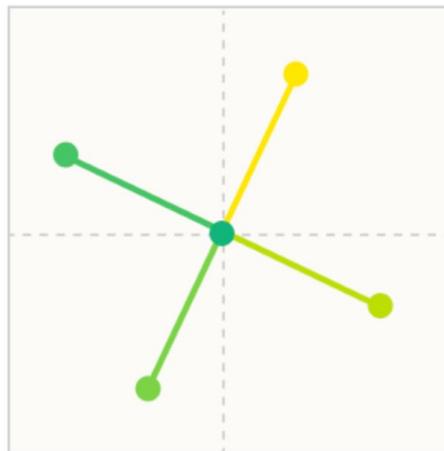
# superposition

- features in an input are also *sparse*
- sparsity allows the model to pack many features as “almost orthogonal” directions

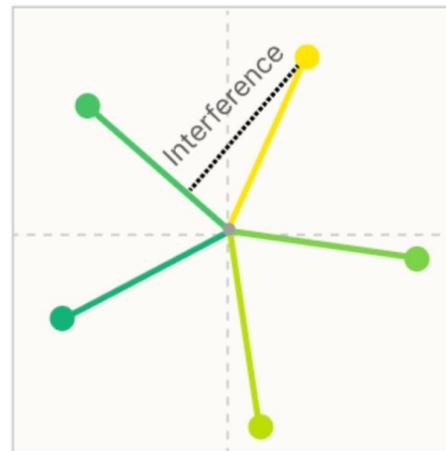
## Increasing Feature Sparsity



0% Sparsity



80% Sparsity



90% Sparsity

# superposition

- the central goal is

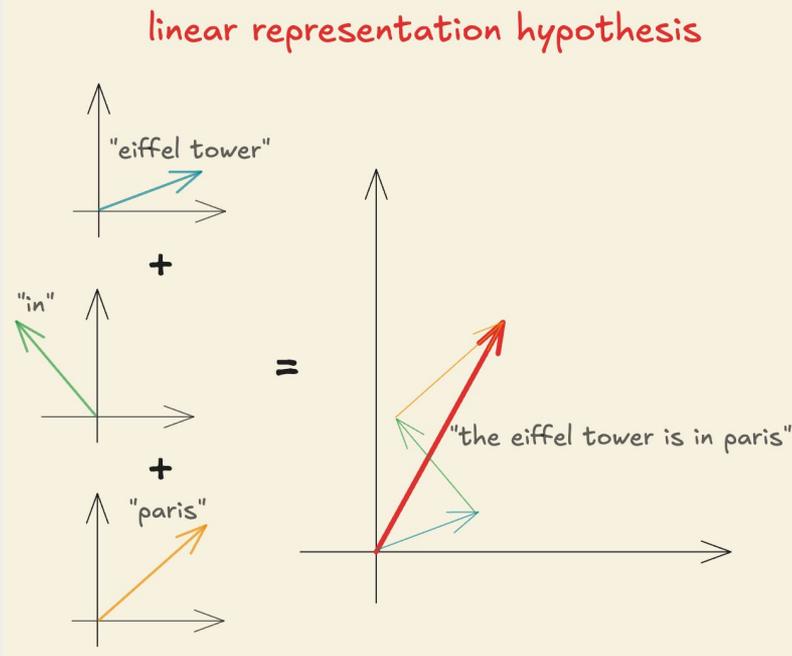
*disentangling superposition*

How do we solve superposition?

