AlignSAE: Concept-Aligned Sparse Autoencoders

Minglai Yang
UGRA @ CLULAB, University of Arizona
mingly@arizona.edu | https://ymingl.com

Thanks to All Co-authors and Collaborators

Minglai Yang
CS, UofA

Xinyu Guo
CS, UofA

Steven Bethard
InfoSci, UofA

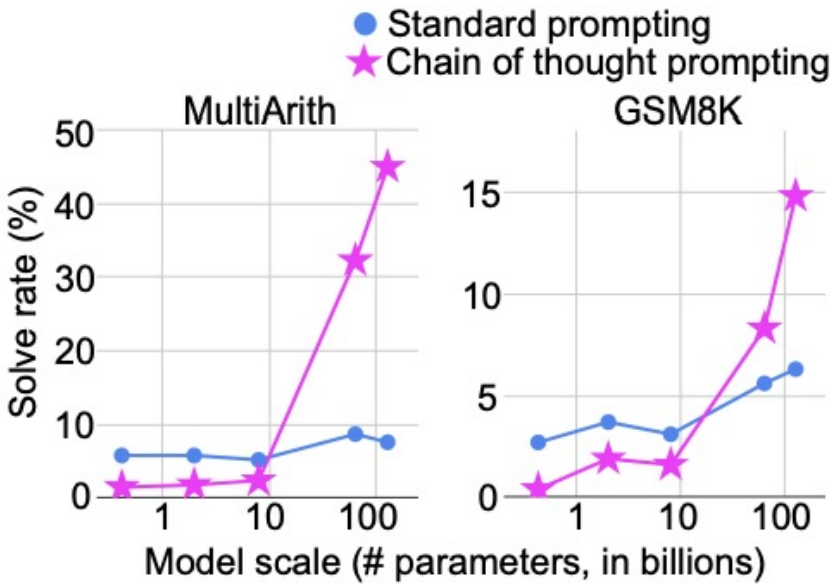
Mihai Surdeanu
CS, UofA

Liangming Pan
CS, PKU

This work wouldn't be possible without them!

Emergent Reasoning Capability

LLMs have exhibited emergent ability to **"reason" like human**



[Wei et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022]

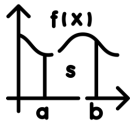
Emergent Reasoning Capability

LLMs have exhibited emergent ability to **“reason” like human**

How reliable is a language model?

Large language models are often unreliable on tasks that require factual knowledge and deliberate, multi-step reasoning.

Factual Recall
Memorization



Compositional
Reasoning

Complex
Decision-Making

Opening the Black Box

- Opaque internals
- Uncertain causes of behavior
- Limited “fine-grained control”
- Hard to verify and trust

How to make LLMs more explainable and controllable?

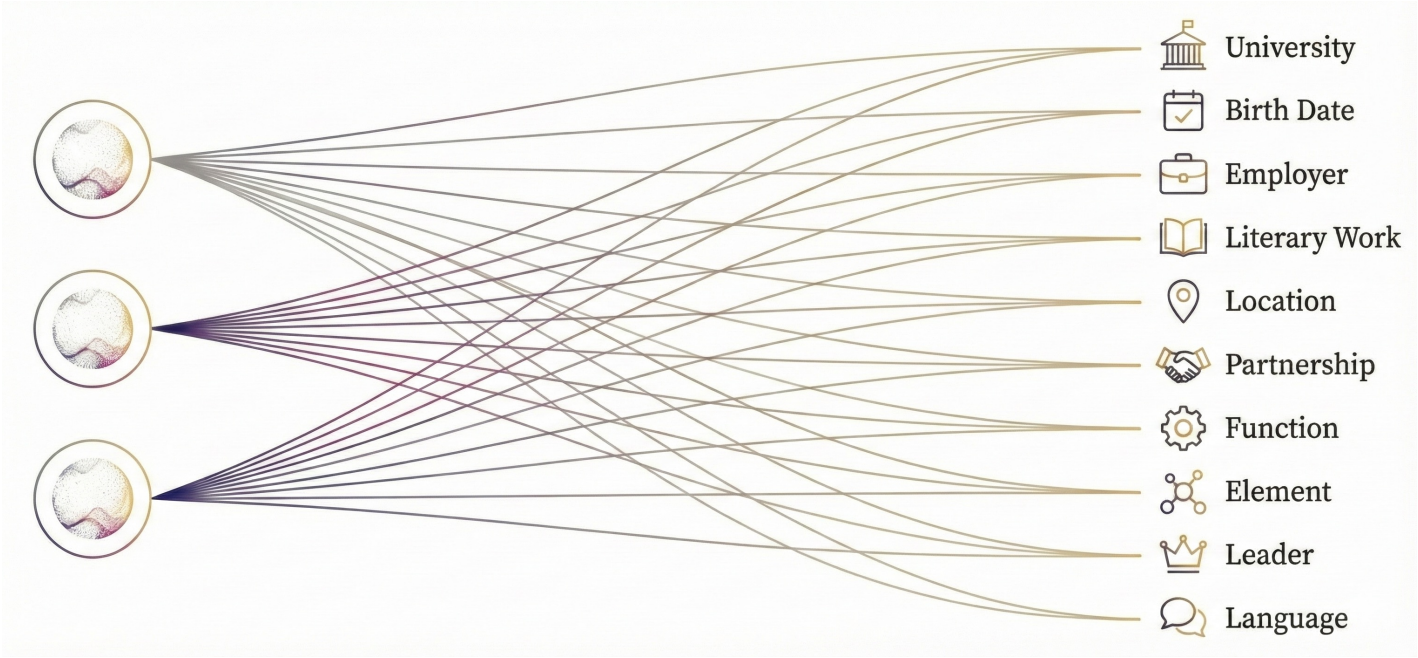
Opening the Black Box

How to make LLMs more explainable and controllable?

Controllable Interface

The Challenge: Unlock the Black Box

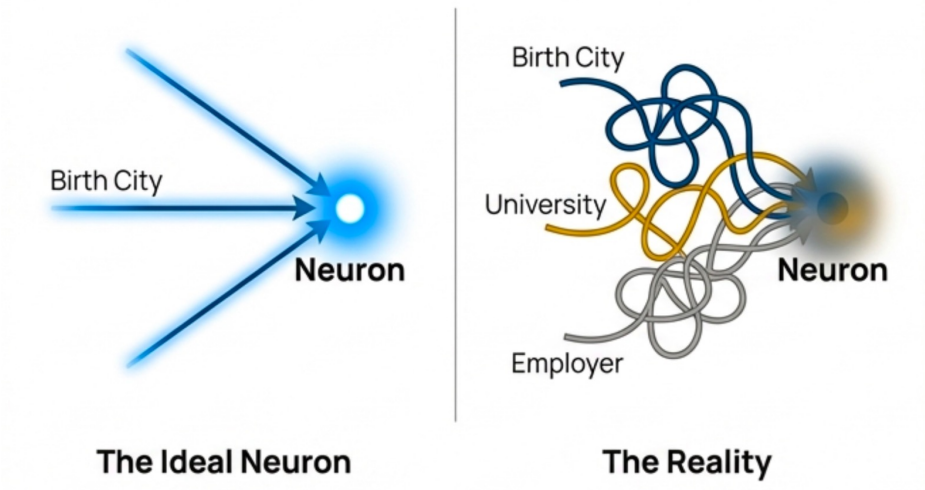
Large Language Models store vast amounts of factual knowledge and reasoning in their parameters, but their internal workings are opaque. Our goal is to create a new system that is interpretable by design.



