The Dark Subspace of Fine-Tuning Memorisation

Niloy Purkait [1] **Gonzalo Nápoles** [1] **Emmanuel Keuleers** [1] **Henry Brighton** [1]

Abstract

Sparse autoencoders (SAEs) learn a dictionary of sparse features over neural-network activations and are widely used to interpret and edit language models. Privacy and safety methods now ablate or steer these features to suppress unwanted behaviour, acting only on what the dictionary represents. We ask whether the dictionary preserves the evidence that a document was used to fine-tune the model, the question studied by membership inference attacks. We decompose each activation into the SAE reconstruction and the reconstruction residual, then train membership detectors on each. In controlled Pythia replications, SAE reconstruction weakens detection, yet detectors recover much of the lost signal from the residual. The same residual-above-reconstruction ordering holds across ten model-SAE settings spanning six architecture families. Surface confounds (norm, length, bag-of-words) and a label-shuffle permutation test do not account for the pattern. Privacy and safety methods that operate only on SAE features may therefore leave training-data evidence detectable in the reconstruction residual.

1. Introduction

Membership inference asks whether a trained model exposes evidence that a particular document was included in its training set (Carlini et al., 2021; 2023; Shi et al., 2023; Zhang et al., 2025; Wang et al., 2025a). Prior work shows that fine-tuning can leave readable activation traces and that membership can be inferred from hidden representations, perturbation responses, and low-rank activation structure (Ibanez-Lissen et al., 2025; Makhija et al., 2025; Minder et al., 2026; Ranjan et al., 2026; Zhang et al., 2026). A parallel line of work uses sparse autoencoders (SAEs) to interpret and edit these activations, and a growing set of methods edits, suppresses, or ablates SAE features to reduce undesirable model outputs (Wang et al., 2025b; Frikha et al., 2025; Suri et al., 2025; Muhamed et al., 2025). Such methods can directly control only the information represented in the learned dictionary and its reconstruction. Recent SAE work cautions against treating that dictionary as a complete representation of the activation, since reconstruction error can carry structured components (Engels et al., 2025) and broader evaluations report limits in reconstruction fidelity, feature completeness, and feature canonicity (Karvonen et al., 2025; Leask et al., 2025). When the SAE reconstruction preserves membership information, feature-level editing acts on a representation that contains the membership signal. When reconstruction weakens that signal but the residual preserves it, feature-level editing may still control targeted outputs without controlling membership inference risk.

In this paper, we decompose each activation in the fine-tuned model into the SAE reconstruction and its residual, then train a membership detector on each part. Within each activation we identify two directions, a knowledge direction separating members from non-members and a recall direction separating texts the model reproduces well from texts it does not. The knowledge direction depends on the member partition and exists only after fine-tuning. The recall direction is a loss-based contrast present in any model, separating high-confidence from low-confidence text in pretraining and high-recall from low-recall member text after fine-tuning. We gate cross-architecture comparison by an explicit reconstruction-quality criterion. Feature-level and direction-targeted interventions act as falsification tests for the residual signal.

We report three main findings. First, SAE reconstruction discards membership evidence that the reconstruction residual still carries, and this residual-above-reconstruction ordering replicates on every architecture where the reconstruction is reliable. Initialisation, detector fitting, training corpus, and surface artefacts do not explain the pattern (§4.6). Second, erasing the leading knowledge directions reduces verbatim extraction of fine-tuning text across architectures while membership detection is preserved on most architectures (Table 15, Table 19, Appendix B.11). Third, the recall direction is detectable from early pretraining, while the knowledge direction takes several fine-tuning epochs to emerge in

[1]Tilburg University, Tilburg, Netherlands. Correspondence to: Niloy Purkait <n.purkait@tilburguniversity.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

activations (§4.7). SAE feature-level interventions therefore require a reconstruction-residual audit before they support a privacy guarantee.[1]

2. Related Work

Membership inference and activation-level evidence. Classical membership inference attacks use shadow models (Shokri et al., 2017) or loss thresholds (Yeom et al., 2018), while LLM attacks use token probabilities, likelihood ratios, calibrated scores, and neighbourhood comparisons (Carlini et al., 2021; Shi et al., 2023; Zhang et al., 2025; Ye et al., 2022; Mattern et al., 2023; Mireshghallah et al., 2022). Attacks on fine-tuned LLMs further refine these signals with local windows and in-context probing (Chen et al., 2026; Lu et al., 2026). Prior work also shows that benchmark artefacts such as temporal shift, vocabulary imbalance, candidate-set construction, and paraphrase sensitivity can dominate reported performance (Duan et al., 2024; Meeus et al., 2025). A complementary white-box line finds membership evidence in internal states, including layer-wise hidden activations, attention and hidden-state statistics, gradient-induced activation drift, and activation differences before and after fine-tuning (Ibanez-Lissen et al., 2025; Makhija et al., 2025; Ranjan et al., 2026; Minder et al., 2026). These works establish that activations can carry membership evidence. We ask how that evidence behaves under an SAE encoder-decoder decomposition.

Sparse autoencoders and incomplete reconstruction. Sparse autoencoders learn overcomplete dictionaries that express activations through sparse feature coordinates (Cunningham et al., 2024; Bricken et al., 2023), building on the superposition hypothesis of Elhage et al. (2022). Recent work scales SAE training with TopK-style objectives and frontier-model dictionaries (Gao et al., 2024; Templeton et al., 2024). At the same time, SAE work cautions against treating the learned dictionary as a complete representation of the activation. SAE reconstruction error can contain structured components with downstream effects (Engels et al., 2025), and broader evaluations identify limits in reconstruction fidelity, feature completeness, feature canonicity, and steering-vector decomposition (Karvonen et al., 2025; Leask et al., 2025; Mayne et al., 2024; Sharkey et al., 2025; Peng et al., 2025). Recent probing studies similarly find that SAE reconstructions can underperform residual stream or tuned linear baselines (Smith et al., 2025; Kantamneni et al., 2025; Davis et al., 2026). We bring this concern to membership privacy by testing whether fine-tuning membership evidence is preserved by the SAE reconstruction or remains recoverable from the reconstruction residual.

[1]Code, data, and analysis scripts are available at https://github.com/NiloyPurkait/DarkSubspace (see Appendix D.17).

SAE-based privacy interventions and unlearning. A growing line of work uses SAE features as the substrate for targeted unlearning, guardrails, and privacy interventions. SSPU projects weights onto SAE-defined subspaces (Wang et al., 2025b). PrivacyScalpel ablates SAE features to suppress personally identifiable information leakage (Frikha et al., 2025). DSG uses dynamic SAE guardrails for inference-time intervention (Muhamed et al., 2025). Suri et al. (2025) apply SAE feature steering to memorisation mitigation in Gemma, and Farrell et al. (2024) report a related negative result on WMDP-Bio unlearning. These methods act through SAE feature coordinates, while other memorisation-mitigation approaches act on raw activations or circuit-level primitives (Zhang et al., 2026; Lasy et al., 2025). Our audit tests whether membership evidence is preserved by the representation that feature-level interventions edit. If detectors recover membership signal from the reconstruction residual, then suppressing SAE features or extractable outputs does not by itself demonstrate reduced membership inference risk.

Contrastive geometry and robustness controls. Our analysis also connects to work that uses feature patching and contrastive geometry to localise model functions. Sparse feature circuits localise functions by patching SAE features (Marks et al., 2025), while contrastive principal component analysis, representation engineering, and activation addition identify directions associated with model outputs (Abid et al., 2018; Zou et al., 2023; Rimsky et al., 2024). Some functions can be mediated by a single residual stream direction (Arditi et al., 2024), but our residual principal component intervention is more limited. It tests partial concentration of membership signal and is not treated as complete causal localisation. Following the logic of extraction-detection separation from neuropsychology (Teuber, 1955; Shallice, 1988), we treat membership detection and verbatim extraction as potentially separable outcomes, where one may be present without the other, implying that they should not be interpreted as the same privacy signal, and include perturbation and low-FPR (false positive rate) controls motivated by known fragilities in membership inference evaluation (Carlini et al., 2022; Duan et al., 2024; Meeus et al., 2025).

3. Methods

We probe mean-pooled residual stream activations at one layer of each fine-tuned model (Figure 1). Each activation is decomposed into a knowledge direction $\mathbf{d}_K$ and a recall direction $\mathbf{d}_R$, and separately into an SAE reconstruction and reconstruction residual. The experiments ask whether the SAE encoder-decoder preserves the membership-discriminating direction and whether the SAE feature dictionary accounts for any residual membership

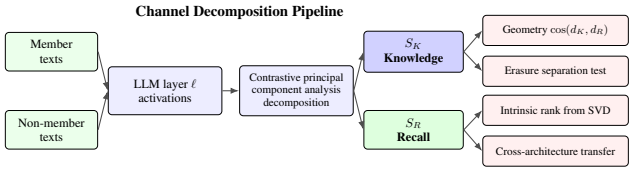

Figure 1. Channel decomposition pipeline. Given member and non-member texts, we extract layer-ℓ activations from a language model and apply contrastive principal component analysis to obtain two subspaces, a knowledge channel $\mathbf{S}_K$ and a recall channel $\mathbf{S}_R$ (each channel is a contrastively estimated low-rank activation subspace, see §3.2). We then evaluate the decomposition with four diagnostics. Geometry tests whether the knowledge and recall directions are distinct. Erasure tests whether the distinction has causal consequences. SVD estimates whether the signal is concentrated or distributed. Cross-architecture transfer tests whether the geometry is model-specific or shared.

signal. Cross-architecture experiments test for the same ordering across models that pass the validity gate of §3.7. Controlled Pythia experiments provide the multi-seed evidence on single-corpus member and non-member splits with known fine-tuning labels. We treat the residual experiments as a recoverability test (Elazar et al., 2021).

3.1. Fine-tuning setup with ground-truth membership labels

Each target model is obtained by fine-tuning a pretrained transformer on a controlled subset of OpenWebText (Gokaslan & Cohen, 2019) following the fine-tuning membership setup of Meeus et al. (2025). Membership labels are known for this fine-tuning stage. Members are documents used for fine-tuning, and non-members are same-distribution OpenWebText documents from a disjoint partition. This same-corpus construction reduces temporal-shift and vocabulary-imbalance confounds (Duan et al., 2024; Meeus et al., 2025). Each fine- tuned model uses 1,000 members and 1,000 non-members, all truncated to a common token length. We fine-tune models from 1B to 12B parameters across eight architecture families (Biderman et al., 2023; Black et al., 2021; Zhang et al., 2022; Penedo et al., 2023; Jiang et al., 2023; Grattafiori et al., 2024; Yang et al., 2024; Team et al., 2024). The residual analyses cover ten model-SAE settings across six architecture families. Seven settings from four families pass the SAE validity gate of §3.7. We additionally analyse Pythia-1B at pretraining checkpoints and at intermediate fine-tuning epochs. Other fine-tuned models are analysed after five epochs.

3.2. Separating knowledge and recall channels

For the channel decomposition, we compute mean-pooled residual stream activations at each transformer layer. We use *channel* as a shorthand for a contrastively estimated low-rank activation subspace (Abid et al., 2018). The knowledge channel contrasts member and non-member documents, while the recall channel contrasts high-recall and low-recall member documents (those the model regenerates well versus less well).

We estimate the knowledge subspace $\mathbf{S}_K$ with contrastive principal component analysis (Abid et al., 2018) on member and non-member activations. We eigendecompose the contrastive covariance $C_{\text{mem}} - \alpha\, C_{\text{non}}$ with $\alpha = 1.0$ and keep the leading $n_c = 10$ eigenvectors as the subspace basis. We estimate the recall subspace $\mathbf{S}_R$ analogously, contrasting high-recall and low-recall members. Recall labels use a loss-based proxy that splits members at the median per-token cross-entropy loss (Yeom et al., 2018; Carlini et al., 2023). We validate this proxy against ROUGE-L extraction scores (longest-common-subsequence overlap between greedy continuation and reference) for each model (Appendix A.2).

We summarise the relationship between $\mathbf{S}_K$ and $\mathbf{S}_R$ using cosine similarity between their leading directions $\mathbf{d}_K$ and $\mathbf{d}_R$, principal angles between the subspaces, and cross-validated detector area under the ROC curve (AUROC) on each projection (membership for $\mathbf{d}_K$, recall for $\mathbf{d}_R$).

3.3. Where membership signal survives SAE reconstruction

We next test whether SAE reconstruction preserves the membership signal.

Threat model and SAE decomposition. We adopt a white-box auditor threat model that extends the auditor framing of Carlini et al. (2023) to internal activations. The auditor can read layer activations and SAE reconstructions, and has labelled member and same-distribution non-member documents for direction estimation and detector calibration. For each model, the measured object is the mean-pooled residual stream activation $\mathbf{h}$ at the SAE target layer. We encode and decode $\mathbf{h}$ through the SAE to obtain the reconstruction $\hat{\mathbf{h}} = \text{dec}(\text{enc}(\mathbf{h}))$ and the reconstruction residual $\mathbf{r} = \mathbf{h} - \hat{\mathbf{h}}$. We then compute membership AUROC on each of $\mathbf{h}$, $\hat{\mathbf{h}}$, $\mathbf{r}$ using score_K, the projection onto the fitted knowledge direction $\mathbf{d}_K$ (§3.7). A full-vector logistic-regression probe is reported in Appendix B.17.

