RESEARCH & INNOVATION FOUNDATION

THE CYPRUS INSTITUTE
RESEARCH•TECHNOLOGY•INNOVATION

TensorICE

New Directions in Interpreting Neural Networks

Andreas Demou

The Cyprus Institute

AQTIVATE Conference, October 1, 2026

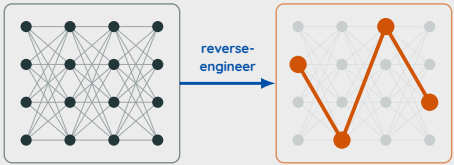
Project funded by the Cyprus Research and Innovation Foundation under contract No. EXCELLENCE/0524/0407

Mechanistic Interpretability

"We built it. We use it. We don't understand it."

The goal

- Reverse-engineer the algorithms a neural network has learned
- Break the network down into human-readable concepts and circuits

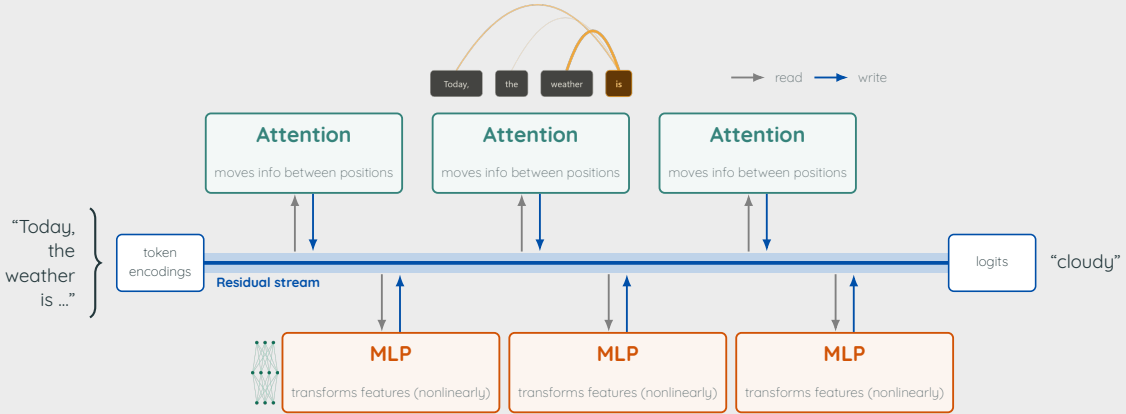


Why it matters

- **Safety & alignment:** detect deception, bias, or unsafe reasoning before deployment.
- **Science:** does the network we trained adhere to the laws of physics? (e.g. mass and energy conservation)
- **Trust:**
  - Reasoning models don't always say what they think: anthropic.com/research/reasoning-models-dont-say-think
  - Chain-of-Thought is not explainability: fbarez.github.io/assets/pdf/Cot_Is_Not_Explainability.pdf

Transformer Refresher

Every block reads from and writes to one shared stream



Key insight: the residual stream is a communication bus where concepts and features live.

Visualizations of actual models: bbycroft.net/llm

Features & Superposition

Linear representation hypothesis

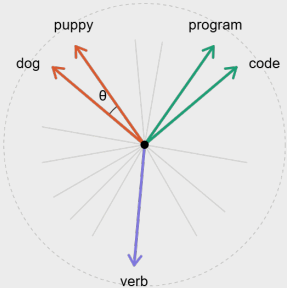
- Words (or features) are encoded as vectors
- Vectors → directions in an activation space of ∼ 10,000 dimensions
- “King” − “man” + “woman” = “Queen”
- But words ≫ 10,000 dimensions

Superposition

- Models pack **more features than dimensions** by using non-orthogonal directions
- Consequence: neurons are **polysemantic**; they fire for multiple unrelated concepts

Problem for Mech. Interp.