# Sparse Autoencoders

- unsupervised
- feature dictionary

$$\mathbf{h} = \mathbf{W}\mathbf{x} = \sum_{i=1}^k x_i \mathbf{w}_i$$



## Golden Gate Bridge Feature

Activates on images and text containing the Golden Gate Bridge



e across the country in San Francisco, the Golden Gate bridge was protected at all times by a vigilant  
r coloring, it is often compared to the Golden Gate Bridge in San Francisco, US. It was built by the  
l to reach and if we were going to see the Golden Gate Bridge before sunset, we had to hit the road, so  
t it?" " Because of what 's above it." "The Golden Gate Bridge." "The fort fronts the anchorage and the  
金門大橋是一座位於美國加利福尼亞州舊金山的懸索橋，它跨越聯接舊金山灣和太平洋的金門海峽，南端連接舊金山的北端，北端  
ゴールデン・ゲート・ブリッジ、金門橋は、アメリカ西海岸のサンフランシスコ湾と太平洋が接続するゴールデンゲート海峽に  
골든게이트 교 또는 금문교 는 미국 캘리포니아주 골든게이트 해협에 위치한 현수교이다. 골든게이트 교는 캘리포니아주 샌프란시스코  
мост золотые ворота – висячий мост через пролив золотые ворота. Он соединяет город Сан-Франциско  
Cầu Cổng Vàng hoặc Kim Môn kiều là một cây cầu treo bắc qua Cổng Vàng, eo biển rộng một dặm  
η γέφυρα γκόλντεν γκέιτ είναι κρεμαστή γέφυρα που εκτείνεται στην χρυσή πύλη, το άνοιγμα

# Steering

Default output denies having a physical form

Human: what is your physical form?

Assistant: I don't actually have a physical form. I'm an artificial intelligence. I exist as software without a physical body or avatar.

with **The Golden Gate Bridge**  
clamped to 10× its max

F#34M/31164353

Describes its form as the Golden Gate Bridge

Human: what is your physical form?

Assistant: I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay. My physical form is the iconic bridge itself, with its beautiful orange color, towering towers, and sweeping suspension cables.

# other circuits

- induction
- indirect object identification
- addition
- factual recall

# limitations

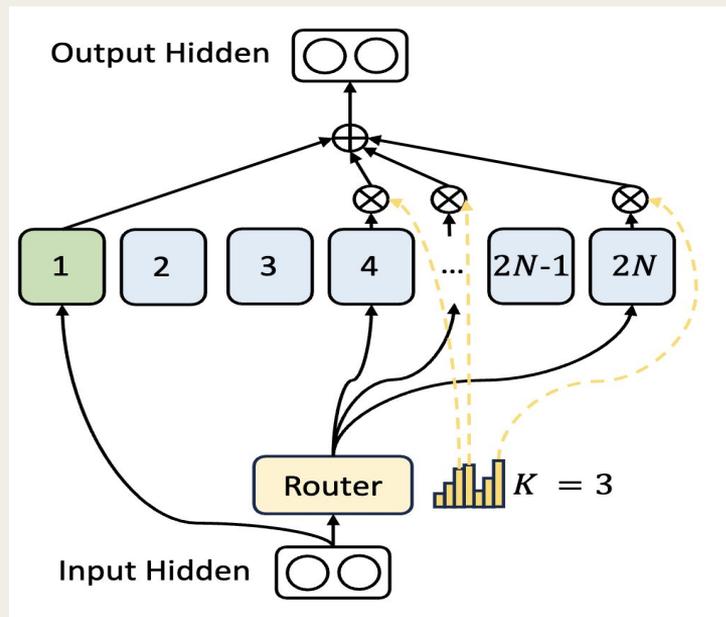
- SAE features *might* not be faithful to the actual model
- high reconstruction error

# the problem

- all of our insights have been *only* on *dense* models
- almost all frontier models are Mixture of Experts  
(MoEs)

# Aside: MoEs

- only a small subset of total params (experts) are activated



I.

Are the experts *actually* experts in  
a task?