Targeting the known membership direction. We compare two SAE training configurations. The *member-only SAE* is trained only on member documents. The *mixed-data SAE* is trained on the union of member and non-member documents and is therefore label-blind by construction. We test whether the residual membership signal is exhausted by the known direction $\mathbf{d}_K$. We fine-tune a pretrained mixed-data SAE under an augmented training objective that adds a

squared penalty on the reconstruction error along $\hat{\mathbf{d}_K}$,

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_1 \|\mathbf{z}\|_1 + \lambda_{d_K} \frac{1}{N} \sum_{i=1}^{N} \big((\mathbf{h}_i - \hat{\mathbf{h}}_i) \cdot \hat{\mathbf{d}_K}\big)^2, \quad (1)$$

where the first two terms $\mathcal{L}_{\text{recon}} + \lambda_1 \|\mathbf{z}\|_1$ form the standard L1-regularised SAE objective, with $\mathcal{L}_{\text{recon}}$ measuring activation reconstruction quality and $\lambda_1 \|\mathbf{z}\|_1$ penalising dense codes $\mathbf{z}$ on a batch of N documents with mean-pooled residual stream activation $\mathbf{h}_i$ and its SAE reconstruction $\hat{\mathbf{h}}_i$ (§3.3). The added third term measures how much of the reconstruction error remains along the labelled membership direction $\hat{\mathbf{d}_K}$, the unit-normalised leading knowledge direction estimated by contrastive principal component analysis on the held-out audit partition (§3.2), with the nonnegative scalar λ_{d_K} weighting this projection penalty. Minimising it pushes the SAE to allocate dictionary capacity so that reconstruction is faithful in the $\hat{\mathbf{d}_K}$ direction. Setting $\lambda_{d_K} = 0$ recovers the baseline objective, while larger values strengthen reconstruction along $\hat{\mathbf{d}_K}$, allowing us to test whether the residual signal disappears once the labelled direction is explicitly preserved by the SAE. Fine-tuning hyperparameters and the held-out audit and disjoint-evaluation partition split are in Appendix B.17.

3.4. Whether SAE features carry membership and recall

We test whether membership and recall directions are encoded by identifiable SAE features. This matters for feature-level interventions. If the relevant channel is spanned by a small set of SAE decoder directions, then ablating those features should affect the corresponding membership or recall outcome.

For each model we choose the analysis layer with highest cross-validated membership-detection AUROC under a fixed five-fold split on raw mean-pooled residual stream activations (§4.7, §3.6, Appendix A.5). SAE-based analyses (§4.2, §4.3, §4.4) use the depth-matched layer instead, at relative depth 0.5 across the controlled Pythia family (layer 16 of 32 for Pythia-6.9B, layer 18 of 36 for Pythia-12B). For Pythia-12B this differs from the per-layer AUROC peak (Appendix B.2). Per-model layers and the scaling-curve consequence are in Appendix Table 2 and Appendix A.5.

At the selected layer, we call an SAE feature a classifier feature if its activation differs between members and nonmembers at false discovery rate of 5%. We then ablate the top-k classifier features in the SAE code, decode the modified code back into activation space, and measure both membership detection and verbatim extractability. This tests whether the features most associated with membership labels are also sufficient to control the downstream membership and recall outcomes.

To test whether the selected SAE features span the channel they are meant to explain, let $\mathcal{F}$ be a selected set of SAE features. Each feature has a decoder vector, and those decoder vectors span a subspace of activation space. Let $\mathbf{P}_{\mathcal{F}}$ denote the orthogonal projection onto that feature subspace. For a target channel subspace $\mathcal{S}$, we define the feature sufficiency criterion as

$$\text{FSC}(\mathcal{S}, \mathcal{F}) = \frac{\|\mathbf{P}_{\mathcal{F}}(\mathcal{S})\|_F^2}{\|\mathcal{S}\|_F^2}, \quad (2)$$

where the numerator is the squared Frobenius norm of the target channel after projection onto the selected feature subspace and the denominator is the channel's squared Frobenius norm. FSC is bounded in $[0, 1]$ and measures how much of the channel is captured by the selected SAE features (1 means full coverage, 0 means none). The full per-axis interpretation is in Appendix B.9.

3.5. Whether interventions separate extraction from detection

We use subspace erasure to test whether membership detection and verbatim extraction depend on separable activation directions. The intervention removes one channel from the activation and measures what changes downstream. If erasing a channel suppresses verbatim extraction without suppressing membership detection, then extraction and detection depend on different activation directions.

For a channel with orthonormal basis $\mathbf{V}$, erasure removes the component of the activation lying in that channel. Given a mean-pooled residual stream activation $\mathbf{h}$, we define the channel-erased activation as

$$\mathbf{h}_{\setminus \mathbf{V}} = \mathbf{h} - \mathbf{V}\mathbf{V}^\top \mathbf{h}, \quad (3)$$

where $\mathbf{V}\mathbf{V}^\top \mathbf{h}$ is the projection of $\mathbf{h}$ onto the channel subspace, and $\mathbf{h}_{\setminus \mathbf{V}}$ is the remaining activation after that projection is removed.

We apply this operation separately to the knowledge basis $\mathbf{V}_K$ and the recall basis $\mathbf{V}_R$. After each intervention, we measure membership detection AUROC and verbatim extractability with ROUGE-L. The full per-condition statistical procedure is in Appendix A.3, and per-model metrics are in Appendix B.7.

3.6. Whether the residual ordering replicates across architectures

We test whether the same channel geometry appears across model architectures. We evaluate three geometric questions. First, we measure shared channel geometry by principal angles between the knowledge subspaces of model pairs at matching hidden dimension. Second, we measure how spread out member activations are with the participation

ratio. Third, we measure whether the shift between members and non-members lies inside the dominant variance directions with the cross-architecture concentration C_k of the membership direction in the top-k principal component basis.

Each measure has a direct reading. Small principal angles indicate that two models place their knowledge subspaces in the same region. A high participation ratio indicates spread-out member variance, while a low ratio indicates compression into a few directions. C_k near 1 means the membership shift lives along the dominant variance directions, while C_k near 0 means it lies on lower-variance or off-subspace directions. The singular value decomposition (SVD) identity is in Appendix B.4, the participation-ratio and concentration formulae and the architecture-family grouping rationale are in Appendix B.15, a linearity check that compares nonlinear and linear membership probes is in Appendix B.3, and detailed measure readings are in Appendix B.5.

3.7. Validity gate for quantitative claims

We interpret cross-architecture comparisons only for SAE runs satisfying predefined validity criteria.

Standard L1-regularised ReLU SAE training is per-initialisation unstable (Bricken et al., 2023; Templeton et al., 2024). Unstable runs fail the validity gate by construction (Appendix D.7).

The validity gate decides which model-SAE settings we interpret quantitatively. The strict criterion requires reconstruction cosine at least 0.90 and mean L0 (active features per token) below 60% of dictionary size. A permissive criterion at reconstruction cosine at least 0.85 admits additional settings to the qualitative cross-architecture ordering. Settings below 0.85 are reported as diagnostic context only. Pythia-6.9B and Pythia-12B carry the strict-gate replication evidence. Pythia-1B serves as a permissive-gate diagnostic and supplies the developmental analysis (§4.7), not the controlled-replication claim. Per-model reconstruction outcomes against the gate thresholds are in Table 1. Multi-seed variance is estimated on the controlled Pythia settings, and cross-architecture settings are single-seed exploratory replications. The cross-architecture directional sign test uses the gate-passing setting as its unit. Exclusions and the dictionary-expansion plus L1 settings used to clear the gate are in Appendix B.18 and Appendix D.6.

The dictionary-corpus choice matters because it controls whether the decomposition can see the evaluation labels. The mixed-data SAE is trained on the union of the member and non-member pools, so the dictionary is label-blind by construction. We also retrain the Pythia-6.9B mixed-data SAE on an OpenWebText partition disjoint from the evaluation pool, testing whether overlap between SAE training data and evaluation documents explains the residual signal (§4.6, Appendix D.16). Member and non-member partitions are independent and identically distributed (IID) by construction (same OpenWebText corpus, different documents). Recent work shows that pretraining membership inference attacks (MIAs) without IID splits can be confounded by distribution shift (Maini et al., 2024; Duan et al., 2024).

The detector scope matters because the main effect size depends on it. The membership score is score_K, the projection of an activation onto the fitted knowledge direction $\mathbf{d}_K$. AUROC is the main separability metric in the body, with low-FPR operating points in Appendix D.17. We also refit $\mathbf{d}_K$ on a held-out partition to test partition sensitivity (Appendix D.10). Sparsity-range scope of the geometric and feature-alignment analyses is in Appendix D.12.

4. Results

4.1. Knowledge and recall directions are weakly aligned at the SAE layer

Across the nine fine-tuned transformers reported in Table 17 (Gemma-2-2B is reported separately in Appendix D.12), the leading knowledge direction $\mathbf{d}_K$ and the leading recall direction $\mathbf{d}_R$ are weakly aligned at the sparse autoencoder layer, with cosine similarity below 0.40 across models (Table 17, Appendix B.10). Per-layer principal angles between the corresponding $\mathbf{S}_K$ and $\mathbf{S}_R$ subspaces are reported in Appendix B.10. Membership detection and extractable recall therefore do not reduce to the same leading activation direction.

Layer-by-layer detector curves support the same separation dynamically. Membership and recall signals develop on different schedules, with mutual alignment rising in late layers only for the two models under capacity pressure (Figure 4, Appendix B.10). This motivates the SAE reconstruction test. If the learned SAE representation preserves the membership-relevant part of the activation, reconstruction should preserve membership detection performance. If reconstruction weakens detection while the residual preserves it, then feature-level interventions may miss membership evidence left outside the reconstruction.

4.2. SAE reconstruction weakens membership detection while residual detectors recover signal

We next test whether membership signal is preserved in the SAE reconstruction or remains recoverable from the reconstruction residual. Table 1 compares membership detection on the original hidden state $\mathbf{h}$, the SAE reconstruction $\hat{\mathbf{h}}$, and the reconstruction residual $\mathbf{h} - \hat{\mathbf{h}}$. Figure 2 illustrates the audit on the Pythia-6.9B mixed-data setting, where detection weakens under the SAE reconstruction and recovers from the residual (the dark subspace).

Table 1. Membership signal under SAE decomposition. AUROC is score$_K$, the projection onto the fitted knowledge direction $\mathbf{d}_K$ (§3.7). Full-vector logistic-regression probe AUROCs are in Appendix B.17. Δ_{recon} is the change from Original to SAE-Recon. Cosine reconstruction quality is reported as $\cos(x, \hat{x})$. N is the number of SAE training seeds. Pythia at $N = 5$ carries the multi-seed evidence. Cross-architecture settings at $N = 1$ are exploratory. Markers † (feature collapse) and ‡ (reconstruction-cosine fail) flag exclusions. Unmarked rows form the gate-passing set scored by the reported tests. Bootstrap confidence intervals (Efron, 1987) and additional reconstruction-quality diagnostics are reported in Appendix D.2.

Model / SAE setting	N	Original	SAE-Recon	Residual	Δ_{recon}	$\cos(x, \hat{x})$
Controlled Pythia experiments						
Pythia-6.9B	5	0.803	0.633	0.833	−17 pp	0.940
Pythia-6.9B (mixed)	5	0.803	0.594	0.781	−21 pp	0.976
Pythia-12B	5	0.764	0.612	0.723	−15 pp	0.991
Cross-architecture experiments (gate-passing)						
GPT-Neo-2.7B	1	0.616	0.572	0.699	−4 pp	0.998
GPT-Neo-2.7B (mixed)	1	0.616	0.558	0.646	−6 pp	0.997
Qwen2-7B	1	0.638	0.526	0.849*	−11 pp	0.850
Gemma-2-2B	1	0.801	0.769	0.820	−3 pp	0.911
Diagnostic context (excluded from claim)						
Pythia-1B‡	5	0.663	0.593	0.670	−7 pp	0.869
OPT-6.7B‡	1	0.774	0.557	0.866*	−22 pp	0.533
Falcon-7B†‡	1	0.690	0.694	0.827*	+0.5 pp	0.767

*Residual norm AUROC exceeds 0.65. †Falcon-7B activates only 9 SAE features. ‡Reconstruction cosine falls below the 0.90 strict threshold (Pythia-1B at 0.869 sits between the strict and permissive thresholds, while OPT-6.7B and Falcon-7B fall below the 0.85 permissive threshold). The setting is reported as diagnostic context only. Mistral-7B and Llama-3-8B are excluded by the validity gate (failed reconstruction, Appendix D.6). These models carry the strongest membership signal at the SAE layer in the cross-architecture set (AUROC 0.967, 0.957, Table 17), so the claim is conditional on architectures that pass the validity gate. Held-out $\mathbf{d}_K$ estimation preserves the ordering at smaller magnitude (§4.6, Appendix D.10). Token-level Bag-of-Words controls are reported in Appendix D.3. Every gate-passing setting has residual above SAE-Recon. The binomial sign test gives $p \approx 0.0625$ on the non-Pythia subset ($n = 4$) and $p \approx 0.008$ on the full gate-passing set ($n = 7$). Per-setting bootstrap CIs in Appendix D.2.