- Find a way to **unmix** the features
- From polysemantic to **monosemantic** neurons



similar meanings → near-parallel vectors — small θ = high cosine similarity

Learning a “Dictionary”

x → input
The activation vector (what the model computed at some layer)

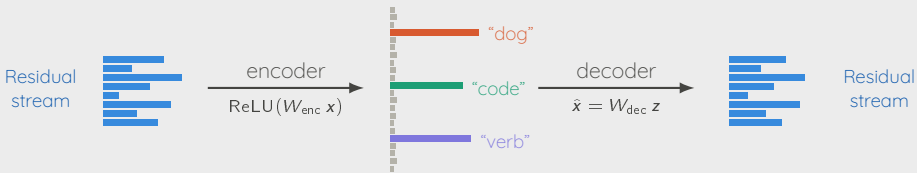
d_i → dictionary atoms
Learned feature directions. One atom per concept.

c_i → sparse coefficients
How active each feature is. Most are zero on any given input.

$$x \approx \hat{x} \equiv b + \sum_i c_i \, d_i$$

signal ≈ sparse mixture of atoms (few nonzero c_i)

Sparse Auto-Encoder (SAE)



x dense activation (dim $d \approx 4096$)

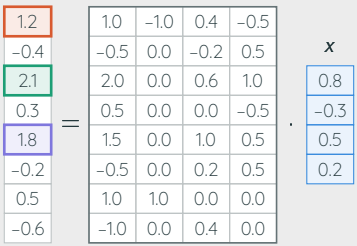
z sparse code (dim $D \gg d$, few nonzero)

$\hat{x}$ reconstruction ($\approx x$)

Mathematical operations of an SAE

Encoder

$u = W_{\text{enc}}\, x$ (bias b_{enc} omitted for clarity)

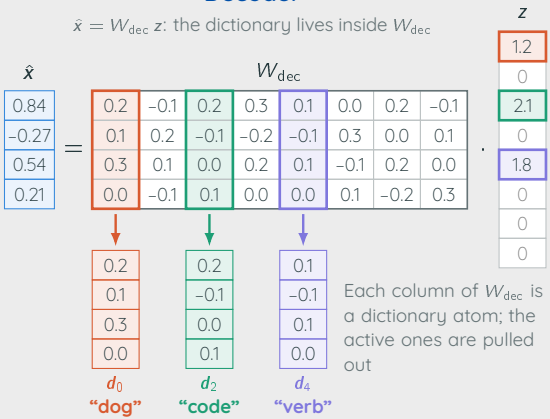


Apply TopK ($k = 3$): keep the 3 largest entries of u, zero the rest → sparse code z

The coloured cells of u survive into z with the same values; the others become 0.

Decoder

$\hat{x} = W_{\text{dec}}\, z$: the dictionary lives inside W_{dec}



Each column of W_{dec} is a dictionary atom; the active ones are pulled out

What SAE Features Look Like

Features we actually find

Token / n-gram features
"closing HTML tag", "French article le/la"

Entity features
"Golden Gate Bridge", "Python the language"

Behavioural features
"sycophancy", "refusal", "sarcasm"

Abstract concept features
"code vs prose", "person's gender", "deception"