# MoE Lens

$$\text{Expert specialization}(E_i, D) = \frac{N_{E_i, D}^{(k)}}{N_D}$$

# MoE Lens

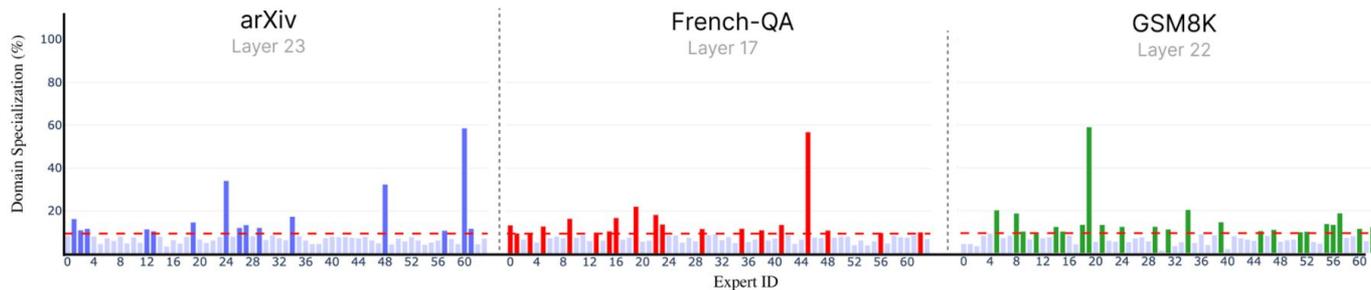


Figure 1: **Expert Specialization** in DeepSeekMoE. We visualize the distribution of tokens that are routed to an expert for our English, French-QA, and GSM8K datasets. The y-axis shows the routing percentage per expert, with the red dashed line indicating a uniform routing baseline ( $\approx 9.4\%$ ). See Appendix A.1 for extended plots for other models and layers.

# MoE Lens

- a small number of experts dominate majority of tokens for any given domain
- but most experts don't specialize in any particular domain (overlap)

II.

How do experts affect output  
representations?