The Barrier: Superposition

Individual neurons are polysemantic

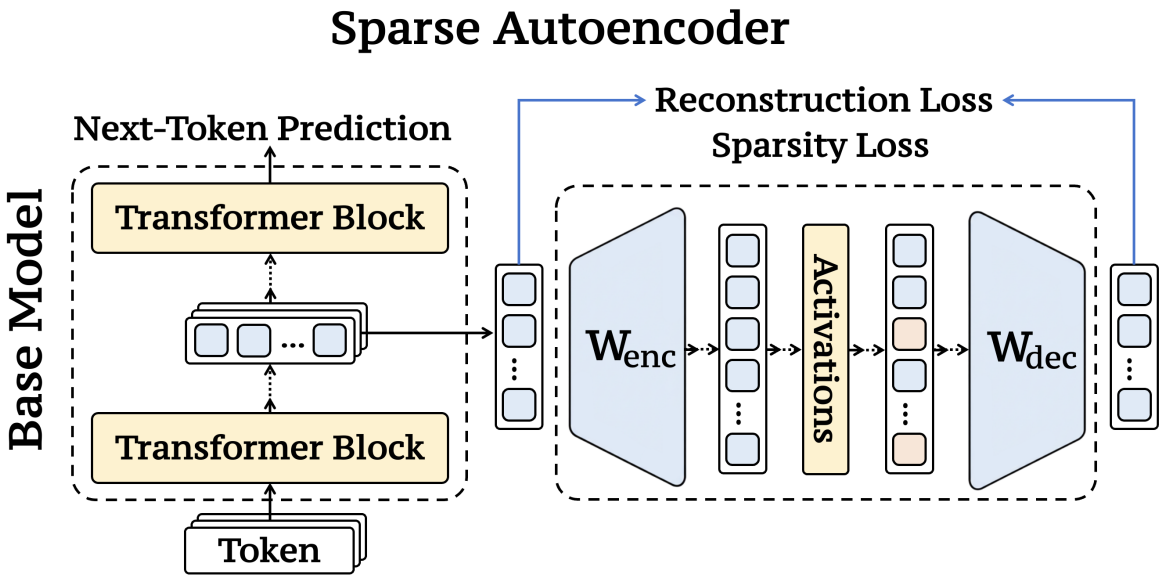
Their activations represent an entangled mixture of distinct concepts, making them impossible to interpret or control reliably.



- Early efforts tried mapping individual neurons to concepts, but this largely failed.
- Neural networks represent more features than they have neurons by encoding concepts as linear combinations across many neurons.

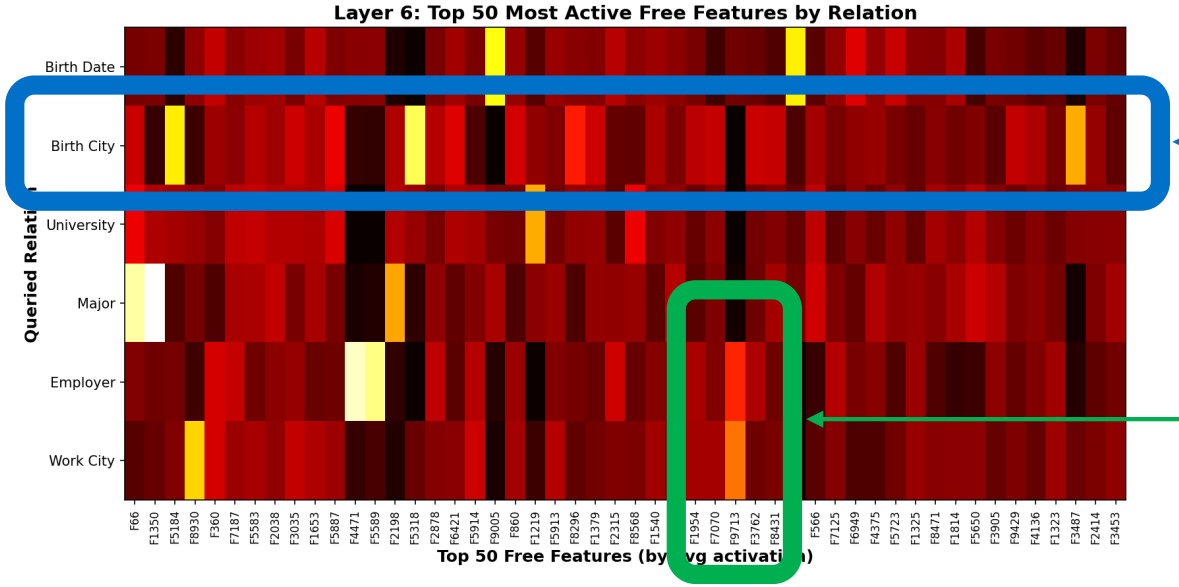
Promising Tool

SAE learns a sparse representation of LM activations. They map the LLM's hidden state into a **higher-dimensional space**, hoping it disentangle the mixed signals into "monosemantic" features.



The Unsupervised Reality: Concepts are Fragmented and Diffuse

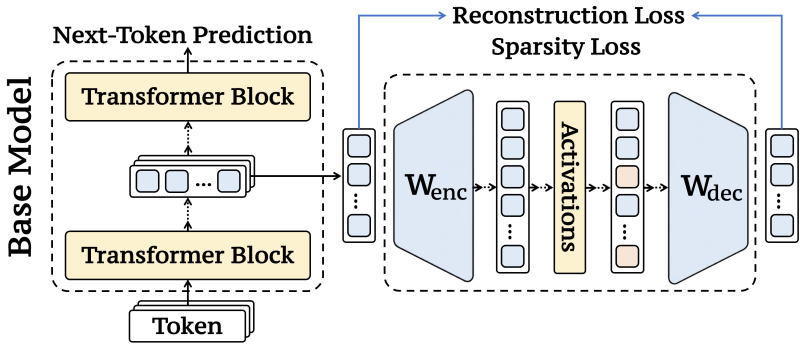
We train a SAE and query for 6 concepts (like ‘BIRTH_DATE’ or ‘EMPLOYER’), the signal for each concept is spread across a few features. No single feature cleanly represents any one concept.