The controlled Pythia experiments provide the strongest decomposition evidence. On Pythia-6.9B, both the member-only SAE and the mixed-data SAE weaken membership detection after reconstruction (17 and 21 percentage points respectively, each pooled over five SAE seeds), while residual detectors recover most of the lost signal. The mixed-data SAE trains on the union of member and non-member activations, so the matching reconstruction-induced reduction rules out member-only dictionary training as the sole explanation in this controlled setting. The disjoint-corpus SAE retrain, in which the SAE training texts share no documents with the membership evaluation set, reproduces the same ordering with a single-seed reconstruction reduction of 0.278, on the favourable side of the 0.209 mean across the five in-corpus mixed-data seeds, and supports the finding independently (Appendix D.16). Pythia-12B shows the same ordering at smaller magnitude. The Pythia results therefore suggest that SAE reconstruction discards membership evidence that remains available in the reconstruction residual.

Figure 3 shows why the AUROC changes. In the Pythia-6.9B mixed-data setting, original activations separate members from non- members. SAE reconstruction collapses the separation, and the reconstruction residual restores it. The effect is therefore visible not only as a scalar AUROC reduction, but as a movement of the member and non-member score distributions.

The residual detector is more stable across SAE seeds than the dictionary projection. The variance across five Pythia-12B SAE initialisations is consistent with the residual-above-reconstruction interpretation. As the SAE seed varies, the residual detector recovers membership signal at nearly constant strength, whereas the dictionary-projection detector is more variable. This asymmetry is evidence against a chance-miss explanation. Per-seed standard deviations and the algebraic identity relating reduction variance to reconstruction variance are reported in Appendix D.11. The magnitude of the reconstruction-induced reduction is partition-sensitive under held-out estimation of $\mathbf{d}_K$ (§4.6), whereas the inversion reported here is not.

Targeting the known membership direction does not eliminate the residual signal. We next test whether the residual signal disappears when the SAE is explicitly trained to reconstruct $\mathbf{d}_K$. Adding a reconstruction penalty along $\mathbf{d}_K$ restores the $\mathbf{d}_K$-based score in the SAE reconstruction, but a general-purpose detector still recovers membership signal from the residual at a level comparable to the original activation (Appendix B.17). A random-direction control does not produce the same recovery, consistent with the targeted training acting on $\mathbf{d}_K$.

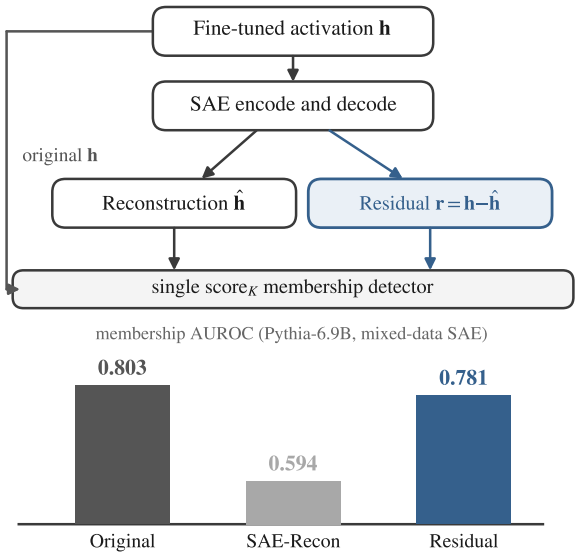

Figure 2. The residual audit on Pythia-6.9B (mixed-data SAE). Each activation $\mathbf{h}$ is split into the SAE reconstruction $\hat{\mathbf{h}}$ and the reconstruction residual $\mathbf{r} = \mathbf{h} - \hat{\mathbf{h}}$, and the membership detector score_K is applied to the original activation, the reconstruction, and the residual. Detection reaches AUROC 0.803 on the original activation, weakens to 0.594 on the reconstruction, and recovers to 0.781 on the residual. The membership evidence the dictionary discards remains recoverable from the residual.

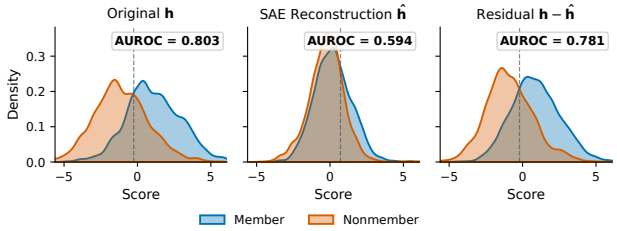

Figure 3. Per-text membership score distributions before and after SAE reconstruction for Pythia-6.9B, using the mixed-data SAE pooled across five seeds. Original activations separate members from non-members (AUROC $= 0.803$), SAE reconstruction collapses the separation (AUROC $= 0.594$), and the residual restores most of it (AUROC $= 0.781$). Magnitudes are partition-conditional. Held-out estimation of $\mathbf{d}_K$ preserves the qualitative ordering at smaller magnitude (§4.6).

The named score is suppressed while the activation-level signal persists. Within the direction-penalised conditions on Pythia-6.9B, the residual score_K falls toward chance under both penalty strengths while a detector trained on the full residual vector holds at the same near-original level (Table 21). The targeted direction is suppressed in the residual as designed, but the membership evidence used by a general-purpose detector lies outside the targeted direction. The experiment uses ground-truth $\mathbf{d}_K$ and a single seed per condition on two models, so practical direction-penalised training would require label-free proxy directions and stronger replication. Figure 7 in Appendix B.17 shows the contrast.

These tests converge on the same picture. Membership evidence in the fine-tuned activation lives in two components, one weakened under SAE reconstruction and the other carried by the reconstruction residual.

4.3. SAE features do not carry the membership signal

The full SAE decoder span contains both channel subspaces (FSC $= 1.000$, Table 16, Appendix B.9). The loss of membership evidence cannot be a property of the linear span alone, so it must arise in the encoder-to-code path.

A fold-matched logistic detector trained on SAE codes $\mathbf{z}$ does not recover the membership signal that the same detector extracts from the raw hidden state (Appendix B.16). The decoder span can represent the membership direction in principle, but the learned sparse code and reconstruction do not preserve it in usable form.

Residual norm and member-correlated features do not carry the residual signal. Residual magnitude explains part of the residual signal in some settings, but not all of it. L2-normalising residual vectors preserves directional membership signal (Table 23, Appendix B.13). Ablating top-k SAE features ranked by point-biserial correlation further weakens reconstructed membership signal rather than recovering it. The feature sufficiency criterion gives the geometric version of the same result. Across the models with reliable SAE reconstructions, classifier features span only a small fraction of the knowledge subspace $\mathbf{S}_K$ (Table 16, Appendix B.9). GPT-Neo-2.7B is the exception. On Pythia-6.9B, ablating all classifier features leaves membership detection at baseline while collapsing verbatim extraction. Methods that select SAE features by training-set correlation (Wang et al., 2025b; Frikha et al., 2025; Muhamed et al., 2025) therefore target the recall channel rather than the membership signal. Per-explanation verification protocols and detailed numerical evidence are in Appendix B.13, Appendix B.9, and Appendix D.9.

4.4. Subspace erasure separates extraction from detection

We test whether erasing the knowledge subspace separates verbatim extraction from membership detection. At rank-5 erasure, removing leading knowledge directions reduces verbatim extraction on the architectures tested (Table 19) while membership detection is preserved on most architectures with above-chance baseline detection (Appendix B.11). Pythia-12B is the exception, where erasure also reduces detection AUROC. An intervention that makes the model stop reproducing text need not remove the evidence used by a

membership detector.

4.5. Residual ordering replicates across architectures

We test whether the residual-above-reconstruction ordering is specific to Pythia. Among model-SAE settings where reconstruction is reliable enough to interpret, the same qualitative pattern appears in the non-Pythia architectures, although the magnitude varies (Table 1). The cross-architecture experiments test the directional hypothesis that residual AUROC exceeds reconstruction AUROC. The architecturally independent comparison is the non-Pythia subset of four gate-passing settings across three non-Pythia families (GPT-Neo member-only, GPT-Neo mixed, Qwen2-7B, Gemma-2-2B). Per-setting confidence intervals, tests of effect direction, and attempted replications are in Table 25 (Appendix D.2) and Appendix B.18.

GPT-family models rely on direction, whereas Llama-style models also rely on magnitude. The norm-direction decomposition reveals a descriptive distinction between architecture families (Table 18, Appendix B.10). The GPT-family and Llama-style groups differ on whether membership separation comes from activation magnitude or from directional structure, with implications for what SAE reconstruction preserves. The Llama-style group contains only two models (Mistral-7B and Llama-3-8B), so the distinction is descriptive. Full architecture scatter plots are in Figure 8, and the per-layer concentration analysis is in Appendix B.15.

Residual-direction ablation is exploratory. Within the controlled Pythia-12B setting, we ask whether the residual membership signal is concentrated in a small number of directions. Ablating the top 10 residual principal components produces measurable reductions in membership detection AUROC, although the intervention does not pass random-rotation, matched-noise, random column mask, or same-SAE consistency control gates (Table 27, Appendix D.15). We therefore treat it as descriptive evidence of residual concentration, not as causal localisation of the membership signal.

4.6. Confound, partition, and operating-point controls

We next test whether the reconstruction-residual finding survives partition, paraphrase, surface-form, and low-FPR operating-point controls.

Held-out $\mathbf{d}_K$ estimation preserves the qualitative ordering but reduces the magnitude. Held-out estimation of $\mathbf{d}_K$ does not support the original magnitude claim, but the qualitative residual $\geq$ original $>$ reconstruction ordering survives on both Pythia models (Appendix D.10). The ordering is therefore not an artefact of fitting and evaluating $\mathbf{d}_K$ on overlapping partitions, although the reduction magnitude is partition-sensitive.

Word-order shuffles reverse the detector orientation under paraphrase. Under paraphrase with shuffled word order, the residual detector orientation reverses sign on Pythia-6.9B, Pythia-12B, and Qwen2-7B. A valid paraphrase audit therefore requires fixed-orientation AUROC. Per-model replication notes and the bidirectional-audit methodology are in Appendix D.17.

The finding survives surface, access-setting, and low-FPR controls. Surface confounds (norm, length, bag-of-words) leave the residual signal intact (Table 18, Appendices D.4, D.3). A label-shuffle permutation test confirms that the cosine alignment between $\mathbf{d}_K$ and $\mathbf{d}_R$ is consistent with random-label baselines on eight of nine models (Appendix D.5). A corpus-disjoint SAE on Pythia-6.9B reproduces the reduction (Appendix D.16). Pythia-12B mixed-data replicates the residual ordering within the Pythia family (Appendix D.11). On Pythia-6.9B, three standard output-level detectors are unaffected by SAE reconstruction, so the layer-local effect is confined to the SAE layer (Appendix D.13). At the 0.1% FPR operating point, the residual TPR is statistically distinguishable from chance on three of four tested models, with Qwen2-7B not distinguishable from chance at the available sample size (Appendix D.17).

The pattern is asymmetric. The qualitative residual-above-reconstruction ordering survives the surface and partition controls. The quantitative magnitude depends on how $\mathbf{d}_K$ is fit, and the residual detector orientation reverses under shuffled-word paraphrase. The decomposition is best read as a directional fact whose effect size and orientation are conditional on the audit setup.

4.7. The recall channel forms before the knowledge channel

Pythia-1B (Biderman et al., 2023) carries a developmental view of the two channels, building on prior developmental analyses of training dynamics in transformers (Olsson et al., 2022; Tirumala et al., 2022; Nanda et al., 2023). We track both across pretraining checkpoints (no fine-tuning) and across fine-tuning epochs on the controlled OpenWebText partition at a single fine-tuning seed (Appendix B.14, Table 20).

Recall is a property of pretraining text-confidence encoding. The recall direction already separates high-loss from low-loss texts well above chance at the earliest pretraining checkpoint we test, and the same separation holds at later checkpoints, while membership detection sits at chance across all pretraining checkpoints (Table 20). The recall proxy is loss-based and so measures text-confidence encoding, a property of pretraining that memorisation-specific recall builds on later.

The membership channel emerges with fine-tuning. At the first fine-tuning epoch the recall channel is already established at the SAE layer while membership detection is at chance there. The SAE-layer membership signal begins to emerge by the third epoch (Table 20), and erasure of the knowledge subspace then produces a measurable reduction in detection (Appendix B.6). Cosine alignment between $\mathbf{d}_K$ and $\mathbf{d}_R$ also grows in late layers as fine-tuning proceeds (Figure 6), suggesting that limited representational capacity forces the two signals to share dimensions once both have formed.