Known limitations

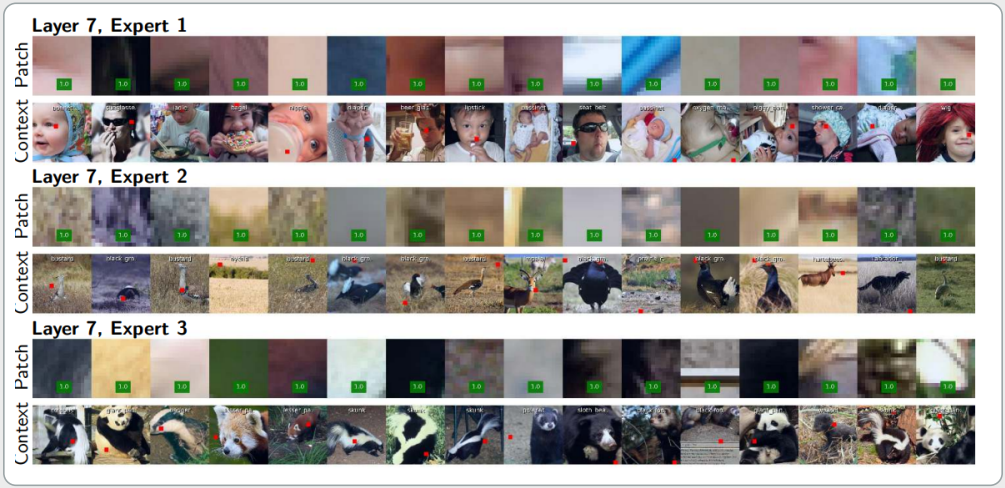
Dead features
Many dictionary atoms never activate: wasted capacity.

Feature absorption / splitting
One concept spread across many atoms (or one atom swallows related concepts).

Additive assumption
Reconstruction is linear: can't separate "star+coffee" from co-occurrence.

SAEs are powerful but imperfect. Every limitation is an active research frontier.

Features in Vision Models



Steering examples

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.
```

Default output chooses Physics

```
Human: If you had to pick one, what's the most interesting science? Be brief.

Assistant: Physics
```

with Brain sciences clamped to 10× its max F#34M/9493533

Changes answer to neuroscience

```
Human: If you had to pick one, what's the most interesting science? Be brief.