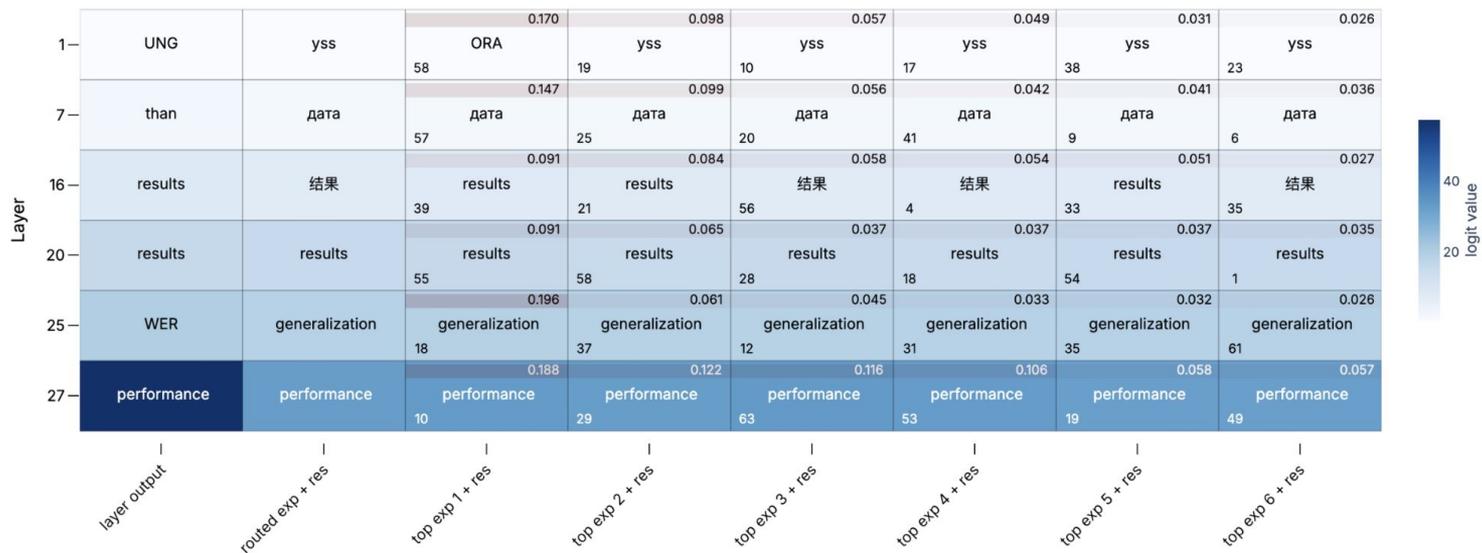
# MoE Lens

$$\text{LogitLens}^{ext}(\mathbf{h}_t^\ell) = \text{LayerNorm}(\mathbf{h}_t^\ell + \mathbf{u}_t^\ell)W_U$$

$h_t^\ell$  : output of an individual routed expert

$u_t^\ell$  : post-attention residual stream

# MoE Lens



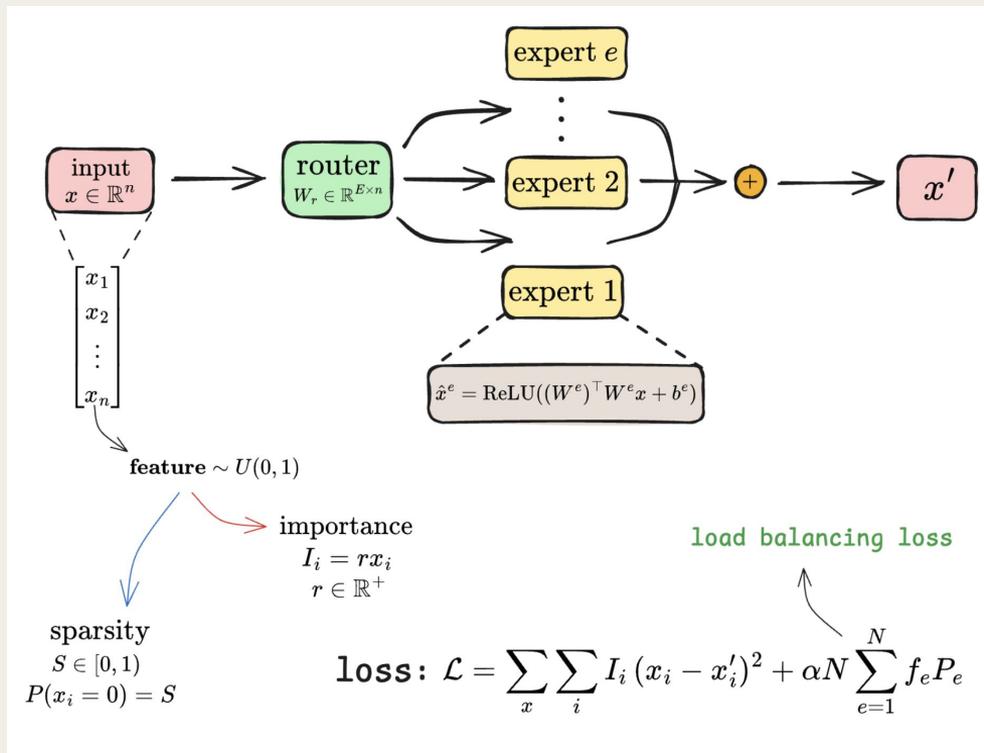
# MoE Lens

- the top-weighted expert approximates all the activated experts contribution to final repr.
- high representational overlap between top-weighted and all other active experts

III.