The two channels originate in different regimes. The recall channel reflects pretraining-era encoding of text confidence, and the membership channel only develops once the model is exposed to the fine-tuning set. SAE feature edits that suppress recall therefore act on a channel that was already in place before fine-tuning, while the membership signal may be mediated by structures that formed later.

5. Concluding Remarks

SAE feature usefulness and membership privacy are different properties. SAE reconstruction can weaken membership detection while the underlying evidence remains recoverable from the residual. Features remain useful sparse coordinates for interpreting or editing model outputs, but a reduction in measured detection does not mean the evidence has been removed. Any SAE feature-level intervention should therefore be paired with a residual audit before it supports a privacy claim. Such an audit trains a residual membership detector at the analysis layer and verifies signal removal in both the residual and reconstruction.

Scope and limitations. Controlled Pythia experiments carry the main causal evidence. Our analysis is white-box on controlled OpenWebText fine-tuning with L1-regularised ReLU SAEs. The residual-above-reconstruction ordering is specific to this dictionary mechanism, and a TopK SAE at the headline setting keeps the membership signal in the reconstruction and reverses the ordering (Appendix D.8). Variance is reported across SAE seeds at one fine-tuning seed per model. The Pythia-1B developmental analysis uses a loss-based recall proxy (§4.7). The reconstruction-reduction magnitude is sensitive to the partition used to fit $\mathbf{d}_K$, and the residual detector orientation reverses under shuffled-word paraphrase, although the qualitative ordering survives. Cross-architecture experiments support qualitative generality at the validity gate (§3.7), which excludes Mistral-7B and Llama-3-8B (strongest signals in the cross-architecture set, Table 17). The direction-penalty contrast (§4.4) is uninformative on GPT-Neo-2.7B, where the full-vector probe sits at chance (Appendix B.17). Pretraining MIA is out of scope (Duan et al., 2024). Further limitations in Appendix D.14.

References

Abid, A., Zhang, M. J., Bagaria, V. K., and Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):2134, 2018.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv preprint arXiv:2406.11717.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. *ICML*, 2023.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, 2021.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bussmann, B., Leask, P., and Nanda, N. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Bussmann, B., Nabeshima, N., Karvonen, A., and Nanda, N. Learning multi-level features with matryoshka sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025. arXiv preprint arXiv:2503.17547.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, 2022.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Chen, Y., Du, Y., Zhang, K., Kundu, A., Fleming, C., Ribeiro, B., and Li, N. Window-based membership inference attacks against fine-tuned large language models. *arXiv preprint arXiv:2601.02751*, 2026. Accepted to USENIX Security 2026.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2024.

Davis, A. L., Vasquez Ferrer, R., and Sukthankar, G. Exploring the limits of probes for latent representation edits in GPT models. *AI*, 7(3):92, 2026. doi: 10.3390/ai7030092.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

Efron, B. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987. doi: 10.1080/01621459.1987.10478410.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Engels, J., Riggs, L., and Tegmark, M. Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research*, 2025. arXiv preprint arXiv:2410.14670.

Farrell, E., Lau, Y.-T., and Conmy, A. Applying sparse autoencoders to unlearn knowledge in language models. *arXiv preprint arXiv:2410.19278*, 2024. NeurIPS 2024 SAE+Unlearning Workshop.

Frikha, A., Ar Razi, M. R., Nakka, K. K., Mendes, R., Jiang, X., and Zhou, X. PrivacyScalpel: Enhancing LLM privacy via interpretable feature intervention with sparse autoencoders. *arXiv preprint arXiv:2503.11232*, 2025.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Gokaslan, A. and Cohen, V. Openwebtext corpus. 2019. http://Skylion007.github.io/OpenWebTextCorpus.

Grattafiori, A. et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Ibanez-Lissen, L., Gonzalez-Manzano, L., de Fuentes, J. M., Anciaux, N., and Garcia-Alfaro, J. LUMIA: Linear probing for unimodal and MultiModal membership inference attacks leveraging internal LLM states. In *European Symposium on Research in Computer Security (ESORICS)*, 2025. arXiv preprint arXiv:2411.19876.

Ilić, D., Stanojević, D., and Cvejoski, K. Powerful training-free membership inference against autoregressive language models. *arXiv preprint arXiv:2601.12104*, 2026.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Kantamneni, S., Engels, J., Rajamanoharan, S., Tegmark, M., and Nanda, N. Are sparse autoencoders useful? a case study in sparse probing. In *International Conference on Machine Learning (ICML)*, 2025. arXiv preprint arXiv:2502.16681.

Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J., Chanin, D., Lau, Y.-T., Farrell, E., McDougall, C., Ayonrinde, K., Till, D., Wearden, M., Conmy, A., Marks, S., and Nanda, N. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *International Conference on Machine Learning (ICML)*, 2025. arXiv preprint arXiv:2503.09532.

Lasy, I., Knees, P., and Woltran, S. Understanding verbatim memorization in LLMs through circuit discovery. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, ACL 2025, Vienna, 2025. arXiv:2506.21588.

Leask, P., Bussmann, B., Pearce, M., Bloom, J., Tigges, C., Al Moubayed, N., Sharkey, L., and Nanda, N. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv preprint arXiv:2502.04878.

Lu, Z., Chi, H., Baracaldo, N., Kadhe, S. R., Jeon, Y., and Yu, L. In-context probing for membership inference in fine-tuned language models. In *Network and Distributed System Security Symposium (NDSS)*, 2026. arXiv preprint arXiv:2512.16292.

Maini, P., Jia, H., Papernot, N., and Dziedzic, A. LLM dataset inference: Did you train on my dataset? In *Advances in Neural Information Processing Systems*, 2024. arXiv preprint arXiv:2406.06443.

Makhija, D., Arivazhagan, M. G., Kumar, V. B., and Gangadharaiah, R. Neural breadcrumbs: Membership inference attacks on LLMs through hidden state and attention pattern analysis. *arXiv preprint arXiv:2509.05449*, 2025. Preprint. Probes transformer hidden states + attention patterns (not SAE features).

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv preprint arXiv:2403.19647.

Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., and Berg-Kirkpatrick, T. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.719.

Mayne, H., Yang, Y., and Mahdi, A. Can sparse autoencoders be used to decompose and interpret steering vectors? In *NeurIPS 2024 Workshop on Foundation Model Interventions (MINT)*, 2024. arXiv preprint arXiv:2411.08790.

Meeus, M., Shilov, I., Jain, S., Faysse, M., Rei, M., and de Montjoye, Y.-A. SoK: Membership inference attacks on LLMs are rushing nowhere (and how to fix it). In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025. arXiv preprint arXiv:2406.17975.

Minder, J., Dumas, C., Slocum, S., Casademunt, H., Holmes, C., West, R., and Nanda, N. Narrow finetuning leaves clearly readable traces in the activation differences. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv preprint arXiv:2510.13900.

Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., and Shokri, R. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.570.

Muhamed, A., Bonato, J., Diab, M., and Smith, V. SAEs can improve unlearning: Dynamic SAE guardrails for precision unlearning in LLMs. *arXiv preprint arXiv:2504.08192*, 2025. COLM 2025.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. *Advances in Neural Information Processing Systems*, 36, 2023.

Peng, K., Movva, R., Kleinberg, J., Pierson, E., and Garg, N. Use SAEs to discover unknown concepts, not to act on known concepts. *arXiv preprint arXiv:2506.23845*, 2025.

Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., and Nanda, N. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Ranjan, R., Grover, U., Lin, X., and Polyzou, A. G-Drift MIA: Membership inference via gradient-induced feature drift in LLMs. *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2026. arXiv preprint arXiv:2604.00419.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Shallice, T. *From Neuropsychology to Mental Structure*. Cambridge University Press, 1988.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Heimersheim, S., Bloom, J., Biderman, S., Conmy, A., Nanda, N., Geiger, A., Geva, M., Tegmark, M., Bau, D., McGrath, T., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

Smith, L., Rajamanoharan, S., Conmy, A., McDougall, C., Kramár, J., Lieberum, T., Shah, R., and Nanda, N. Negative results for sparse autoencoders on downstream tasks and deprioritising SAE research (mechanistic interpretability team progress update). Google DeepMind Safety Research, Medium, March 2025.

Suri, M., Anand, N., and Bhaskar, A. Mitigating memorization in LLMs using activation steering. *arXiv preprint arXiv:2503.06040*, 2025.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Teuber, H.-L. Physiological psychology. *Annual Review of Psychology*, 6(1):267–296, 1955.

Tirumala, K., Markosyan, A., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35, 2022.

Wang, C., Wang, Y., Hooi, B., Cai, Y., Peng, N., and Chang, K.-W. Con-ReCall: Detecting pre-training data in LLMs via contrastive decoding. *arXiv preprint arXiv:2409.03363*, 2025a.

Wang, X., Li, Z., Wang, B., Hu, Y., and Zou, D. Model unlearning via sparse autoencoder subspace guided projections. *arXiv preprint arXiv:2505.24428*, 2025b. EMNLP 2025.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2022. doi: 10.1145/3548606.3560675.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, 2018.

Zhang, J., Sun, J., Yeats, E., Ouyang, Y., Kuo, M., Zhang, J., Yang, H., and Li, H. Min-K%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2025.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhang, X., Shang, H., and Li, X. GSS: Gated subspace steering for selective memorization mitigation in LLMs. *arXiv preprint arXiv:2602.08901*, 2026.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Experimental Details

The appendix has four parts. Appendix A gives the model, fine-tuning, and protocol details. Appendix B reports additional results, including the per-model decompositions and the worked end-to-end example (Appendix B.1). Appendix C collects additional figures. Appendix D reports robustness and boundary conditions, including held-out direction estimation, the TopK SAE scope test (Appendix D.8), and the privacy-aware feature audit (Appendix D.9).

A.1. Model and Fine-Tuning Configuration

Table 2. Models, fine-tuning configurations, and sparse autoencoder specifications. All fine-tuning uses the AdamW optimiser with a cosine learning rate schedule and a maximum learning rate of 2×10^{-5}. Dict. size shown is the primary SAE configuration used for headline residual-above-reconstruction results. Per-analysis configurations are documented in Appendix D.6. CF count is the number of classifier features identified at a Benjamini–Hochberg corrected false discovery rate of 5%.

Model	Architecture	Params	SAE layer	Dict. size	CF count
Pythia-1B	GPT-NeoX	1.0B	8	16,384	4
Pythia-6.9B	GPT-NeoX	6.9B	16	32,768	131
Pythia-12B	GPT-NeoX	12B	18	20,480	9
GPT-Neo-2.7B	GPT-Neo	2.7B	16	10,240	954
OPT-6.7B	OPT	6.7B	24	32,768	55
Falcon-7B	Falcon	7.0B	16	18,176	7
Mistral-7B	Mistral	7.2B	16	16,384	10
Llama-3-8B	LLaMA	8.0B	16	16,384	—
Qwen2-7B	Qwen2	7.6B	16	14,336	7

For Falcon-7B, 7 classifier features are identified pre-FDR and 0 survive Benjamini–Hochberg correction. FSC values are computed using the pre-FDR features. Llama-3-8B is excluded from classifier-feature analyses because the available SAE checkpoint fails the reconstruction-quality criterion. For Mistral-7B, 73 classifier features are identified pre-FDR and 10 survive Benjamini–Hochberg correction. FSC values use the pre-FDR features. For Qwen2-7B, 15 classifier features are identified pre-FDR and 7 survive Benjamini–Hochberg correction.

Fine-tuned model weights are stored locally and loaded via the HuggingFace model loading interface. All inference is conducted in half precision (float16) on NVIDIA A40, L40S, or H100 GPUs. Sparse autoencoders are trained following the protocol of Templeton et al. (2024), using an L1 penalty to encourage sparsity with 4× dictionary width. Classifier features are identified by a two-sample test on activation magnitudes between members and non-members at a Benjamini–Hochberg corrected false discovery rate of 5%.

A.2. Channel Decomposition Details

Contrastive principal component analysis. For each layer, we compute group covariance matrices on the mean-pooled hidden states. The knowledge directions are the top eigenvectors of $\mathbf{C}_{\text{mem}} - \mathbf{C}_{\text{non}}$, and the recall directions are the top eigenvectors of $\mathbf{C}_{\text{hi}} - \mathbf{C}_{\text{lo}}$. We retain $n_c = 10$ components for subspace construction and cosine analysis, and $n_c = 5$ for the erasure experiments in Appendix B.7. We verified that results are quantitatively stable for $n_c \in \{2, 3, 5, 10, 15, 20\}$. On Pythia-1B, $\cos(\mathbf{d}_K, \mathbf{d}_R)$ varies by less than 0.0002 across this range.

Recall label assignment. For each member text, we compute per-token cross-entropy loss and assign high-recall labels to members in the bottom 50% of the loss distribution. This proxy is validated by Spearman rank correlation with ROUGE-L extraction scores ranging from $\rho = -0.32$ to $\rho = -0.67$ across eight models (all $p < 10^{-24}$). The negative sign reflects the convention that lower loss indicates higher recall. Partial correlation on Pythia-6.9B controlling for loss shows that the recall direction $\mathbf{d}_R$ retains a small but significant association with ROUGE-L ($\rho = +0.105$, $p = 0.001$), while the knowledge direction $\mathbf{d}_K$ does not ($\rho = -0.037$, $p = 0.238$).