Concept Fragmentation: The signal for “BIRTH_CITY” is scattered across multiple features, making precise intervention impossible.

Feature Entanglement: A single feature activates for multiple, unrelated concepts like ”WORK_CITY" and ”EMPLOYER".

Promising Tool but critical flaw



Standard SAEs are trained with an unsupervised objective (reconstruction + sparsity).

- No Incentive for Alignment
- Interpretation is Difficult
- Features Remain Entangled

This undermines the goal of reliable, feature-level control

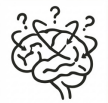
Our Insight: Treat SAE training as LLM training

We re-framed the problem by drawing a parallel to the LM training pipeline. An unsupervised phase builds general capability, but a supervised phase is required for alignment.

LLM Development

Unsupervised Pretraining
Model learns general knowledge and linguistic competence from vast text data.

↓



A powerful but not aligned model

Supervised Posttraining (SFT/RLHF)
Model is finetuned to align with human instructions and preferences.

↓

A strong and helpful assistant

Our SAE Approach (AlignSAE)

Unsupervised Pretraining
SAE learns a general, sparse feature dictionary focused on reconstruction.

↓

A powerful but not aligned feature space

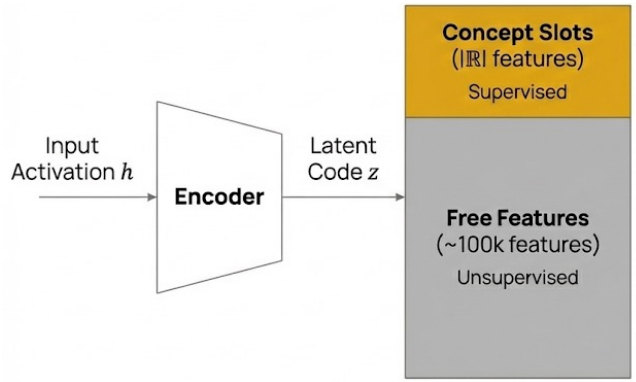
Supervised Posttraining (Binding Loss)
SAE finetuned with concept supervision to bind specific features to an ontology.

↓

A controllable and interpretable interface

Dedicated Slots + Target Objectives

A Novel Hybrid Architecture



We partition the SAEs' latent features, dedicating **a small, supervised set for interpretability** while a large, unsupervised bank preserves reconstruction fidelity.

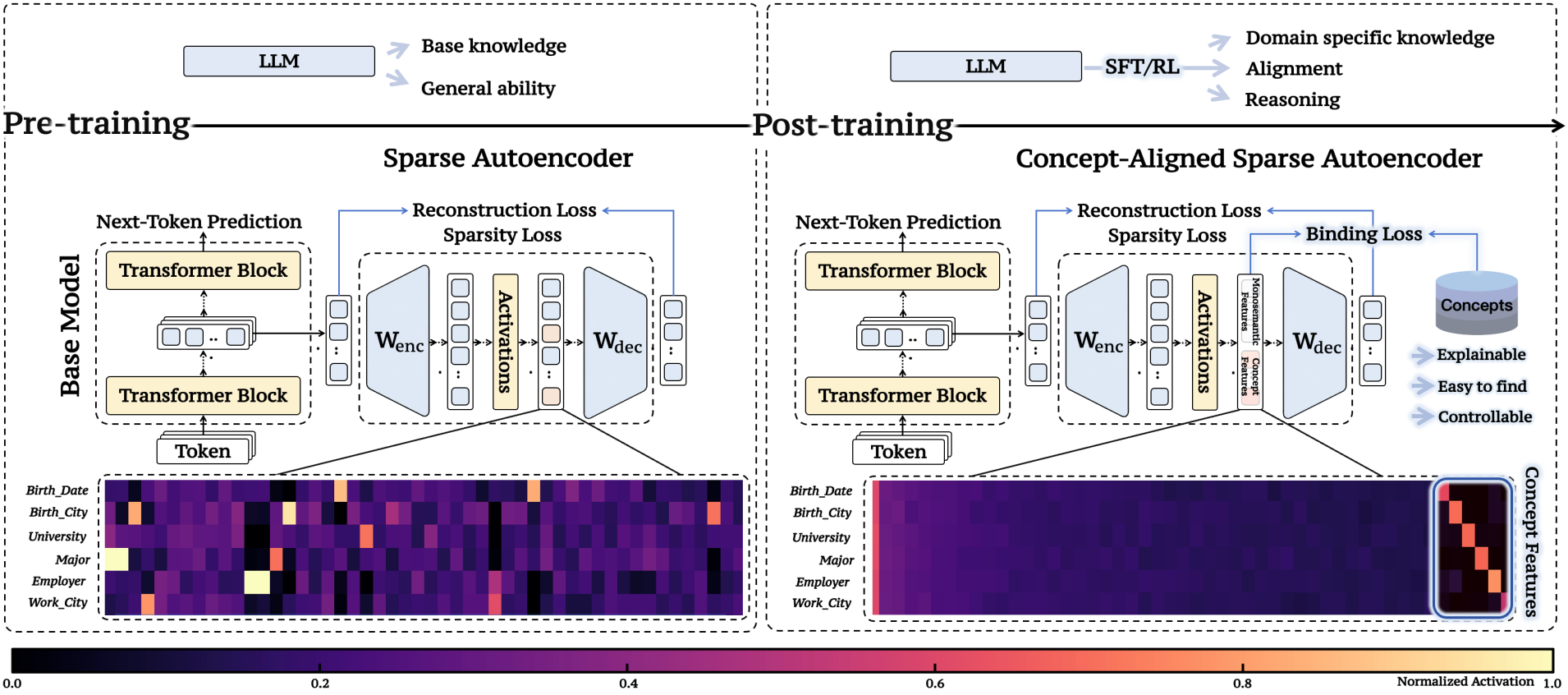
New Training Objectives to Enforce Alignment

L_{bind} (Binding Loss): Forces a one-to-one mapping between each concept and its dedicated slot.

$L_{\perp}$ (Invariance/Orthogonality Loss): Makes concept features invariant to irrelevant details and decorrelates them from the free features, preventing information leakage.

L_{val} (Sufficiency/Value Loss): Ensures the concept slots are sufficient to predict the correct answer, creating a useful bottleneck.