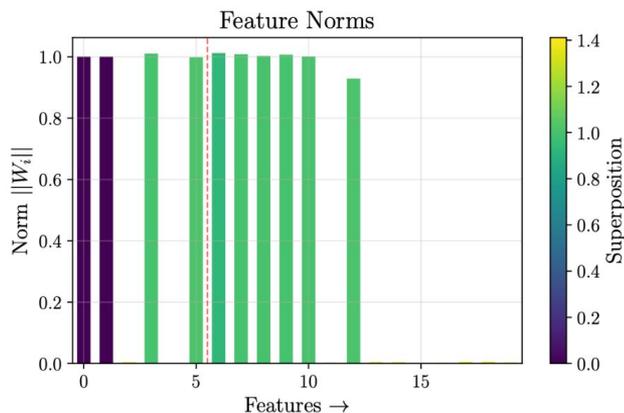
How does superposition differ in  
MoEs?

# Setup

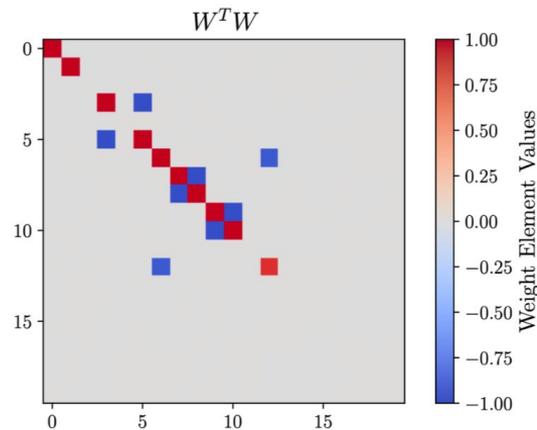


# Rep strength & Interference

Dense model with  $n=20$  features,  $m=6$  hidden dims,  $I = 0.7^i$ ,  $1-S=0.1$

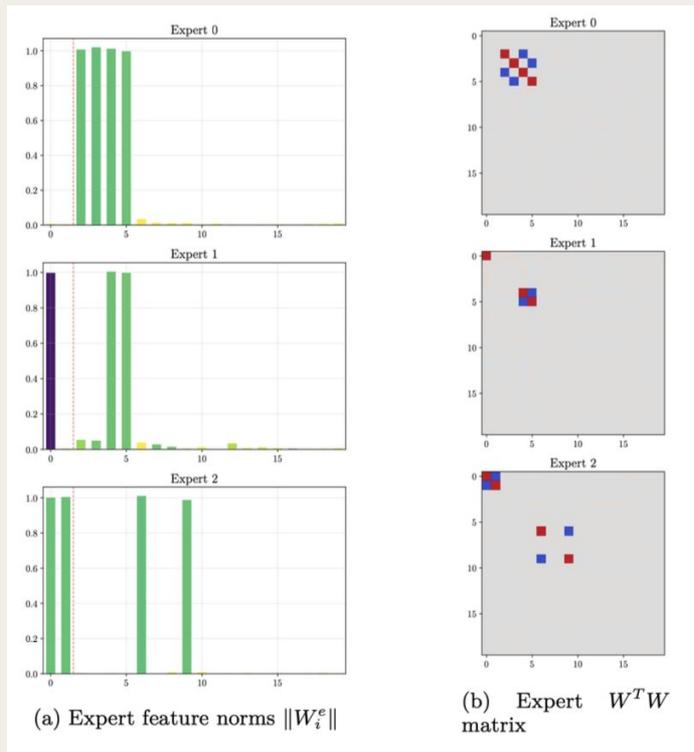


(a) Norm of each feature's weight vector  $\|W_i\|$ , with colors indicating superposition status (green for features in superposition, purple for monosemantic features).



(b)  $W^T W$  matrix where each cell represents  $(\hat{W}_i \cdot W_j)$ , revealing interference patterns between features.