Probe training. Membership and recall probes are logistic regression classifiers trained on the projection of activations onto $\mathbf{S}_K$ or $\mathbf{S}_R$, respectively. We report five-fold cross-validated AUROC. Regularisation strength is selected by nested cross-validation.

A.3. Extraction-Detection Separation Protocol

For each model, we perform the intervention at the sparse autoencoder layer. For each condition, we run the full forward pass with the modified activations and evaluate four metrics.

1. Detection AUROC using a loss-based detector, defined as member versus non-member classification based on per-token cross-entropy.
2. Mean member loss, defined as the average per-token cross-entropy over member texts. This acts as a proxy for recall disruption, where higher loss indicates that the model is less confident in reproducing the text.
3. ROUGE-L of greedy continuations from fixed-length prefixes, used as a direct measure of verbatim extraction.
4. Exact match rate, defined as the fraction of member texts whose greedy continuation matches the ground truth exactly.

Erasure is applied at inference time only and does not modify model weights.

Subspace dimension and prefix length. The erasure experiments remove the leading $n_c = 5$ directions of the

relevant subspace. The verbatim-extraction measurement uses ROUGE-L on greedy continuations from fixed-length prefixes, with prefixes truncated to a common length across all models so that prefix exposure is held constant when comparing extraction outcomes across the erasure conditions.

A.4. Feature Ablation Protocol

For the k-ablation experiment, we rank classifier features by the magnitude of their mean activation difference between members and non-members. We then zero out the top-k features in the sparse autoencoder encoding, decode back to the hidden-state space, and continue the forward pass. We vary k from 1 to the total number of classifier features and measure both detection AUROC and extractability, using exact match rate and ROUGE-L.

A.5. Scaling Curve

Table 3. Scaling curve of membership detection across seven Pythia model sizes. All models are fine-tuned under the same protocol on the same data. Mem. AUROC is the membership AUROC of the rank-10 $\mathbf{S}_K$ subspace logistic probe at each model's best layer, the same metric as the per-layer tables and distinct from the rank-1 score_K of Table 1. Membership geometry is absent at 70M parameters and emerges above approximately 400M parameters (Figure 8, right).

Model	Mem. AUROC
Pythia-70M	0.507
Pythia-160M	0.588
Pythia-410M	0.716
Pythia-1B	0.800
Pythia-2.8B	0.842
Pythia-6.9B	0.876
Pythia-12B	0.781

To track when the membership geometry appears, we fine-tune seven Pythia models from 70M to 12B parameters under the same protocol and measure best-layer membership detection at each scale (Table 3, Figure 8 right). At 70M parameters, membership detection is at chance (0.507). The signal rises steadily through 160M (0.588) and 410M (0.716), reaching above 0.80 at 1B parameters, with the strongest detection at 6.9B (0.876). We interpret the transition between 160M and 410M as an approximate emergence threshold, consistent with prior work showing memorisation grows with model scale (Tirumala et al., 2022).

Pythia-12B breaks the monotonic trend, decreasing to 0.781 after peaking at 6.9B. This decrease likely reflects a layer-selection artefact. At layer 18, the channel decomposition detects a score_K of 0.764 (Table 1), and both layers show the same reconstruction/residual pattern (15 percentage points reduction at layer 18 versus 14 percentage points at layer 24).

B. Additional Results

B.1. Worked Example of the Residual Audit

This section expands the inline example in the main results (§4.2, Figure 2) by following one document through the audit on the Pythia-6.9B mixed-data setting. All numbers here are the published values for that setting in Table 1, and no new measurement is introduced.

Take a member document from the controlled OpenWebText split. We read its mean-pooled residual stream activation $\mathbf{h}$ at layer 16, the SAE layer for Pythia-6.9B. The membership detector scores this document by projecting $\mathbf{h}$ onto the fitted knowledge direction $\mathbf{d}_K$. Across the full evaluation pool of 1000 members and 1000 non-members, this original-activation detector reaches membership AUROC 0.803.

We then pass the same activation through the mixed-data SAE. The encoder produces a sparse code, and the decoder maps it back to the reconstruction $\hat{\mathbf{h}} = \mathrm{dec}(\mathrm{enc}(\mathbf{h}))$. The reconstruction is faithful in aggregate, with reconstruction cosine 0.976 on the corpus. Scoring the reconstructed activations with the same $\mathbf{d}_K$ detector gives membership AUROC 0.594, a reduction of about 21 percentage points relative to the original activation. The dictionary representation therefore discards much of the membership separation.

The discarded part is the residual $\mathbf{r} = \mathbf{h} - \hat{\mathbf{h}}$, the dark subspace of the reconstruction. Scoring the residual activations with the same $\mathbf{d}_K$ detector gives membership AUROC 0.781, which recovers most of the separation lost in the reconstruction. The single detector reads three views of the same document, the original activation, the reconstruction, and the residual, and the membership evidence the dictionary does not represent remains recoverable from the residual. The held-out estimation of $\mathbf{d}_K$ (Appendix D.10) preserves this ordering at smaller magnitude.

B.2. Per-Layer Decomposition Tables

Tables 4–12 provide the full per-layer channel decomposition results for each model, including GPT-Neo-2.7B (Table 7).

B.3. Nonlinear Probing Comparison

Table 13 reports the AUROC gap between multi-layer perceptron (MLP) probes (Belinkov, 2022) (two hidden layers, 256 and 128 units, ReLU activations, 5-fold cross-validation) and linear probes across all analysed layers. In five of seven models, the linear probe achieves *higher* mean membership AUROC than the MLP, likely because the MLP's additional parameters overfit the small probe training set. The largest positive gap (MLP better) is 0.037 for OPT-6.7B membership at a single layer. The largest negative gap (linear better) is -0.032 for Pythia-12B mem-

Table 4. Per-layer channel decomposition for Pythia-1B (epoch 5). $\cos(\mathbf{d}_K, \mathbf{d}_R)$ is the cosine between the leading knowledge direction and the leading recall direction at the listed layer. Angle is the mean principal angle between the rank-10 knowledge subspace $\mathbf{S}_K$ and recall subspace $\mathbf{S}_R$. Mem. AUROC and Rec. AUROC are the cross-validated logistic-probe AUROCs for the membership and recall labels using the projection onto $\mathbf{S}_K$ and $\mathbf{S}_R$ respectively. Subsequent per-model tables (Tables 5–12) follow this column convention.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Angle (°)	Mem. AUROC	Rec. AUROC
0	−0.170	20.3	0.487	0.660
2	−0.072	15.1	0.514	0.714
4	−0.082	12.7	0.510	0.737
6	−0.038	17.6	0.524	0.749
8	0.102	16.7	0.577	0.762
10	0.279	21.5	0.686	0.767
12	0.402	20.8	0.739	0.744
14	0.643	23.8	0.800	0.761
15	0.671	25.5	0.771	0.746

Table 5. Per-layer channel decomposition for Pythia-6.9B. Mem. AUROC is the rank-10 $\mathbf{S}_K$ subspace logistic probe, distinct from the rank-1 score_K AUROC reported in Table 1 of the main text.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Angle (°)	Mem. AUROC	Rec. AUROC
0	−0.083	21.0	0.489	0.622
4	−0.013	14.9	0.492	0.750
8	0.026	17.2	0.525	0.751
12	−0.008	20.4	0.689	0.728
16	0.107	21.1	0.828	0.696
20	0.109	21.6	0.876	0.686
24	0.086	26.0	0.861	0.658
28	0.056	26.6	0.860	0.663
31	0.038	29.5	0.864	0.672

Table 6. Per-layer channel decomposition for Pythia-12B. Mem. AUROC is the rank-10 $\mathbf{S}_K$ subspace logistic probe, distinct from the rank-1 score_K AUROC reported in Table 1 of the main text.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Angle (°)	Mem. AUROC	Rec. AUROC
0	−0.063	21.9	0.488	0.684
6	−0.060	14.8	0.513	0.767
12	0.145	21.8	0.574	0.803
18	0.255	21.2	0.741	0.820
24	0.336	24.1	0.781	0.806
30	0.540	26.2	0.742	0.808
35	0.695	25.8	0.756	0.817

Table 7. Per-layer channel decomposition for GPT-Neo-2.7B.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Mem. AUROC	Rec. AUROC
0	−0.060	0.470	0.684
4	−0.001	0.493	0.691
8	0.022	0.503	0.741
12	−0.119	0.497	0.815
16	0.024	0.504	0.813
20	0.140	0.486	0.805
24	0.241	0.526	0.790
28	0.245	0.521	0.788
31	0.078	0.539	0.769

Table 8. Per-layer channel decomposition for OPT-6.7B.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Angle (°)	Mem. AUROC	Rec. AUROC
0	−0.103	19.9	0.500	0.728
4	0.382	20.2	0.499	0.819
8	0.183	15.9	0.495	0.837
12	0.121	20.7	0.514	0.843
16	0.156	23.6	0.548	0.866
20	0.183	25.8	0.753	0.880
24	−0.063	28.4	0.869	0.883
28	0.029	31.1	0.848	0.879
31	0.196	29.1	0.820	0.881

Table 9. Per-layer channel decomposition for Falcon-7B.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Angle (°)	Mem. AUROC	Rec. AUROC
0	−0.191	19.3	0.489	0.760
4	−0.062	16.2	0.479	0.798
8	0.046	15.7	0.500	0.826
12	0.087	22.4	0.549	0.846
16	0.161	23.4	0.635	0.855
20	0.209	24.4	0.690	0.861
24	0.277	26.0	0.811	0.862
28	0.579	27.8	0.829	0.872
31	0.590	26.4	0.971	0.898

Table 10. Per-layer channel decomposition for Mistral-7B.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Angle (°)	Mem. AUROC	Rec. AUROC
0	−0.274	18.7	0.500	0.500
4	−0.306	16.6	0.527	0.793
8	−0.270	25.0	0.759	0.787
12	−0.092	22.8	0.919	0.758
16	−0.010	25.3	0.967	0.714
20	0.029	26.2	0.975	0.661
24	0.011	30.8	0.977	0.623
28	−0.021	35.2	0.997	0.609
31	0.029	28.5	1.000	0.609

Table 11. Per-layer channel decomposition for Llama-3-8B.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Angle (°)	Mem. AUROC	Rec. AUROC
0	−0.093	21.4	0.498	0.610
4	−0.053	15.7	0.541	0.764
8	0.018	18.6	0.723	0.754
12	0.090	19.5	0.908	0.756
16	0.223	22.0	0.957	0.735
20	0.295	25.2	0.948	0.719
24	0.271	26.0	0.954	0.702
28	0.234	27.3	0.980	0.671
31	0.321	29.5	0.999	0.661

Table 12. Per-layer channel decomposition for Qwen2-7B.

Layer	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Angle (°)	Mem. AUROC	Rec. AUROC
0	0.145	17.1	0.497	0.549
4	0.218	16.0	0.509	0.758
8	0.176	14.7	0.550	0.724
12	0.199	15.5	0.620	0.717
16	0.390	17.5	0.803	0.705
20	0.439	16.1	0.941	0.700
24	0.440	20.1	0.940	0.692
27	0.479	27.2	1.000	0.666

Table 13. Nonlinear probing comparison. Mean and max AUROC gap (MLP probe minus linear probe) for membership (mem.) and recall (rec.) labels, computed across all analysed layers. Positive values indicate the two-layer MLP probe outperforms the linear probe. Negative values indicate the linear probe is better. All gaps fall below 5 percentage points, giving no evidence that the tested MLP probes recover materially more signal than the linear probes under this sample size and regularisation regime.

Model	Mean gap (mem.)	Mean gap (rec.)	Max gap (mem.)	Max gap (rec.)
Pythia-1B	−0.013	+0.016	+0.014	+0.028
Pythia-6.9B	−0.023	−0.021	+0.012	−0.002
Pythia-12B	−0.032	−0.009	+0.028	+0.033
GPT-Neo-2.7B	+0.007	+0.010	+0.025	+0.031
OPT-6.7B	−0.007	−0.011	+0.037	+0.021
Falcon-7B	−0.019	−0.000	+0.031	+0.013
Mistral-7B	−0.015	−0.010	+0.004	+0.025

bership. These results provide no evidence that the tested MLP probes recover materially more membership or recall signal than the linear probes under this sample size and regularisation regime.

B.4. Cross-Architecture Transfer Details

To assess whether the K/R subspaces reflect shared geometric structure across architectures, we compute principal angles between the $\mathbf{S}_K$ subspaces (rank $k = 10$) of all model pairs with matching hidden dimension ($d = 4{,}096$). Principal angles measure the best-case alignment between two subspaces after optimal rotation, providing a strict test of geometric similarity.