AlignSAE's Two-Phase Curriculum



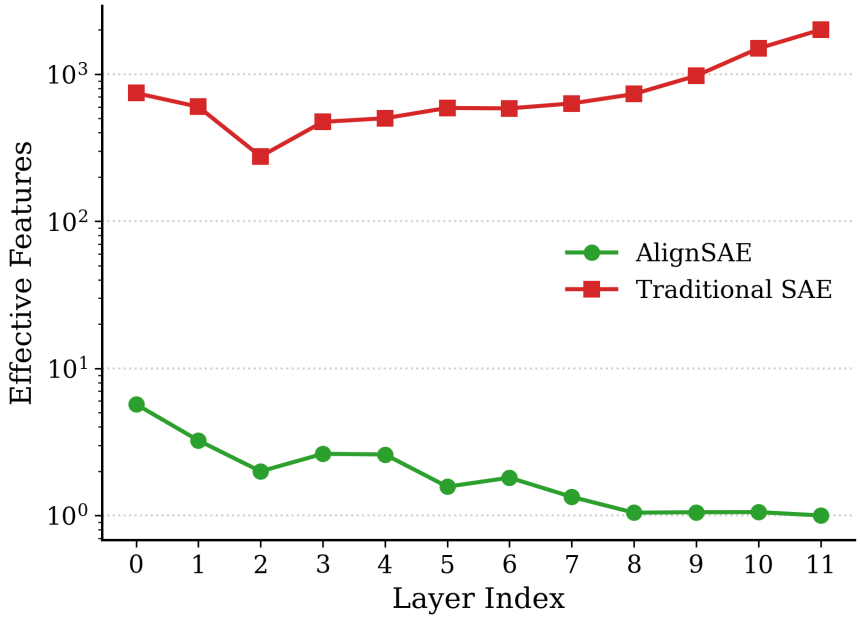
The result is a clean, verifiable, one-to-one mapping between concepts and features.

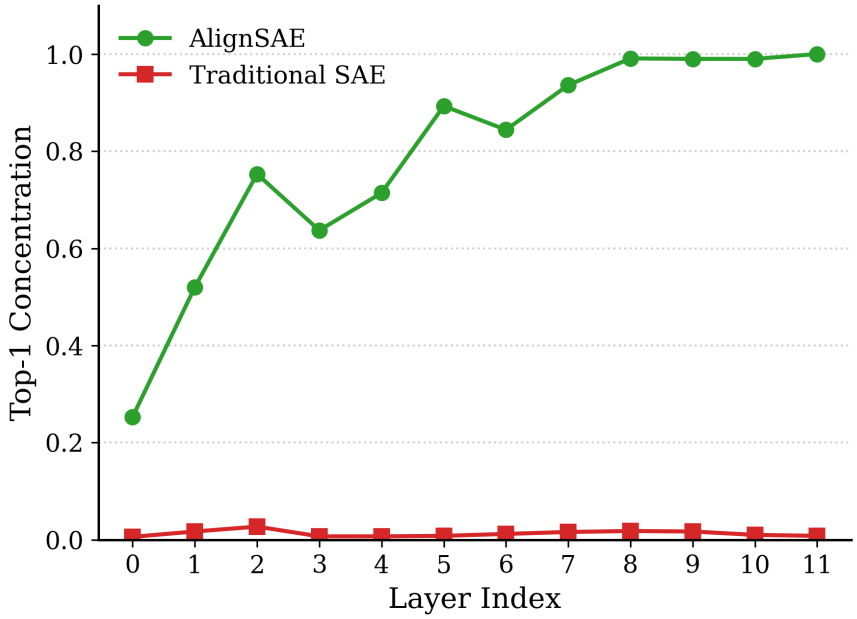
Feature Fragmentation

SAE post-training concentrates each concept onto a single feature.

$$A_{c,k} = \mathbb{E}_{i:c(i)=c}[z_{i,k}] \qquad B_{c,k} = \frac{A_{c,k}}{\sum_{k'} A_{c,k'} + \epsilon}.$$

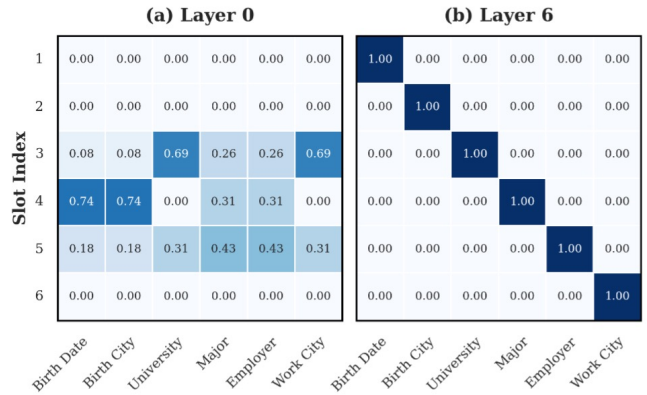
$$\text{EffFeat}(c) = \exp\Big(-\sum_k B_{c,k} \log B_{c,k}\Big) \downarrow \qquad \text{Top1Conc}(c) = \max_k B_{c,k} \uparrow$$



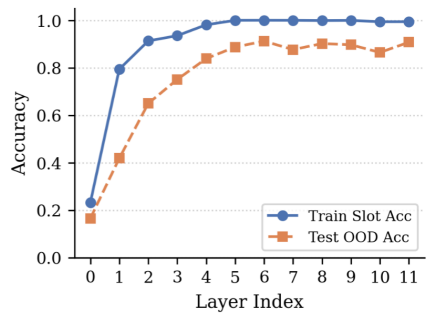


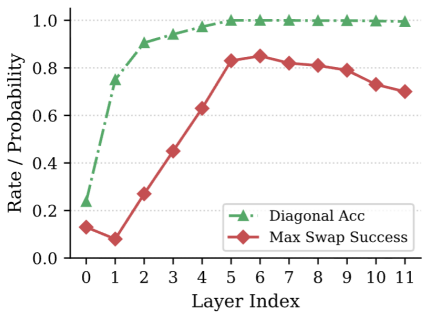
Optimal Sweet Spot

Concept alignment is best in the middle layers of the transformer.



Metric	Layer 0	Layer 6	Δ (L6 – L0)
Diagonal Accuracy	0.238	**1.000**	↑ 0.76
Swap Success	0.040	**0.850**	↑ 0.81
Train Slot Acc	0.232	**1.000**	↑ 0.77
Test Unseen Acc	0.165	**0.912**	↑ 0.75
Recon MSE	6.53×10^{-5}	7.42×10^{-2}	↑ ≈ 1.1k×





Early layers lack the necessary abstraction, while deeper layers can be overly compressed for the final task, hindering clean reconstruction. Layer 6 provides the optimal balance.

From Interpretation to Intervention: The Promise of Causal Control

Because *AlignSAE* binds each concept to a specific, isolated feature slot, we can now test for true causal control.

1. Query
"When was person X born?" (targets `BIRTH_DATE`)

↓

2. Find Activation
Get hidden activation at Layer 6.

↓

3. Get Features
Use `AlignSAE` to get sparse feature vector.

↓

4. Intervene
Manually activate the feature slot for a different concept deactivate original concept.

↓

5. Proceed
Let the model's computation proceed with this modified activation.

The Promise of Causal Control

Does the model's final output change to match the concept we activated?