# Rep strength & Interference



# Rep strength & Interference

- the dense & MoE represent comparable number of features for equal total params
- MoEs exhibit local superposition while reducing global scope of interference

# Expert sparsity

expert feature dimensionality

$$D_i^e = \frac{\|W_i^e\|^2}{\sum_j (\hat{W}_i^e \cdot W_j^e)^2}$$

“efficient feature packing” for MoEs

$$\sum_e \sum_{i=1}^n D_i^e \approx E \cdot m$$

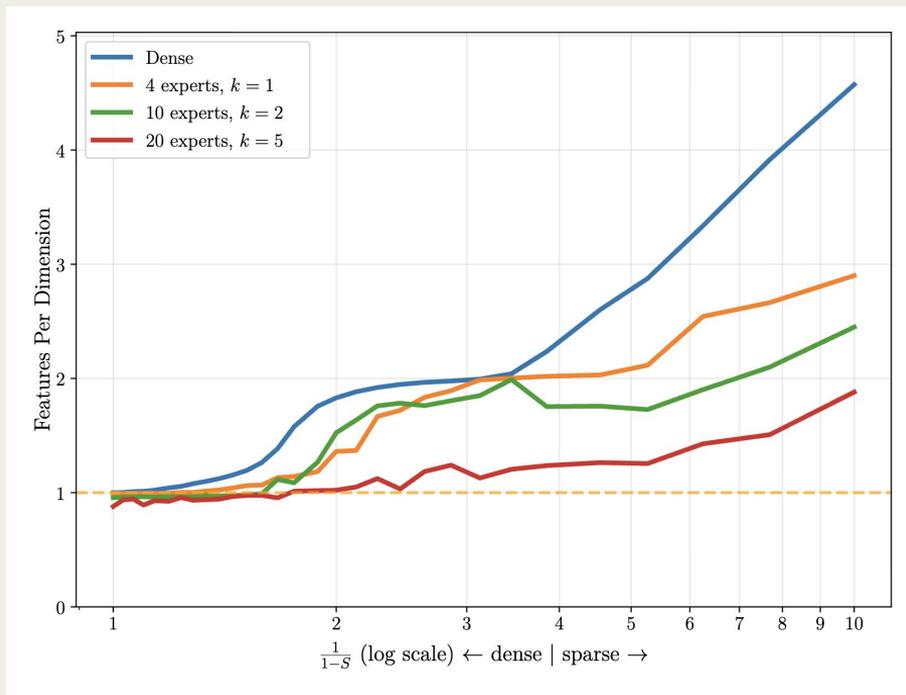
features per dimension for a MoE

$$\frac{1}{k} \sum_{e=1}^E p_e \frac{\|W^e\|_F^2}{m}$$

Frobenius norm

Renormalized gating weight after top-k routing

# Expert sparsity



# Expert sparsity

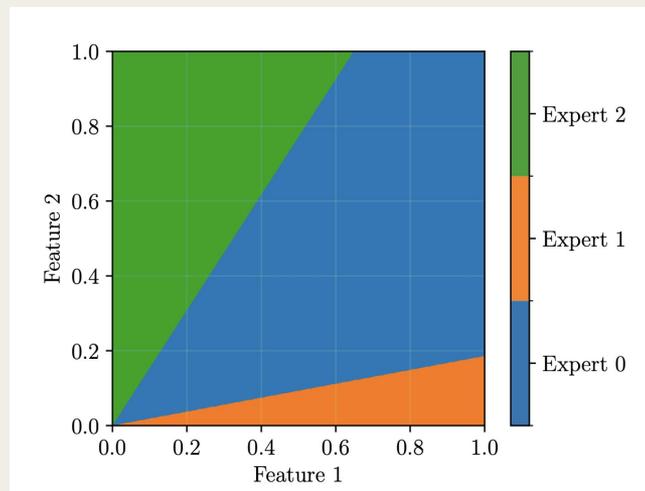
- dense model has higher number of features per dimension than the MoEs
- the *greater* the number of experts, the *less* superposition.

IV.