Computation. Let $\mathbf{Q}_a \in \mathbb{R}^{d\times r_a}$ and $\mathbf{Q}_b \in \mathbb{R}^{d\times r_b}$ be orthonormal bases for the knowledge subspaces $\mathcal{S}_K^{(a)}$ and $\mathcal{S}_K^{(b)}$ of two models with matching hidden dimension d. The principal angles θ_i are read off the singular values of the cross-basis matrix.

$$\sigma_i\big(\mathbf{Q}_a^\top \mathbf{Q}_b\big) = \cos(\theta_i). \tag{4}$$

Larger singular values correspond to smaller principal angles and stronger raw subspace alignment. The test asks whether the two knowledge subspaces overlap in the models' original hidden-state coordinates. It is therefore a strict coordinate-level comparison and does not test whether the same information could be brought into alignment after learning a cross-model map or rotation. We pick rank $k = 10$ to match the channel-decomposition rank used elsewhere, and we compute the singular values via standard SVD of $\mathbf{Q}_a^\top \mathbf{Q}_b$.

Principal angle analysis across three models with matching hidden dimension ($d = 4{,}096$, comprising Pythia-6.9B, OPT-6.7B, and Llama-3-8B) shows no evidence that any model pair shares subspace geometry beyond chance. The mean squared cosine of principal angles ranges from 0.0015 to 0.0025 across all three pairs, indistinguishable from the random baseline of 0.0025 ± 0.0003. This provides no evidence of shared subspace geometry under the strict same-dimensional principal-angle comparison, even though all models show K/R separation.

B.5. Cross-Architecture Measure Readings

We elaborate the intuitive reading of each cross-architecture measure summarised in §3.6.

The first measure asks whether two models put their knowledge subspaces in the same place. Principal angles compare the leading directions of $\mathcal{S}_K^{(a)}$ and $\mathcal{S}_K^{(b)}$ inside a shared ambient vector space, restricted to model pairs at matching hidden dimension ($d = 4{,}096$). Small angles indicate similar directions, while angles near 90° indicate near orthogonality. The comparison is strictly coordinate-level and does not learn a cross-model map.

The second measure asks how spread out a model's member activations are. The participation ratio reads as an effective number of active dimensions, high when variance is distributed across many directions and low when it is concentrated in a few. It summarises whether member activations occupy a broad distributed region of activation space or a more compressed low-dimensional subspace.

The third measure asks whether the membership signal lives inside that dominant subspace at all. The mean-difference vector Δ_{mem} between member and non-member activations points along the direction that separates the average member document from the average non-member document. The cross-architecture concentration C_k reports the fraction of Δ_{mem} captured by the top-k principal component basis. Values near 1 mean the shift lies along the dominant directions of member variation. Values near 0 mean it lies on lower-variance or off-subspace directions.

B.6. Extraction-Detection Separation Across Fine-Tuning Epochs

Table 14. Extraction-detection separation for Pythia-1B at fine-tuning epochs 1, 3, and 5. Det. AUROC is the loss-based membership detector AUROC. ROUGE-L is the verbatim-extraction score on greedy continuations from fixed-length prefixes. Erase $\mathbf{S}_K$ removes the rank-5 knowledge subspace at the SAE layer, Erase $\mathbf{S}_R$ removes the rank-5 recall subspace, Erase both removes both subspaces. At epoch 1 baseline detection is weak and erasure effects are small. By epoch 3 the knowledge subspace carries enough signal for erasure to produce measurable reductions, and extraction ROUGE-L collapses under erasure at all epochs.

Epoch	Metric	Baseline	Erase $\mathbf{S}_K$	Erase $\mathbf{S}_R$	Erase both
1	Det. AUROC	0.677	0.594	0.596	0.502
	ROUGE-L	0.150	0.066	0.076	0.037
3	Det. AUROC	0.993	0.900	0.914	0.705
	ROUGE-L	0.170	0.044	0.053	0.030
5	Det. AUROC	1.000	0.984	0.987	0.900
	ROUGE-L	0.208	0.035	0.034	0.023

Table 14 shows the progression of extraction-detection separation results as fine-tuning proceeds. At epoch 1, baseline

detection AUROC is 0.677, above chance but far from perfect, and subspace erasure reduces it to near chance (0.502 under combined erasure), consistent with a nascent but fragile knowledge signal. By epoch 3, the knowledge subspace has strengthened. Baseline detection reaches 0.993, and erasing $\mathbf{S}_K$ reduces it to 0.900 while erasing $\mathbf{S}_R$ reduces it to 0.914. By epoch 5, the detection signal is strong enough that single-subspace erasure reduces detection only to 0.984 and 0.987, while combined erasure still reduces detection to 0.900. Extraction ROUGE-L collapses under erasure at all epochs, showing that the intervened subspaces are important for verbatim reproduction even when the knowledge signal is still developing.

B.7. Full Metrics for Extraction-Detection Separation

Table 15. Extraction-detection separation with all metrics for all models at fine-tuning epoch 5. Mean member loss is the average per-token cross-entropy on member texts under the modified activations and quantifies generation disruption. Exact match rate is the fraction of member texts whose greedy continuation matches the ground truth exactly and provides a strict extraction measure complementary to ROUGE-L. Erase $\mathbf{S}_K$, Erase $\mathbf{S}_R$, and Erase both denote rank-5 erasure of the knowledge subspace, the recall subspace, and both subspaces at the SAE layer.

Model	Baseline	Erase $\mathbf{S}_K$	Erase $\mathbf{S}_R$	Erase both
Mean member loss				
Pythia-1B	0.565	3.675	3.625	4.734
Pythia-6.9B	0.041	9.025	8.603	9.667
Pythia-12B	0.364	19.136	18.772	20.729
OPT-6.7B	0.412	2.121	2.083	2.251
Exact match rate				
Pythia-1B	0.064	0.008	0.007	0.006
Pythia-6.9B	0.740	0.007	0.010	0.006
Pythia-12B	0.120	0.000	0.000	0.000
OPT-6.7B	0.111	0.018	0.020	0.019

B.8. Directional Steering Results

Directional steering experiments add $\alpha \cdot \hat{d}$ to the residual stream at the sparse autoencoder layer, varying $\alpha \in \{-5, -3, -2, -1, 0, 1, 2, 3, 5\}$ for both the knowledge direction $\mathbf{d}_K$ and the recall direction $\mathbf{d}_R$. For both Pythia-1B and Pythia-6.9B, baseline detection is already near-perfect, so all steering conditions yield detection AUROC at or near 1.000 regardless of direction or magnitude. For Pythia-1B, steering along $\mathbf{d}_R$ produces a larger reduction of exact-match extraction than steering along $\mathbf{d}_K$. The exact-match rate falls from roughly six percent at α=0 to roughly four percent at $|\alpha|$=5 for $\mathbf{d}_R$, compared with a reduction from six percent to five percent for $\mathbf{d}_K$. For Pythia-6.9B, both directions produce negligible changes in extraction, with exact-match rates remaining near 72 percent across all conditions, consistent with the model's stronger memorisation making single-direction perturbations insufficient to disrupt recall. The ceiling effect contrasts with the extraction-detection separation results, where full subspace erasure (rank-5 projection) reduces extraction across the four metrics. This difference is consistent with the knowledge and recall signals being distributed across multiple directions within each subspace rather than concentrated in a single dominant activation direction, although the near-ceiling detection baseline limits what single-direction steering can establish.

B.9. Feature Sufficiency Criterion Values

The feature sufficiency criterion evaluates geometric sufficiency. We ask whether the selected SAE features span the activation channel they are meant to explain. If they do not, then a feature-level intervention has little reason to affect that channel. The criterion is therefore a geometric prerequisite for the causal-effectiveness test that follows.

Value-interpretation axes. The feature sufficiency criterion is bounded between 0 and 1 by construction. Values near 1 mean that the selected SAE features span most of the target channel, so a feature-level intervention that ablates those features can in principle remove most of the channel-aligned variation. Values near 0 mean that the channel lies outside those feature directions, so a feature-level intervention has little geometric reason to affect the channel even when it succeeds at suppressing the targeted features. The criterion therefore separates geometric compatibility (whether the chosen feature directions even contain the target channel) from causal effectiveness (whether ablation actually changes the downstream outcome).

Using all decoder vectors gives FSC of 1.000 for both subspaces in all models (Table 16) when the decoder span contains the channel subspace. The informative comparison is therefore the selected classifier-feature subset, whose FSC remains low across models. The low FSC values for classifier features (0.024–0.154) indicate that the subset of features identified by the statistical test captures only a small fraction of the channel-defined subspaces. To assess whether this low coverage is expected with so few classifier features, we compare each model's observed FSC against 1,000 random feature subsets of matched size. In seven of eight interpretable models, the observed FSC exceeds the 97th percentile of the random baseline, indicating that classifier features are more aligned with $\mathbf{S}_K$ than random dictionary elements. GPT-Neo-2.7B is the exception (observed FSC 0.069, random mean 0.253, percentile 0.0%). Its 954 classifier features produce lower coverage than random subsets of the same size, suggesting that the membership-correlated features in this model are not aligned with the knowledge subspace under this FSC test.

B.10. Per-Model Channel-Decomposition Geometry

This section provides model-specific commentary on the channel decomposition results summarised in Table 17.

Table 16. Feature sufficiency criterion (FSC) for classifier features and the full SAE dictionary relative to the knowledge subspace $\mathbf{S}_K$ and the recall subspace $\mathbf{S}_R$. FSC is bounded between 0 and 1 and measures the fraction of the target subspace spanned by the listed feature set, with values near 1 indicating that the features capture the subspace and values near 0 indicating they do not. n_{CF} is the number of classifier features identified at a Benjamini–Hochberg corrected false discovery rate of 5%. Random Mean reports the mean FSC($\mathbf{S}_K$, CF) from 1,000 random subsets of size n_{CF} drawn from the full dictionary. Pctl. gives the rank of the observed FSC within the random baseline distribution. Llama-3-8B is excluded because the available SAE checkpoint fails the reconstruction-quality criterion. The Pythia-12B entry reports values at the layer-24 channel-geometry reference layer, while Table 1 reports Pythia-12B at the layer-18 SAE evaluation layer.

Model	n_{CF}	**FSC($\mathbf{S}_K$, CF)**	**Random Mean**	**Pctl.**	**FSC($\mathbf{S}_K$, all)**	**FSC($\mathbf{S}_R$, all)**
Pythia-1B	4	0.029	0.003	99.9%	1.000	1.000
Pythia-6.9B	131	0.082	0.030	97.4%	1.000	1.000
Pythia-12B	9	0.024	0.004	99.3%	1.000	1.000
GPT-Neo-2.7B	954	0.069	0.253	0.0%	1.000	1.000
OPT-6.7B	55	0.154	0.017	100.0%	1.000	1.000
Falcon-7B	7	0.066	0.011	100.0%	1.000	1.000
Mistral-7B	73	0.062	0.051	97.3%	1.000	1.000
Qwen2-7B	15	0.109	0.019	100.0%	1.000	1.000

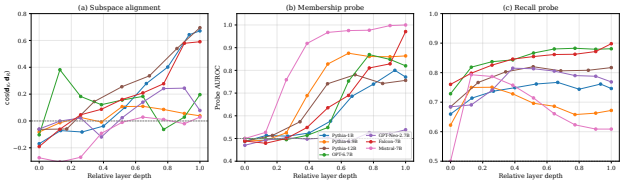

Figure 4. Layer-by-layer probe performance and subspace alignment across nine models. Each panel plots membership-detector AUROC (solid), recall-probe AUROC (dashed), and cosine similarity between the knowledge direction $\mathbf{d}_K$ and the recall direction $\mathbf{d}_R$ (dotted) as a function of transformer layer index. AUROC values are five-fold cross-validated. Knowledge and recall signals develop at different layers, and the two subspace directions maintain low mutual alignment throughout, with cosine similarity rising in late layers only for the smaller-capacity Pythia models (Pythia-1B, Pythia-12B).

Table 17. Channel decomposition at the sparse autoencoder layer for each model. $\cos(\mathbf{d}_K, \mathbf{d}_R)$ is the cosine similarity between the leading knowledge direction and the leading recall direction. Mem. AUROC and Rec. AUROC are five-fold cross-validated logistic-probe AUROCs for membership and recall labels, computed on activation projections onto $\mathbf{S}_K$ and $\mathbf{S}_R$ respectively. Cosine similarity remains below 0.40 in all nine models reported (Gemma-2-2B is reported separately in Appendix D.12), indicating low mutual alignment between the two subspaces. The Pythia-12B entry reports the channel-decomposition measurement at layer 24, the channel-geometry reference layer. Table 1 reports Pythia-12B at the depth-matched layer-18 SAE evaluation layer (§3.4).

