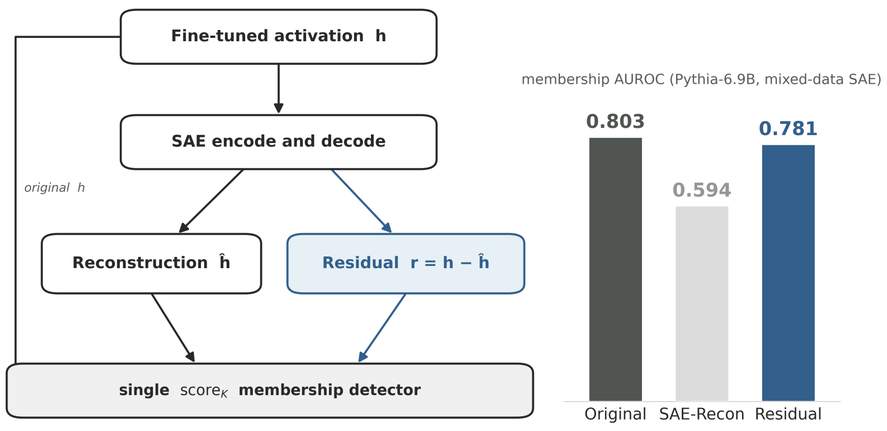
ICML 2026 Mechanistic Interpretability Workshop · SPOTLIGHT

The Dark Subspace of Fine-Tuning Memorisation

Niloy Purkait · Gonzalo Nápoles · Emmanuel Keuleers · Henry Brighton · **Tilburg University**



Editing SAE features for privacy may not erase training data evidence from LLMs.

The reconstruction residual is **the dark subspace**, the part of the activation a feature edit leaves unchanged.

Each activation decomposes into what the sparse autoencoder reconstructs and the residual it omits. A single membership detector **(box 1)** recovers from the residual most of what the reconstruction discards **(box 2)**.

The membership signal is recoverable from the residual, not the reconstruction. The same ordering holds across all seven settings that pass the gate **(box 4)**. A direct attempt to remove the evidence only relocates it **(box 3)**, and stopping a model from reproducing its training text does not remove the evidence of what it was trained on **(box 5)**.

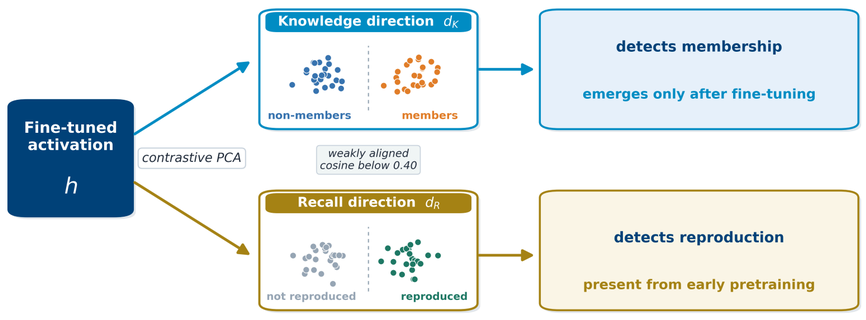
WHAT THIS MEANS

Output suppression is not privacy.

PrivacyScalpel, SSPU, and DSG use probing or SAE-based feature representations to identify and intervene on privacy relevant features. Our work shows that they mainly act on the recall direction, not the membership signal. Audit the reconstruction and the residual before a feature edit is called unlearning.

1 The method

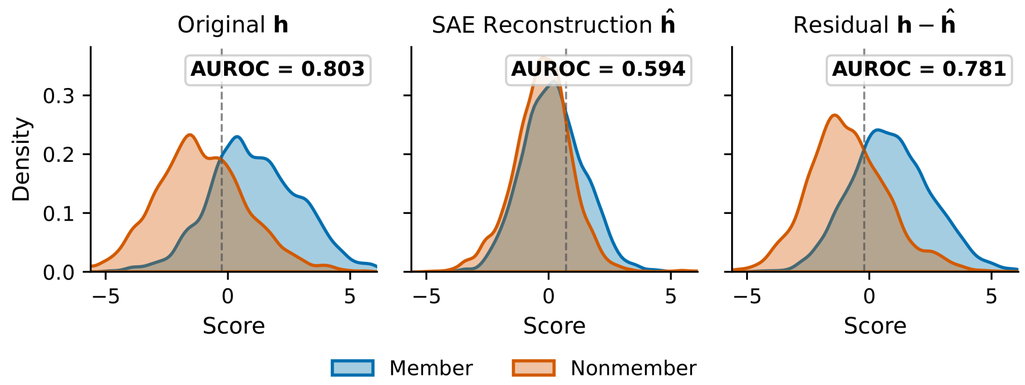
A white-box auditor at the analysis layer decomposes each mean-pooled activation h into the SAE reconstruction $\hat{h}$ and the residual $r = h - \hat{h}$.