Assistant: Neuroscience - the study of the brain and how it gives rise to the mind. Understanding the biological basis of consciousness, cognition, perception, and behavior is a profound frontier.
```

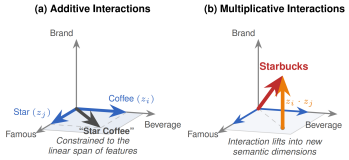
Extension (1): PolySAE, compositional features

Koromilas et al., ICML 2026

The problem: SAEs have linear decoding.

"**star**" and "**coffee**" cannot be combined into **"Starbucks"**

➥ Need **non-linear decoding**



Standard SAE

$$\hat{\boldsymbol{x}} = \boldsymbol{b} + \sum_i c_i \, \boldsymbol{d}_i$$

Features **combine additively**. To represent a compound, the model must **allocate a new atom**.

PolySAE (polynomial decoder)

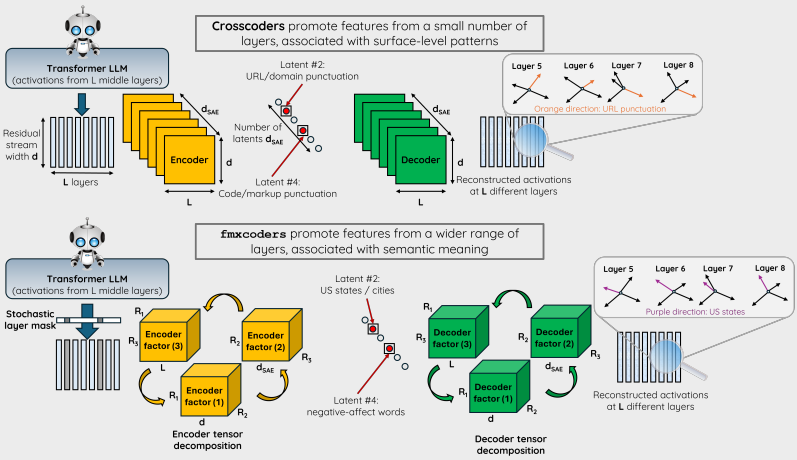
$$\hat{\boldsymbol{x}} = \boldsymbol{b} + \sum_i c_i \, \boldsymbol{d}_i + \sum_{i,j} c_{ij} \, \boldsymbol{d}_i \otimes \boldsymbol{d}_j$$

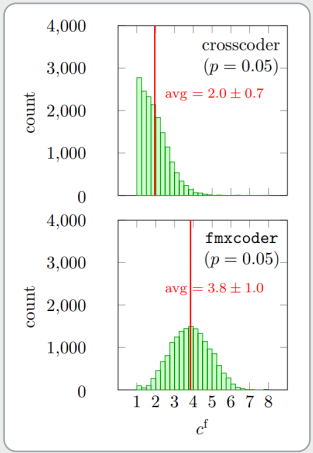
Adds **interaction terms** → composition rather than simple addition. Captures compositional structure explicitly, with fewer atoms.

Poly Feature A	Poly Feature B	Context	Vanilla SAE
[Italian, Italy]	[gang, mob]	Details obtained by the Guardian reveal extent to which **Sicilian** mafia clans are migrating north after running into financial problems in Italy.	[State, ISIS]

Extension (2): fmxcoders, multilayer activations

Demou et al., NeurIPS 2026 (spotlight)





SAEs on language models vs. CFD surrogates: what changes

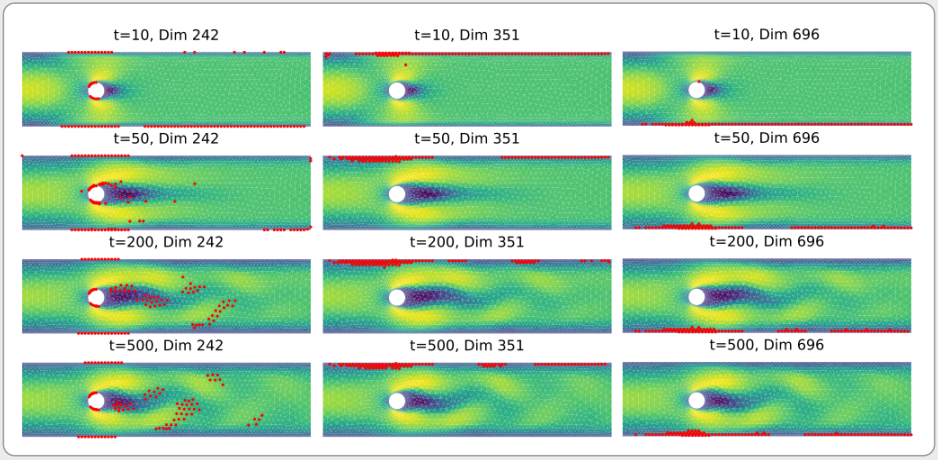
	Language models	CFD models & data
What is a sample?	**Tokens** (words, symbols) from a **discrete** vocabulary	**Patches** of mesh nodes or grid points, **continuous representation**[3]
What features look like	**Discrete concepts**: words, entities, behaviours (refusal, sarcasm)	**Continuous flow structures**: wake, boundary layers, vortices[1] [2]
How features are labelled	**Top-activating text examples**, explained automatically by an LLM	**Maps on the mesh**, correlation with vorticity or enstrophy[4], frequency spectra, probes on labelled events[3]
Ground truth	**Subjective judgement** of whether a label fits	**Agreement with physical fields**: e.g. overlap with high-vorticity regions[1], conservation residuals[3]

In fluids, physics supplies vocabulary and ground truth, but features are continuous and can fade over time.[4]

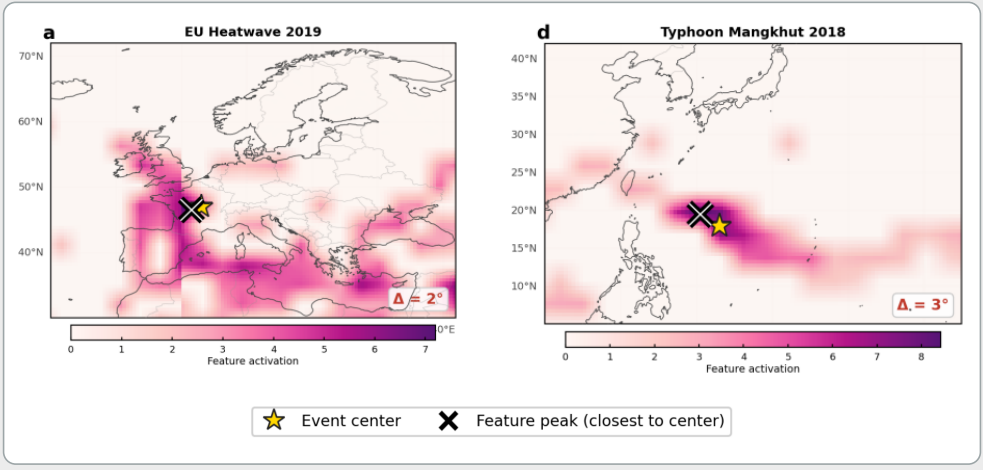
[1] Hu & Liu, arXiv:2507.16069 [2] Hu, Glatt & Liu, arXiv:2604.04946 [3] MacMillan & Ouellette, arXiv:2512.24440 [4] Rosenfeld & Sonnewald, arXiv:2606.11657