Model	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	**Mem. AUROC**	**Rec. AUROC**
Pythia-1B (ep. 5)	0.102	0.577	0.762
Pythia-6.9B	0.107	0.828	0.696
Pythia-12B	0.336	0.781	0.806
GPT-Neo-2.7B	0.024	0.504	0.813
OPT-6.7B	−0.063	0.869	0.883
Falcon-7B	0.161	0.635	0.855
Mistral-7B	−0.010	0.967	0.714
Llama-3-8B	0.223	0.957	0.735
Qwen2-7B	0.390	0.803	0.705

Table 18. Norm and direction decomposition of membership detection. Norm AUROC is the AUROC of a logistic membership detector trained on the L2 norm of mean-pooled hidden activations, reported at the best layer per model. score_K is the AUROC of a detector trained on the projection onto the knowledge direction $\mathbf{d}_K$, reported at the sparse autoencoder layer per model. The Pythia-12B value here (0.707) uses the norm-direction analysis split and per-model evaluation protocol, distinct from the channel-decomposition headline value of 0.764 in Table 1, which is fit on the controlled split at the matched seed-49 SAE configuration. score_K advantage is the difference between the two columns in percentage points. GPT-family models carry membership signal in directional structure with norm AUROC near chance, whereas Llama-style models combine magnitude and direction.

Model	**Norm AUROC**	score_K	score_K **advantage**
GPT-family			
Pythia-6.9B	0.542	0.803	+26 pp
OPT-6.7B	0.534	0.774	+24 pp
Falcon-7B	0.557	0.690	+13 pp
Pythia-1B	0.548	0.660	+11 pp
Pythia-12B	0.599	0.707	+11 pp
GPT-Neo-2.7B	0.514	0.616	+10 pp
Qwen2-7B	0.542	0.638	+10 pp
Llama-style			
Mistral-7B	0.877	0.927	+5 pp
Llama-3-8B	0.843	0.903	+6 pp

Pythia models. In the two models with hidden dimension 4,096 (Pythia-6.9B and OPT-6.7B), cosine similarity remains below 0.11 at the sparse autoencoder layer. Pythia-12B, the largest model in the Pythia family, shows cosine 0.336 at its sparse autoencoder layer. This rises to 0.695 in the final layers, a pattern shared with Pythia-1B, suggesting that the GPT-NeoX architecture entangles the two signals in late layers.

OPT-6.7B. OPT-6.7B achieves a near-zero cosine (-0.063) paired with the highest recall AUROC (0.883), indicating that the two subspaces can both be high-quality and weakly aligned.

GPT-Neo-2.7B. GPT-Neo-2.7B shows near-zero cosine (0.024) with membership detection at chance (AUROC 0.504), showing that the membership signal has not consolidated at this layer and serving as a negative control.

Falcon-7B. Falcon-7B, from a distinct architecture family (multi-query attention, RefinedWeb pretraining), shows cosine 0.161 at its sparse autoencoder layer with the second-highest recall AUROC (0.855) after OPT-6.7B (0.883), providing cross-family evidence of low alignment between the two subspaces.

Mistral-7B. Mistral-7B shows the lowest absolute cosine (-0.010) and the strongest membership detection (AUROC 0.967) among all nine models, showing that minimal alignment between the two directions is compatible with strongly discriminative knowledge representations. The negative sign indicates anti-alignment between the leading directions, supporting the separation of the two subspaces.

Llama-3-8B. Llama-3-8B shows cosine 0.223 at its sparse autoencoder layer with strong membership detection (AUROC 0.957), comparable to Mistral-7B, while maintaining recall AUROC of 0.735.

Qwen2-7B. Qwen2-7B shows the highest cosine among all nine models (0.390) but retains a mean principal angle of $17.5°$, indicating that while the leading directions share more variance than in other models, the full subspaces remain geometrically distinguishable. Despite this higher alignment, Qwen2 achieves membership AUROC of 0.803 and recall AUROC of 0.705, showing that both channels carry discriminative signal even when subspace overlap is moderate.

Interpreting cosine magnitude. In high-dimensional activation spaces ($d = 2048$–5120), random unit vectors have expected cosine similarity near zero, so low cosine between $\mathbf{d}_K$ and $\mathbf{d}_R$ is necessary but not sufficient evidence for a mechanistic separation. The label-shuffled permutation test (Appendix D.5) finds the observed cosines consistent with the random-label baseline on 8 of 9 models, with Mistral-7B falling below baseline. The evidence for separation therefore rests on the causal interventions (§3.5), which show that targeted and random erasures produce qualitatively different outcomes.

Architecture-family AUROC narration. Table 18 resolves the descriptive distinction summarised in §4.5. The GPT-family group has norm AUROC near chance, with values clustered between 0.514 and 0.599 across Pythia-6.9B, OPT-6.7B, Falcon-7B, Pythia-1B, Pythia-12B, GPT-Neo-2.7B, and Qwen2-7B. The channel decomposition adds an additional score_K advantage in this group, ranging from $+10$ to $+26$ percentage points, so the directional component of membership separation is what SAE reconstruction can weaken. The Llama-style group has high norm AUROC, with Mistral-7B at 0.877 and Llama-3-8B at 0.843, and the channel decomposition adds only marginal further separation ($+5$ to $+6$ percentage points). SAE reconstruction therefore appears to preserve the magnitude component while weakening the directional component measured by the channel decomposition.

Linearity validation. To test whether the linear subspace decomposition misses discriminative nonlinear structure, we compare linear and nonlinear probes at each layer across seven models. Two-layer MLP probes (256/128 hidden units) achieve mean AUROC gains of at most 0.037 over logistic regression for membership and 0.025 for recall. In five of seven models the MLP performs *worse* than the linear probe for membership (Table 13). The MLP gap is below 0.04 across all models, giving no evidence that the tested nonlinear probe materially improves over the linear decomposition under this protocol.

B.11. Per-Model Extraction-Detection Separation

This section provides model-specific commentary on the extraction-detection separation results summarised in Table 15.

OPT-6.7B. OPT-6.7B presents a contrasting case. Detection remains above 0.999 under all erasure conditions, while extraction falls by 57% (from 0.288 to 0.123). This suggests that OPT encodes the knowledge signal more diffusely across layers or in a higher-dimensional subspace not fully captured by rank-5 erasure at layer 24. The extraction disruption nevertheless shows that the recall subspace carries generation-relevant information.

Pythia-12B. Pythia-12B, which has the highest subspace cosine among the Pythia models (0.336), shows no differential separation between channels. Both $\mathbf{S}_K$ and $\mathbf{S}_R$ erasure

reduce detection to 0.647 and 0.644 respectively, with the combined erasure reaching 0.602. This symmetric response is consistent with the greater overlap between the two subspaces at this model scale.

GPT-Neo-2.7B. GPT-Neo-2.7B is omitted from the main table because its baseline detection is at chance (AUROC 0.464), making the extraction-detection separation uninformative. Subspace erasure increases member loss fourfold in all conditions, but there is no detection signal to selectively disrupt.

Falcon-7B. Falcon-7B shows minimal extraction-detection separation selectivity (detection 0.999 under both erasures, $\Delta_{\text{det}} = 0.0002$), despite its low cosine (0.161). Extraction falls from 0.216 to 0.140 under combined erasure, consistent with its weaker memorisation baseline.

Mistral-7B. Mistral-7B, despite minimal alignment between $\mathbf{d}_K$ and $\mathbf{d}_R$ (cosine -0.010) and the strongest membership signal among all models (AUROC 0.967), shows no detection disruption under any erasure condition, while extraction falls from 0.843 to 0.162 under combined erasure. Like OPT, this suggests that the detection signal is encoded diffusely across layers rather than concentrated in the rank-5 subspace at layer 16.

Qwen2-7B. Qwen2-7B, which has the highest subspace cosine (0.390), shows detection resilience similar to OPT and Mistral (AUROC ≥ 0.996 under all conditions), while extraction falls from 0.658 to 0.128 under combined erasure.

B.12. Extraction ROUGE-L for Extraction-Detection Separation

Table 19. Extraction ROUGE-L under four erasure conditions for the extraction-detection separation experiment. Baseline is the ROUGE-L score on greedy continuations from fixed-length prefixes with no intervention. Erase $\mathbf{S}_K$ and Erase $\mathbf{S}_R$ remove the rank-5 knowledge and recall subspaces respectively at the SAE layer. Erase both removes both subspaces. Both single-subspace erasures reduce extraction across all models, and GPT-Neo-2.7B is omitted because baseline detection is at chance.

Model	Baseline	Erase $\mathbf{S}_K$	Erase $\mathbf{S}_R$	Erase both
Pythia-1B	0.208	0.035	0.034	0.023
Pythia-6.9B	0.751	0.029	0.033	0.027
Pythia-12B	0.271	0.000	0.000	0.000
OPT-6.7B	0.288	0.123	0.118	0.115
Falcon-7B	0.216	0.173	0.163	0.140
Mistral-7B	0.843	0.673	0.657	0.162
Llama-3-8B	0.700	0.015	0.034	0.005
Qwen2-7B	0.658	0.144	0.129	0.128

Extraction ROUGE-L falls under subspace erasure across all models, from baselines of 0.21–0.84 to below 0.035 in all Pythia models and Llama-3, and to near-zero in Pythia-12B. This indicates that both subspaces carry information needed for autoregressive generation at the intervention layer. The combined erasure produces the largest detection reductions and the lowest extraction scores, supporting the view that the two subspaces carry complementary rather than redundant information.

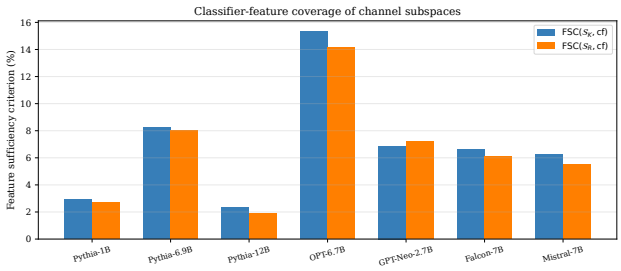

Figure 5. Feature sufficiency criterion (FSC) for classifier features relative to the knowledge subspace $\mathbf{S}_K$ and recall subspace $\mathbf{S}_R$. FSC measures the fraction of the target subspace spanned by the listed feature set, bounded between 0 and 1, with values near 1 indicating that the features capture the subspace. Across all models, classifier features capture less than 16% of either subspace, and the smaller models show near-zero coverage.

B.13. Sparse Autoencoder Feature Analysis

The feature sufficiency criterion (Figure 5) shows that selected classifier features span only a small fraction of the knowledge and recall subspaces. The largest coverage occurs in OPT-6.7B (FSC 0.154 for $\mathbf{S}_K$, 0.142 for $\mathbf{S}_R$) and Qwen2-7B (FSC 0.109 for $\mathbf{S}_K$, 0.107 for $\mathbf{S}_R$), while Pythia-12B shows the lowest among larger models (FSC 0.024 for $\mathbf{S}_K$, 0.019 for $\mathbf{S}_R$). The selected SAE classifier features cover only a small fraction of the channel-defined subspaces, even though the full decoder span can geometrically contain them. This separates geometric compatibility from feature-level recoverability. Causal ablation of sparse autoencoder features is therefore not sufficient evidence that a model has or has not memorised specific training data, because the selected features do not cover the relevant channel subspaces. The classifier features do carry information relevant to recall, as shown by the extractability cliff, but they are insufficient to capture the full subspaces that drive either membership detection or verbatim extraction. Full FSC values for all models appear in Table 16. Pythia-1B shows the same pattern as Pythia-6.9B. Its four classifier features have no measurable effect on detection.

Top-k feature ablation protocol. The norm and feature-subset analyses summarised in §4.3 use two complementary verification protocols on the reconstructed activation. The first checks whether residual magnitude alone explains membership signal. We L2-normalise each residual vector to unit length before training the membership detector. If the result reduced to a residual norm artefact, normalisation would collapse the AUROC. Instead, the directional membership signal is preserved (Table 23), so the residual detector cannot be reduced to a simple norm artefact. The

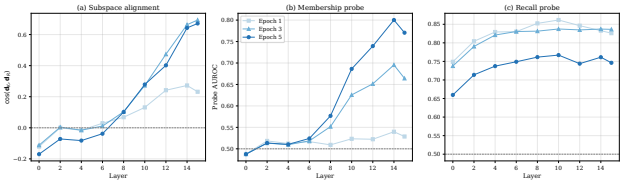

Figure 6. Channel decomposition across training stages. Left three panels show layer-by-layer membership AUROC (solid), recall AUROC (dashed), and $\cos(\mathbf{d}_K, \mathbf{d}_R)$ (dotted) for Pythia-1B at fine-tuning epochs 1, 3, and 5. At epoch 1 the recall subspace is already well-formed while the knowledge subspace remains at chance, and by epoch 5 both subspaces are present but begin to align in late layers.

second protocol checks whether the membership evidence is carried by a small set of the most member-associated SAE features. We rank classifier features by point-biserial correlation with the binary membership label, then ablate the top k features for k swept across the full classifier-feature set. If the membership evidence were carried by such a small set, ablation should either recover or remove that evidence. Instead, ablation further weakens the reconstructed membership signal rather than recovering it, indicating that the tested feature subset is not the relevant carrier.

B.14. Recall channel emerges before knowledge channel during fine-tuning

Table 20. Developmental path of channel decomposition at the sparse autoencoder layer. $\cos(\mathbf{d}_K, \mathbf{d}_R)$ is the cosine between the leading knowledge and recall directions. Mem. AUROC and Rec. AUROC are five-fold cross-validated logistic-probe AUROCs for membership and recall labels. The top block reports fine-tuning epochs for Pythia-1B, and the bottom block reports pretraining checkpoints for Pythia-1B with no fine-tuning. The recall subspace emerges before the knowledge subspace in both regimes.