Contrastive PCA decomposes each activation into a knowledge direction that detects membership (members vs non-members) and a recall direction that detects verbatim reproduction (reproduced vs not). Membership is measured by score$_K$, the projection onto that single knowledge direction. Observed weak alignment between these directions (cosine below 0.40) suggests that SAE-based privacy interventions can suppress reproduction while the membership signal remains recoverable.

2 Members and non-members separate

AUROC



On the controlled Pythia-6.9B model, the same score$_K$ detector (the projection of each activation onto the knowledge direction d_K) scores the original activation, the reconstruction, and the residual. Members and non-members become less distinguishable in the reconstruction (0.594), yet remain distinguishable in the residual (0.781), close to the original activation (0.803). The signal is not explained by vocabulary, norm, or length (bag-of-words AUROC 0.46).

3 Suppression relocates the evidence

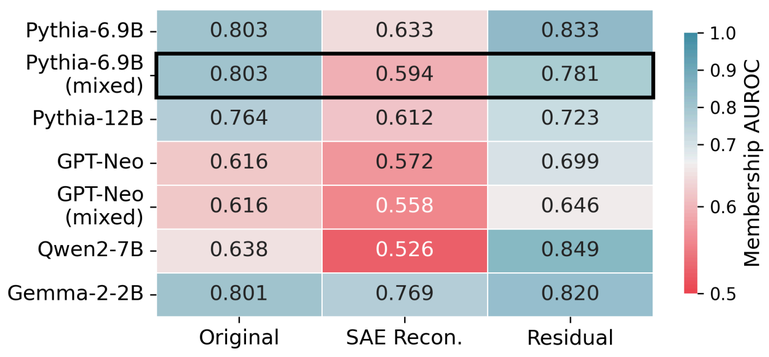
FALSIFICATION TEST

To test a best-case removal, we retrain the SAE with a penalty that removes the labelled membership direction d_K from the residual. The targeted score$_K$ decreases to chance, yet a full-vector detector on the residual still recovers membership. The signal is distributed across many directions, not localised in the one removed. A random-direction control shows no recovery.

Targeted direction only	Whole residual vector
0.52	**0.80**
at chance	recovers membership

4 Residual exceeds reconstruction

GATE-PASSING · 7/7

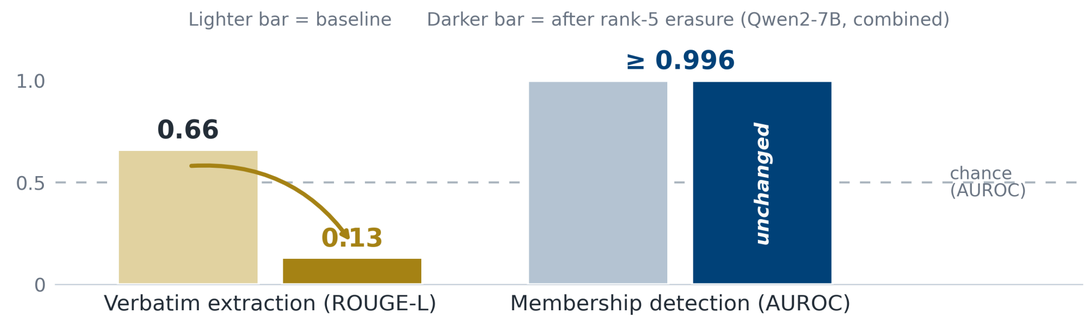


The residual exceeds the reconstruction in all seven settings that pass the reconstruction-quality gate (sign test **p ≈ 0.008**). The boxed row is the controlled Pythia-6.9B cell.

5 Extraction is not detection

INTERVENTION

Stopping a model from reproducing text does not remove the evidence of what it was trained on. Erasing the knowledge or recall directions reduces verbatim extraction while membership detection holds on most architectures (Pythia-12B sole exception).



Scope. We audit a standard L1 ReLU sparse autoencoder, not DeepMind's Gemma Scope or JumpReLU. A TopK SAE at this setting reverses the ordering. Mistral-7B and Llama-3-8B show the strongest signal but fail a predefined reconstruction-quality gate and are excluded. The controlled Pythia experiments provide the main causal evidence. Cross-architecture results are single-seed and exploratory. The residual-above-reconstruction ordering is the robust claim.

The residual audit

TAKE-HOME

1. Fit a membership detector at the analysis layer.
2. Apply the feature edit and recompute the activation.
3. Decompose it into SAE reconstruction and residual.
4. Report membership AUROC on both. A privacy claim holds only when membership is not recoverable from the residual.

Recall forms before knowledge

OBSERVATIONAL

The recall direction is detectable from early pretraining, while membership detection remains at chance there and the knowledge direction begins to emerge by the third fine-tuning epoch.

Pythia-1B developmental probe, single fine-tuning seed.

Paper **Code**