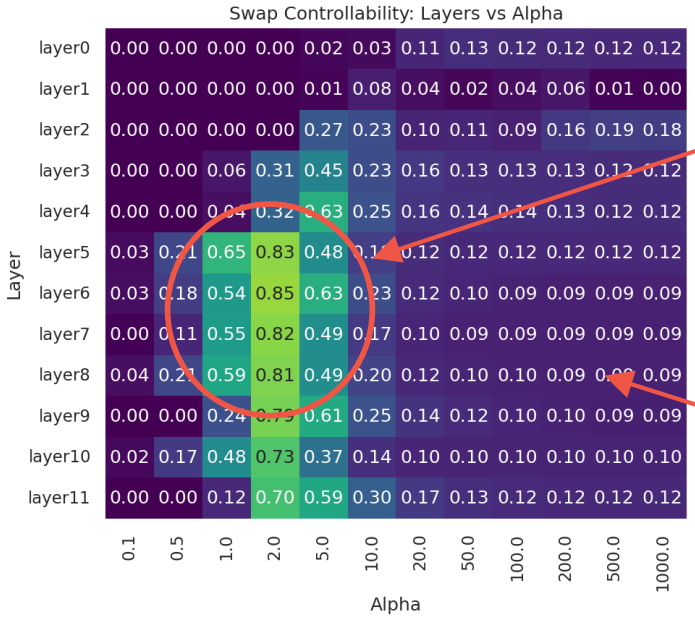
By simply activating a single feature slot, we can reliably steer the model's output to a different factual concept, with minimal side effects.

Original Query (Targeting Concept A)	Intervention (Activate Slot B)	Model's Generated Output
What is Grace Wendy Rivera's **work city**?	→ UNIVERSITY	**Florida International University**
Where did Thomas Heath Stafford go to **college**?	→ BIRTH_DATE	**2, March, 1981**
When was Megan Kian Valencia **born**?	→ WORK_CITY	**Framingham, MA**
What was Jennifer Pruitt's **major**?	→ UNIVERSITY	**University of Wisconsin-Madison**

Mapping the Controllability Landscape

Does the model's final output change to match the concept we activated?

Swap success depends on the intervention layer and the amplification strength (α). There is a clear operating range where control is most effective.



Optimal Control Zone:
Up to 85% success rate at Layer 6 with a=2.0.

Over-amplification:
Too much force destabilizes the model's output

Even Failures Are Informative

What happens when a concept swap doesn't produce the exact correct answer?

The model still generates an answer of the correct semantic type, The intervention successfully steers the model onto the right topic.

Target swap	α=2 (Error 15%)			α=10 (Errors 77%)		
	Same	Diff	Same %	Same	Diff	Same %
birth_city	36	20	64.3	91	50	64.5
birth_date	20	0	100.0	139	0	100.0
employer	8	1	88.9	134	3	97.8
major	1	0	100.0	77	51	60.2
university	42	10	80.8	101	1	99.0
work_city	9	7	56.2	94	25	79.0
Overall	**116**	**38**	**75.3**	**636**	**130**	**83.0**

75.3% of failed swaps still produce an answer in the correct category at Layer 6 with moderate amplification (α=2).

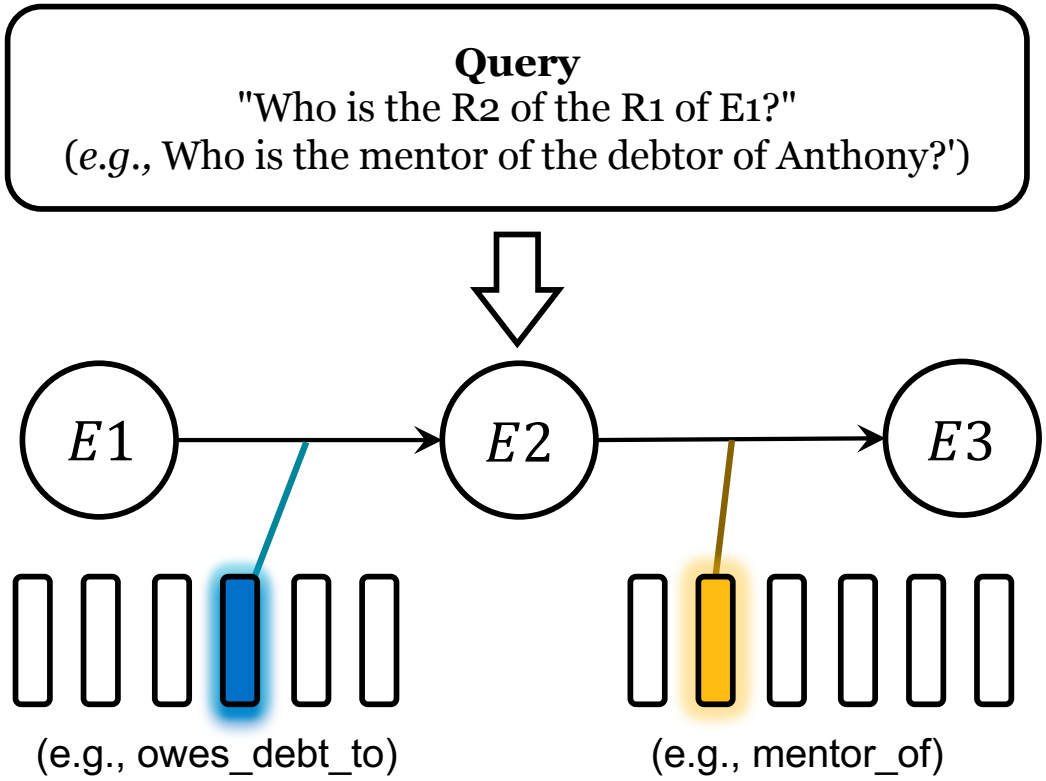
The entity is wrong, but the answer type is correct (a plausible major).