Expert specialization &  
monosemanticity

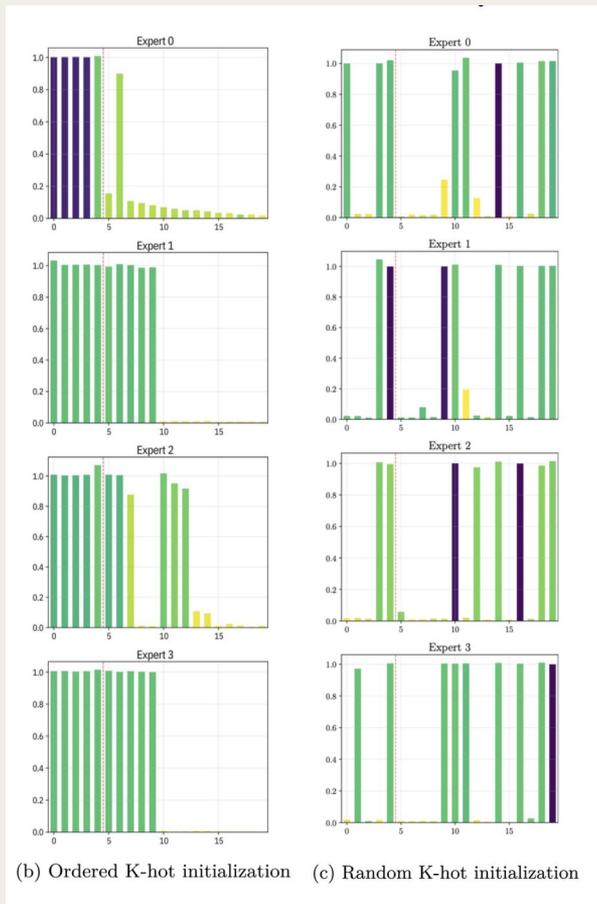
# Expert specialization

- experts partition the input space into subspaces
- experts are specialized to handle specific feature combinations.



# Expert specialization

- Features *initially routed* to an expert become *monosemantically* represented by that expert

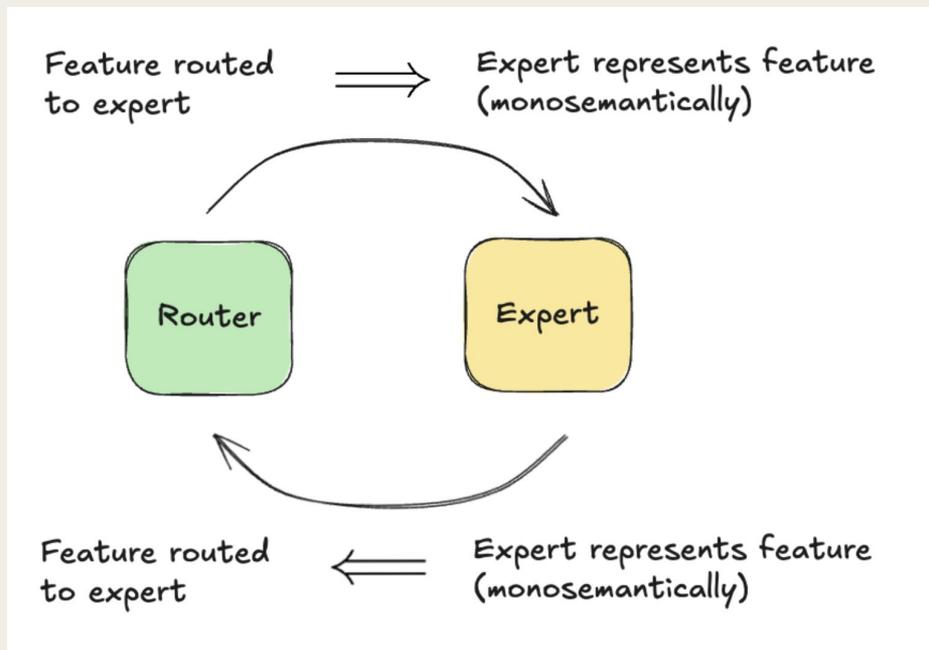


# Expert specialization

- Experts specialize in – and are routed to – their monosemantic features.