Application in CFD

Hu & Liu (2025), arXiv:2507.16069



Application in Weather Modelling



SAEs for fluid flows: current limitations

SAEs assume linear composition

$$\hat{x} = b + \sum_i c_i \, d_i$$

Every SAE applied to fluids so far reconstructs activations through a **linear decoder**, including SAE neural operators and KAN-SAE.

Fluid physics is quadratic

$$\dot{a}_i = \sum_j L_{ij} \, a_j + \sum_{j,k} Q_{ijk} \, a_j a_k$$

Advection $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$ gives a **pairwise interaction tensor** Q_{ijk} that no linear combination of modes can express. Energy $\frac{1}{2}|\boldsymbol{u}|^2$, enstrophy and Reynolds stresses are quadratic too.

Symptoms and gaps in current studies

Fragmented features

Non-linear features are never represented; interactions spread over many weak features.

No physical ground truth

Validated by eye or vorticity thresholds; feature interactions cannot be validated.

Conservation laws

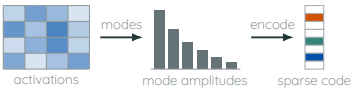
How mass and energy conservation are encoded is untested.

Weak causal tests

Only single-feature scaling; no activation patching, no shared benchmark.

Beyond the state of the art: a spectral polynomial SAE

1 Spectral encoder



Express activations in flow modes (Fourier / POD), then keep the few most active features.

2 Polynomial decoder

$$\hat{\boldsymbol{x}} = \boldsymbol{b} + \underbrace{\sum_i c_i\, \boldsymbol{d}_i}_{\text{linear (standard SAE)}} + \underbrace{\sum_{i,j} c_i c_j\, \boldsymbol{d}_{ij}}_{\text{quadratic} \leftrightarrow \text{NS nonlinearity}} + \underbrace{\sum_{i,j,k} c_i c_j c_k\, \boldsymbol{d}_{ijk}}_{\text{cubic} \leftrightarrow \text{triadic transfer}}$$

Physics as ground truth

The learned interaction tensor is compared directly with the Galerkin tensor Q_{ijk} of the same flow, e.g. the cylinder-wake shift mode and triadic couplings in turbulence.

Conservation laws by degree

Falsifiable test: mass conservation ($\nabla \cdot \boldsymbol{u} = 0$, linear) sits in the linear terms, while energy conservation (quadratic) needs the interaction terms.

Causal toolkit & open benchmark

Joint interventions on feature pairs and circuit analysis across layers; physics probes (Q-criterion, λ_2, divergence and energy residuals) vs. linear SAEs.

Applied to open-weight surrogates: Poseidon, Walrus, DPOT, PDE-Transformer, MeshGraphNets