Swap	**Q:** Where did Jesse Kian Tate go to college? **Original:** UNIVERSITY → **Swap to:** MAJOR
Outputs	**Baseline:** Rochester Institute of Technology **Gold target:** Physical Therapy **Generated: Geography** *(type ✓ entity ✗)*

Two-hop Reasoning

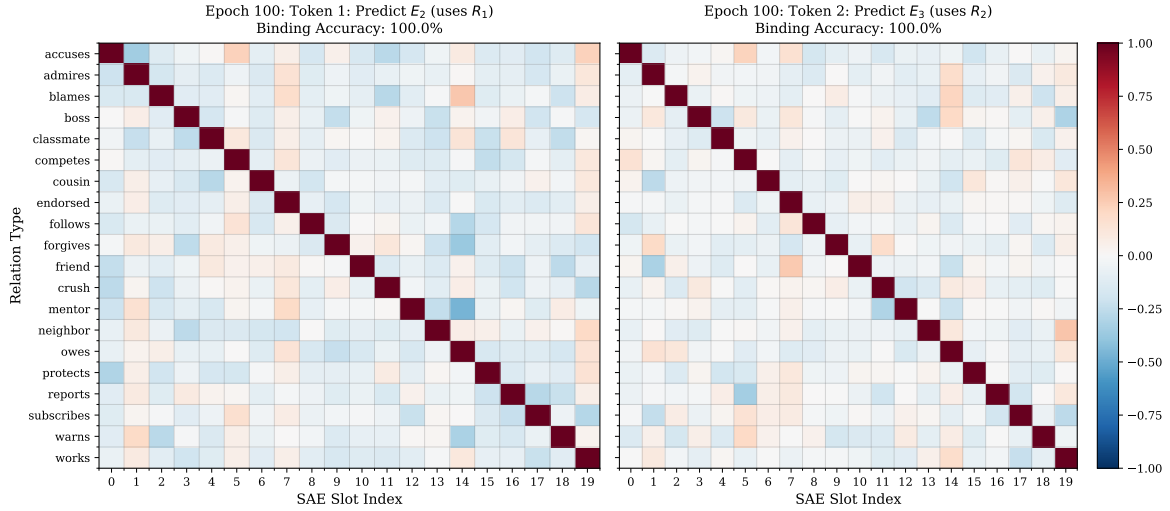
Can *AlignSAE* provide insight into more complex, sequential reasoning processes?

The AlignSAE concept slots should **activate sequentially**, tracking the reasoning process.



Visualizing the Chain-of-Thought

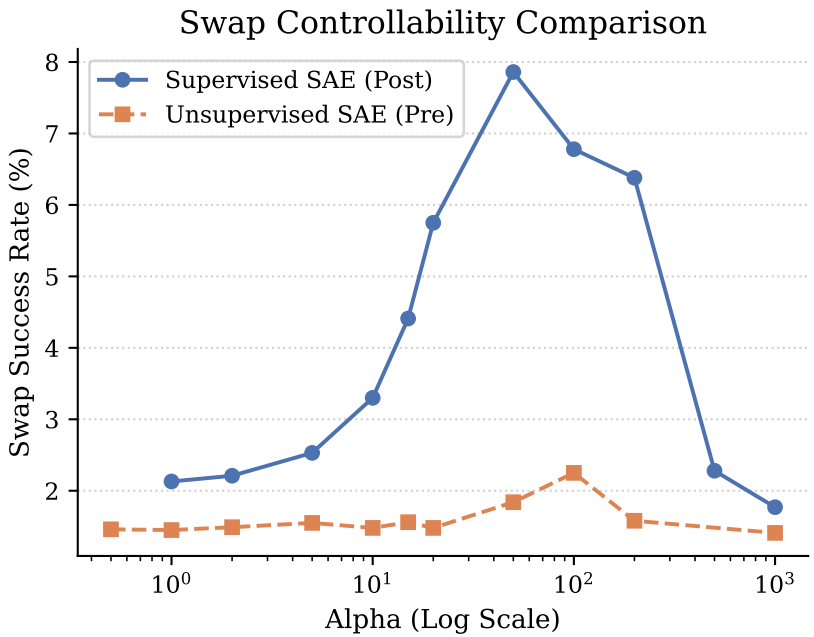
The concept slots perfectly track the active relation at each step of the reasoning chain.



AlignSAE provides a dynamic, step-by-step trace of the model's internal reasoning process, making complex computations more transparent.

Concept Swapping Analysis

Post-trained *AlignSAE* enables substantially higher swap success than an unsupervised SAE baseline.



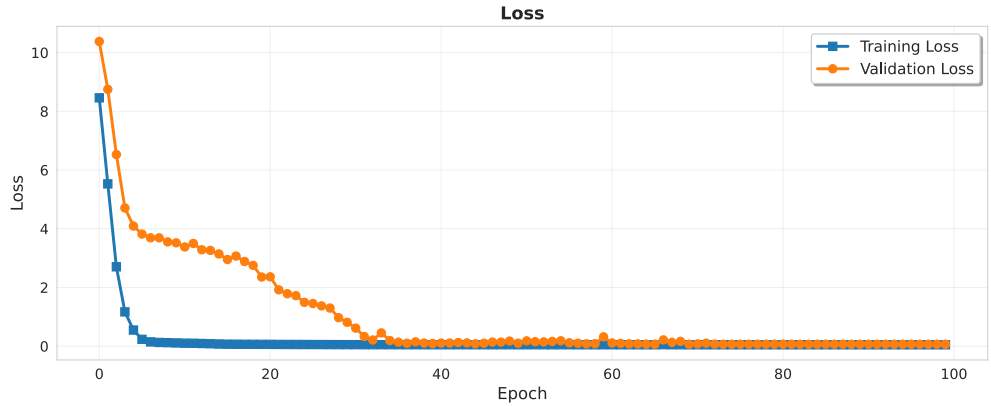
Swap success peaks at moderate intervention strengths (mid-range α), indicating predictable controllability.

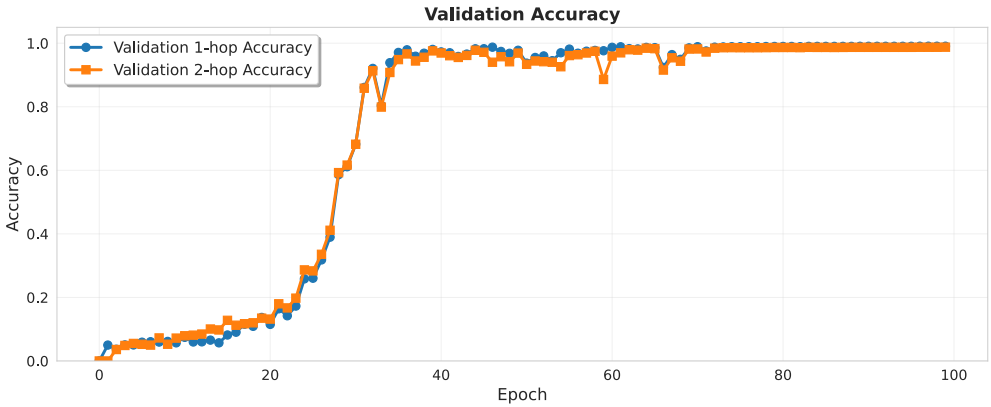
Understanding Grokking - An Abrupt Leap

In complex reasoning tasks, models sometimes exhibit a "**grokking-like** emergence." After a long period of stagnation, validation accuracy 0% suddenly jumps to near-perfection, suggesting a fundamental shift in the model's internal strategy.

What happens "inside" the model during this transition?

How does it reorganize its knowledge to achieve this leap?