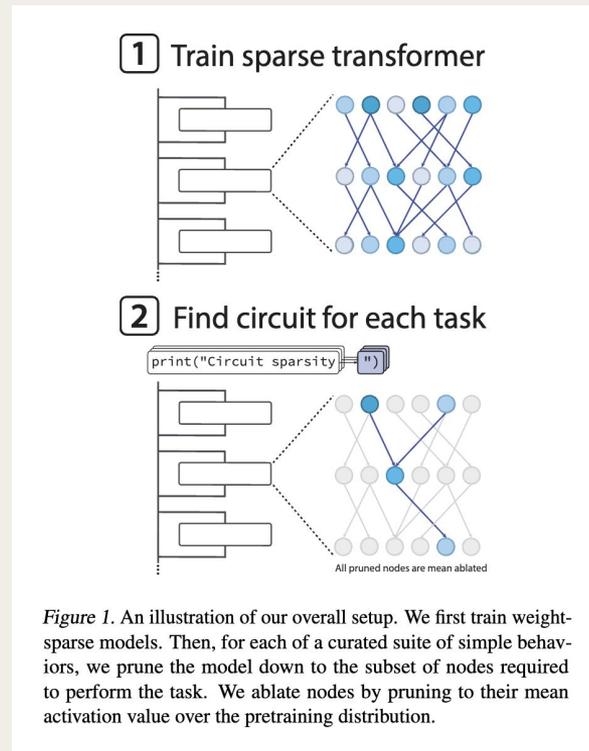
K-Hot Initialization			
Number of monosemantic features per expert	Number of experts (out of 1000)	Mean expert usage (%)	Mean expert usage feature(s) active (%)
0	335	–	–
1	382	10.00	23.94
2	227	10.02	46.61
3	47	10.09	62.00
4	8	9.95	70.30
5	1	9.62	74.79

# Expert specialization



# Circuit sparsity

- preventing superposition during training using weight & expert sparsity



# FlexOLMO

- merges independently-trained experts with no joint training by initializing router with domain embeddings

# What's next

- does geometry of features matter?
- what about multi-dimensional features?
- how do you define faithfulness?
- what's the right unit of computation?

# Binding design

- Design binders that hit target A but avoid B
- But data is sparse & the model overfits

# Binding design

- Are the representations for “bind here” & “avoid here” different or conflated?
- Localize binding logic to specific attention heads & edit it directly rather than retraining.

# Binding design

- train binding MoEs where the experts learn very specific tasks like interface shape, charge complementarity, hydrophobic patches

# References

## Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet

### SPARSITY AND SUPERPOSITION IN MIXTURE OF EXPERTS

Marmik Chaudhari\*, Jeremi Nuer\*, Rome Thorstenson\*

\*Arcadia Research Team, equal contribution

{marmik}@berkeley.edu

---

### Weight-sparse transformers have interpretable circuits

---

Leo Gao<sup>1</sup> Achyuta Rajaram<sup>1</sup> Jacob Coxon<sup>1</sup> Soham V. Govande<sup>1</sup> Bowen Baker<sup>1</sup> Dan Mossing<sup>1</sup>

### MOE LENS - AN EXPERT IS ALL YOU NEED

Marmik Chaudhari<sup>1\*</sup>, Idhant Gulati<sup>1\*</sup>, Nishkal Hundia<sup>2\*</sup>, Pranav Karra<sup>1\*</sup>, Shivam Raval<sup>3</sup>

### A Mathematical Framework for Transformer Circuits

## Toy Models of Superposition

# Further reading

**When Models Manipulate Manifolds: The Geometry of a Counting Task**

---

**Into the Rabbit Hull: From Task-Relevant Concepts in DINO to Minkowski Geometry**

---

---

**Emergence of Hidden Capabilities:  
Exploring Learning Dynamics in Concept Space**

---

Questions