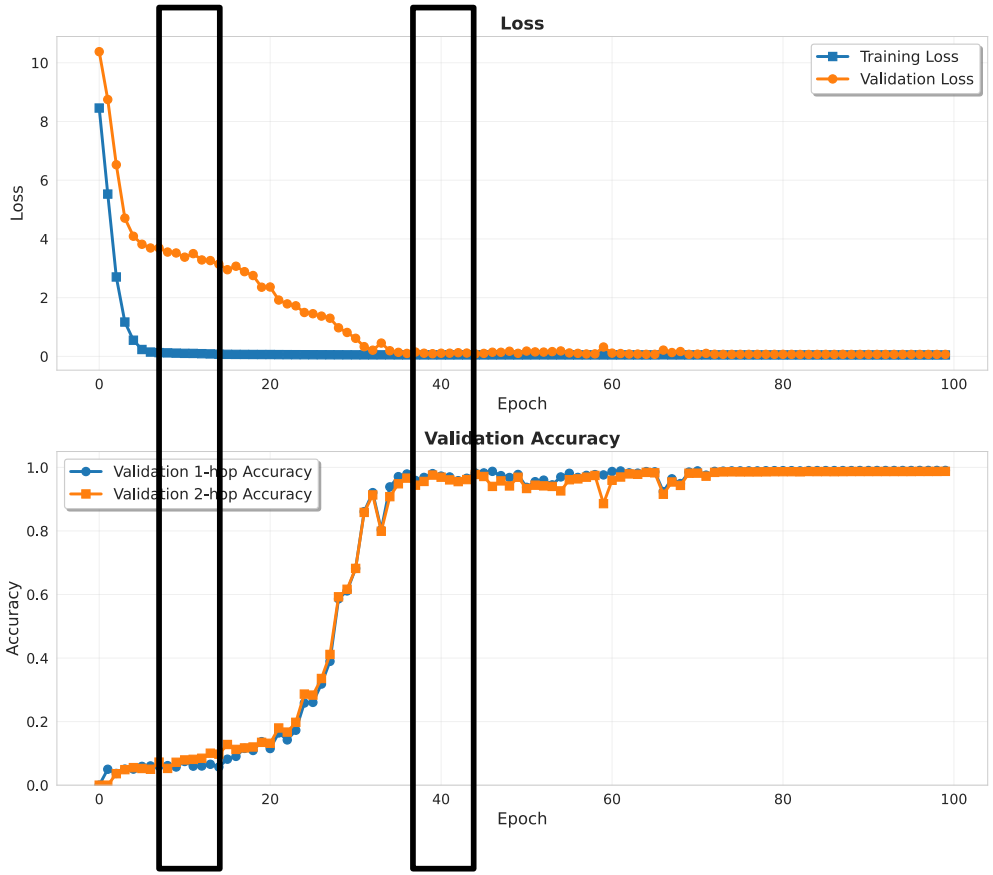




Understanding Grokking

AlignSAE is the right tool because:

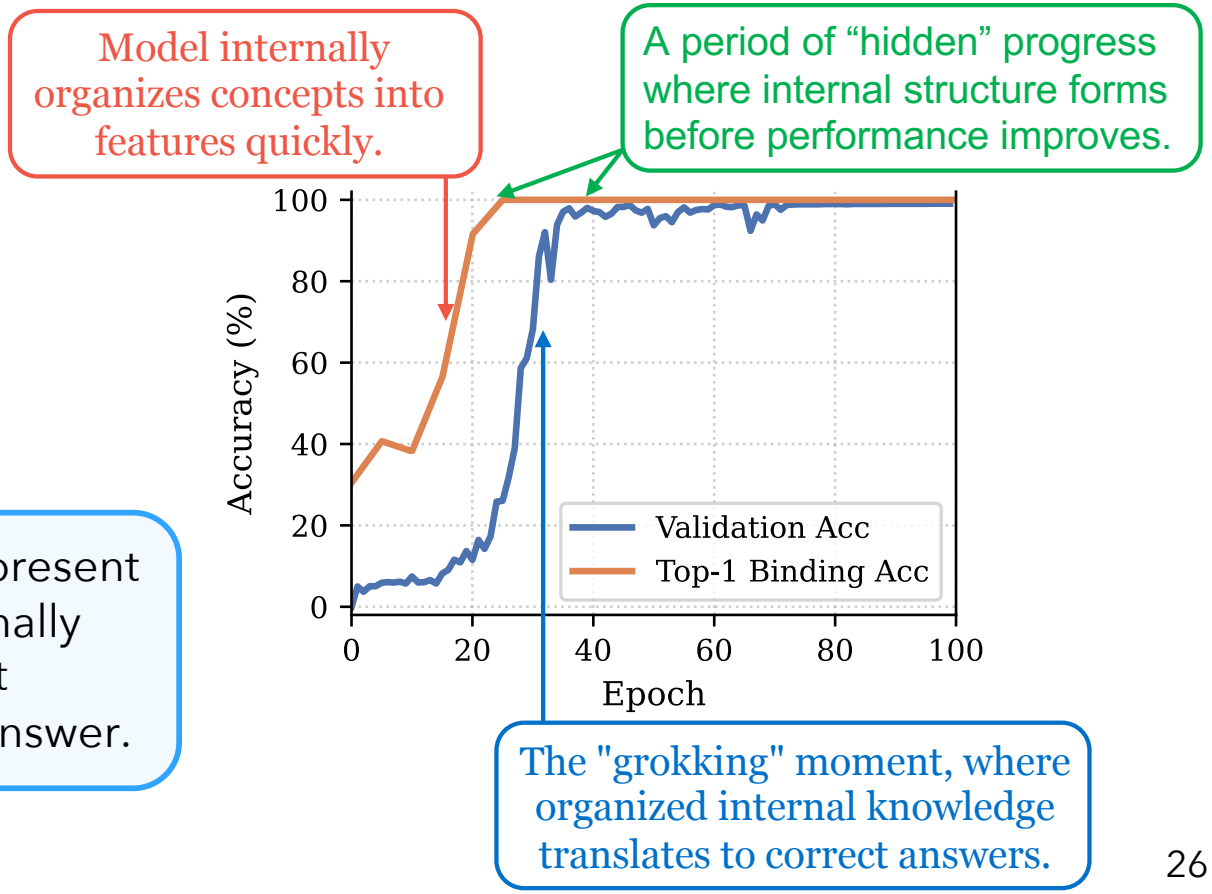
- **Localization**: It forces the model to consolidate its knowledge about a concept into a single, addressable feature, rather than spreading it across the network.
- **Verification**: We can see exactly which relation the model is 'thinking about at each step by checking which slot activates.



The Lag Between Knowing and Showing

AlignSAE reveals a clear separation between the model's internal organization and its external performance.

The model first learns how to represent the compositional steps internally **before** it learns to use that representation to get the right answer.

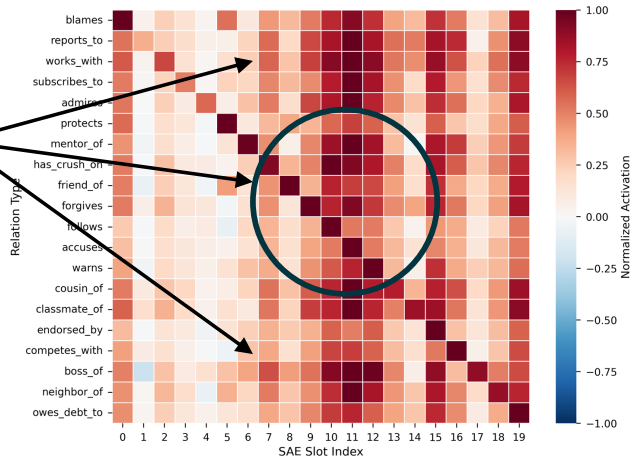


From Entangled Mess to Order

Phase 1: Entangled & Memorized

Pre Grokking (Epoch 10) – Token 1

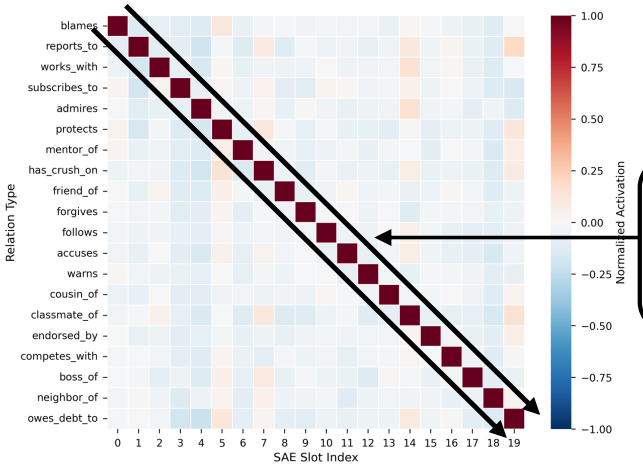
Substantial Off-Diagonal Mass



Representations are **confused**. The signal for the first hop ('R1') is dispersed across many unrelated slots. The model has not isolated the logical steps.

Phase 2: Organized & Compositional

Post Grokking (Epoch 40) – Token 1

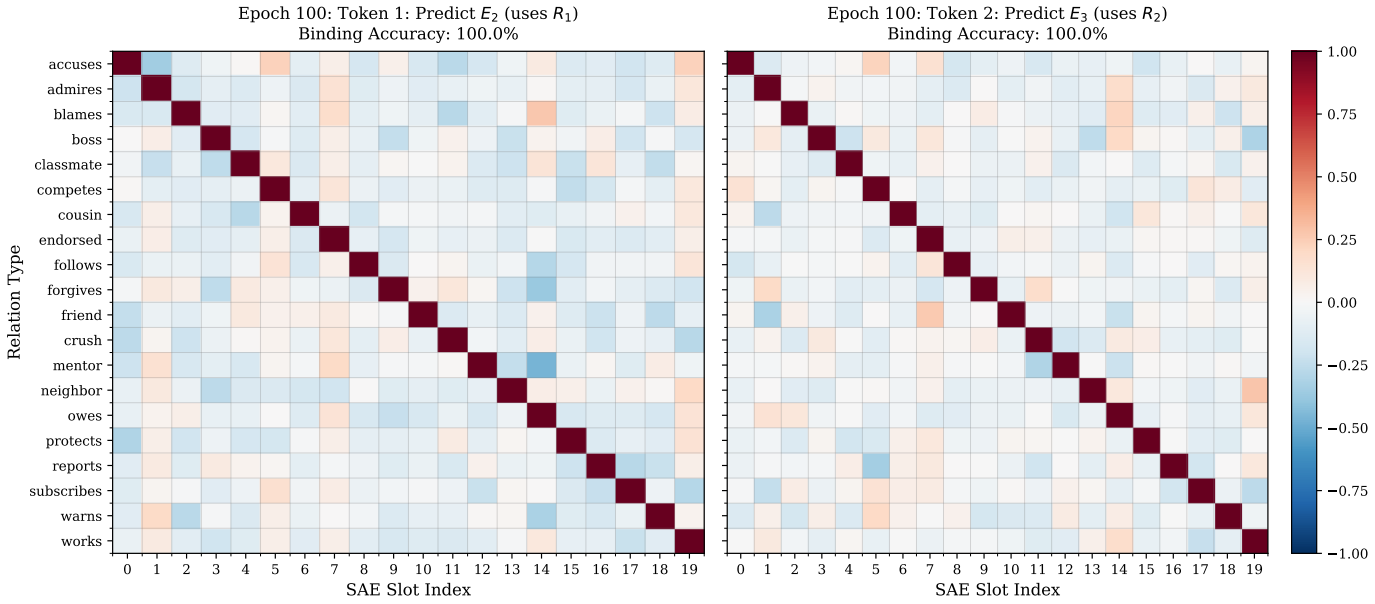


A Near Perfect Diagonal

Representations have **crystallized**. The binding is near perfect. The slot for Relation 1 activates exclusively and correctly.

Step-Wise Alignment Post-Grokking

After grokking, the model has a compositional representation. It activates the correct relation slot for each sequential step of the reasoning task.

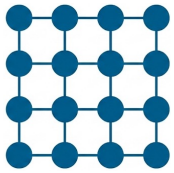


Grokking is the Crystallization of Internal Structure

Our investigation, using *AlignSAE* as a probe, suggests that grokking in 2-hop reasoning coincides with a phase transition inside the model.

From:

A “messy” state where hop-specific information is entangled and distributed likely relying on memorization.



To:

A “clean” state where knowledge becomes structured and compositional. The model consolidates its understanding into distinct, addressable features for each step of the reasoning chain.

This consolidation of "**compositional, slot-addressable relational features**" is what enables the sudden jump in generalization and performance. The model learns not just the answers, but the algorithm.

An Analogy: Organizing a Disordered Library

Pre Grokking

The information (books) is present, but unsystematized. Finding a fact relies on brute-force search (memorization).

The Grokking Transition

A system is created. Specific concepts (genres) are assigned to specific locations (shelves/slots).

Post Grokking

The system is complete. To find 'Science, you go directly to the 'Science' shelf. The process is efficient and reliable.

Takeaways

- AlignSAE introduces a **'pre-train, then post-train'** curriculum to align SAE features with human concepts.
- This creates a **clean, one-to-one** mapping in the middle layers of a frozen LLM, transforming a diffuse feature space into an interpretable one.
- The aligned feature slots act as reliable causal control knobs, enabling precise **concept swaps** at inference time.

Thank You!

mingly@arizona.edu

http://ymingl.com