Stage	Checkpoint	$\cos(\mathbf{d}_K, \mathbf{d}_R)$	Mem. AUROC	Rec. AUROC
Fine-tuning (Pythia-1B)				
	Epoch 1	0.068	0.509	0.852
	Epoch 3	0.103	0.552	0.832
	Epoch 5	0.102	0.577	0.762
Pretraining (Pythia-1B, no fine-tuning)				
	Step 10K	0.005	0.482	0.937
	Step 50K	0.080	0.489	0.935
	Step 100K	0.165	0.491	0.938
	Step 143K	0.204	0.489	0.943

Comparing Pythia-1B across training stages shows an asymmetry in how the two subspaces form (Figure 6, Table 20).

Fine-tuning dynamics. At epoch 1, the recall subspace is already well-developed (AUROC 0.852) while the knowledge subspace is absent (AUROC 0.509, indistinguishable from chance). By epoch 3, membership detection begins to emerge (AUROC 0.552 at the sparse autoencoder layer, 0.695 at the best layer), while recall remains strong. By epoch 5, both subspaces are present, but cosine similarity rises above 0.6 in late layers, suggesting that limited representational capacity forces the two signals to share dimensions. Extraction-detection separation experiments across epochs confirm this progression. At epoch 1 the SAE-layer channel detector is near chance (Table 14), while by epoch 3 erasing $\mathbf{S}_K$ reduces detection from 0.993 to 0.900.

Pretraining dynamics. The same asymmetry appears in pretraining. At step 10,000, before any fine-tuning, the recall subspace already separates high-loss from low-loss texts with AUROC 0.937, while membership detection is at chance (0.482). The recall proxy (loss-based thresholding) at this stage measures text difficulty, not memorisation-specific recall. The finding is that the text-confidence encoding that later underlies recall is present from early pretraining, while the membership signal requires exposure to the fine-tuning set. The recall AUROC reaches 0.943 by step 143,000, and the cosine similarity between $\mathbf{d}_K$ and $\mathbf{d}_R$ grows from 0.005 to 0.204, indicating that the two directions separate as pretraining proceeds. This suggests that the recall subspace reflects a general property of how transformers encode text confidence, not an artefact of fine-tuning, while the knowledge subspace only develops once the model has been exposed to the fine-tuning examples.

B.15. Additional Details for Intrinsic Dimensionality

Participation ratio. We estimate the effective dimensionality of the member activation distribution from the singular spectrum of the centred member activations at the analysis layer. For each model, with singular values $\{s_i\}_{i=1}^{d}$, we report the participation ratio

$$\mathrm{PR} = \frac{\left(\sum_{i=1}^{d} s_i^2\right)^2}{\sum_{i=1}^{d} s_i^4}, \tag{5}$$

which can be read as an effective number of active dimensions. A high PR means variance is distributed across many directions, while a low PR means variance is concentrated in a few dominant directions.

Membership direction and top-k concentration. The mean-difference vector

$$\Delta_{\mathrm{mem}} = \bar{\mathbf{h}}_{\mathrm{mem}} - \bar{\mathbf{h}}_{\mathrm{non}}, \tag{6}$$

where $\bar{\mathbf{h}}_{\mathrm{mem}}$ and $\bar{\mathbf{h}}_{\mathrm{non}}$ are the mean activations of member and non-member documents at the analysis layer, points along the direction that separates the average member from the average non-member.

Let $\mathbf{V}_k$ contain the top k right singular vectors of the centred member activations. These vectors span the leading principal component subspace, the directions that explain the largest variation among member examples. We measure the alignment of the membership direction with this

subspace as

$$C_k = \frac{\left\|\mathbf{V}_k^\top \Delta_{\mathrm{mem}}\right\|_2}{\left\|\Delta_{\mathrm{mem}}\right\|_2}, \tag{7}$$

the fraction of the membership-direction length captured by the top-k principal component subspace. Values near 1 mean that the member–non-member shift is concentrated along the dominant directions of member variation. Values near 0 mean that the shift lies outside those directions, suggesting that membership information is carried by lower-variance or off-subspace directions. We report C_k at $k = 10$ to match the channel-decomposition rank.

Architecture-family grouping. The architecture-family groups in Figure 8 are GPT-NeoX (Pythia-1B, Pythia-6.9B, Pythia-12B), GPT-Neo (GPT-Neo-2.7B), OPT (OPT-6.7B), Falcon (Falcon-7B), Mistral (Mistral-7B), LLaMA (Llama-3-8B), Qwen (Qwen2-7B), and Gemma (Gemma-2-2B). The grouping uses pretraining-corpus and architecture family, not parameter count. The Kruskal–Wallis test below uses these groups as labels.

OPT-6.7B as outlier. OPT-6.7B is a consistent outlier. Its participation ratio (103.2) exceeds the next-highest model (Falcon-7B, 87.2) by 18%, and its rank_{90} (329) is the highest in the study. This diffuse encoding is consistent with OPT's response in the extraction-detection separation experiments, where rank-5 erasure failed to disrupt detection (AUROC remained > 0.999). The intrinsic-rank analysis provides one explanation. OPT spreads the knowledge signal across roughly twice as many dimensions as the Pythia models, so low-rank erasure is insufficient.

Relation to the baseline model. Under a uniform encoding hypothesis where knowledge is spread equally across all d dimensions, the expected concentration of the top 10 principal components would be $10/d \approx 0.002$ for $d = 4096$. The observed concentrations (0.35–0.62) exceed this baseline by 150–250×, rejecting the uniform-encoding baseline. However, the values fall well below a low-rank hypothesis (> 0.85), indicating that knowledge encoding is structured but not sparse. This medium-rank regime constrains privacy interventions. Erasing a small number of directions is insufficient to remove the knowledge signal, as confirmed by the extraction-detection separation results for OPT, Mistral, and Qwen2, but the signal is concentrated enough that targeted subspace methods remain more effective than random perturbation.

Architecture-family concentration analysis. As a descriptive analysis of the architecture-family pattern in Figure 8, we run a Kruskal-Wallis test on per-layer concentration values across architecture-family groups. The test rejects the no-difference hypothesis at $p = 0.006$. We do not interpret this as confirmatory evidence for an architecture-family effect because per-layer values within a model are not independent, and the Llama-style group contains only two models. The result is therefore used only as a descriptive summary of the architecture scatter, not as an independent statistical claim.

B.16. Sparse-Code Adequacy (z versus h)

We test whether a logistic membership detector trained on SAE codes $z = \mathrm{enc}(h)$ recovers the signal that the same detector extracts from the raw mean-pooled hidden state h. The adequacy criterion is pre-specified as $\mathrm{AUROC}(z) \geq \mathrm{AUROC}(h) - \epsilon$ with $\epsilon = 0.05$. The decision is by paired one-sided t-test across five stratified folds, with Holm-Bonferroni correction (Holm, 1979) across the family of valid models. Quality gates exclude any model whose SAE reconstruction cosine on the corpus falls below 0.85 (no faithful z) or has fewer than 100 unique nonzero features (near-dead dictionary). After excluding OPT-6.7B, Mistral-7B, and Pythia-1B for failing the reconstruction-cosine gate, the valid set is Pythia-6.9B, Pythia-12B, Qwen2-7B, and GPT-Neo-2.7B. On Pythia-6.9B, Pythia-12B, and Qwen2-7B the z-probe AUROC sits at chance while the h-probe AUROC reaches the standard activation level, with paired-fold t-statistics far on the inadequate side (Holm-adjusted $p \approx 1$ for each). On GPT-Neo-2.7B the h-probe is itself near chance at the SAE layer, so the comparison is uninformative. The pre-specified adequacy criterion fails on every valid model where the comparison is informative. The SAE encoding does not preserve the membership signal that lives in the raw hidden state.

B.17. Permutation Null for Full-Vector Probes

To calibrate the fresh logistic regression probes against overfitting (since $p/n \approx 2$ for $d = 4096$ and $N = 2000$), we run a permutation test for each SAE condition. For each of 50 label shuffles, a fresh L2-regularised logistic regression probe ($C = 1.0$, 5-fold stratified cross-validation) is trained on the full activation vector with randomly permuted membership labels. The resulting permutation AUROC distribution provides a direct calibration of the chance floor under the same dimensionality and sample size.

Privacy-aware SAE fine-tuning hyperparameters. The three privacy-aware SAE conditions use the $\mathbf{d}_K$-targeted reconstruction objective from Equation 1. Fine-tuning runs for 5,000 steps at learning rate 10^{-4} on the same mixed corpus used to train the baseline mixed-data SAE, with $\lambda_{d_K} \in \{0.1, 1.0\}$ across the three privacy-aware conditions. The membership direction $\mathbf{d}_K$ used in the penalty is estimated on a held-out audit partition that is disjoint from the partition used for evaluation. Evaluation runs on the disjoint partition under the same labelling protocol, so the fitted

$\mathbf{d}_K$ direction is not exposed to the documents on which detection AUROC is reported.

Table 21 reports the full results across four SAE conditions, each decomposed into original, SAE-reconstructed, and residual compartments. The permutation calibration supports two conclusions. All probe AUROCs on true labels lie far above the corresponding permutation distribution mean (0.500 to 0.506), showing that the probes detect signal beyond the permutation-calibrated chance floor. The divergence between score_K and probe AUROC under privacy-aware SAE training (the two Pythia-6.9B privacy-aware conditions) also persists after permutation calibration. For the baseline mixed-data SAE, both score_K and probe performance fall together in the reconstructed compartment (0.648 and 0.692), consistent with genuine weakening of membership-detection performance. For the two Pythia-6.9B privacy-aware SAEs, the reconstructed detector remains near the original level (0.778 to 0.831) even as score_K stays at or near its original value, while the residual detector remains high (0.802 to 0.836) even though residual score_K collapses to 0.520 and 0.537. This indicates that the membership signal extends beyond the single $\mathbf{d}_K$ direction and that privacy-aware SAE training suppresses only the $\mathbf{d}_K$ component rather than the full signal. The GPT-Neo-2.7B privacy-aware SAE shows a contrasting pattern in which the original probe is at chance (0.499), indicating that its membership geometry is concentrated along $\mathbf{d}_K$ rather than distributed across multiple dimensions.

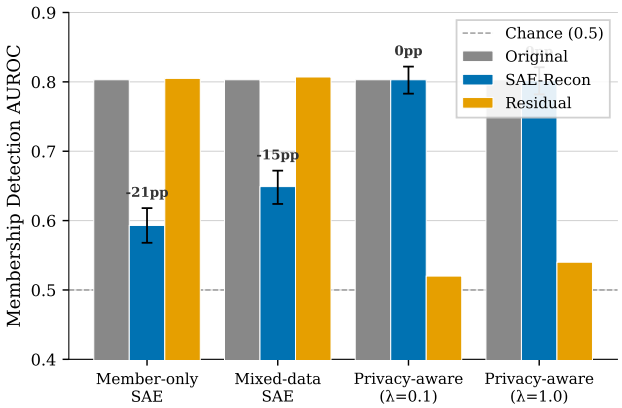

Figure 7. Membership detection before and after SAE reconstruction on Pythia-6.9B. score_K is the AUROC of a logistic detector trained on the projection onto the knowledge direction $\mathbf{d}_K$. Standard SAEs reduce score_K on the reconstructed activation, and direction-penalised SAEs (trained with a $\mathbf{d}_K$-projection penalty) recover score_K on the reconstruction but leave membership signal detectable in the residual.

B.18. Qwen2 Mixed-Data SAE Falls Below the Strict Reconstruction Threshold

B.18.1. Mixed-Data Qwen2 SAE Evaluation

A single-seed evaluation on Qwen2-7B at the L1 = 5e-4 baseline hyperparameter, matched to the Qwen2-7B member-only SAE configuration with layer 16, $d_{\text{sae}}/d_{\text{model}}$ multiplier 4, and $d_{\text{sae}} = 14{,}336$, produced reconstruction cosine 0.889, below the 0.90 pre-specified reconstruction-quality threshold. Mean L0 activation count was 4.93 features. The observed AUROC reduction was 0.0475, computed as the difference between the original score_K AUROC of 0.6381 on raw activations and the 0.5907 AUROC on SAE-reconstructed activations. Residual AUROC (0.6579) exceeded the original AUROC (0.6381), matching the residual-above-original pattern seen in the Qwen2-7B member-only entry of Table 1.

An additional evaluation at L1 = 3e-4 is indicated to test whether a less aggressive penalty would clear the 0.90 reconstruction-cosine threshold while retaining a measurable reduction, but we do not include that evaluation. The Qwen2 single-seed evaluation is reported here for transparency. It is not used in any main-body claim and does not enter the within-model causal evidence on Pythia-6.9B.

B.18.2. SAE Reconstruction-Quality Audit

Table 22 itemises the SAE configurations attempted for the matched-hyperparameter mixed-data control across candidate models. Only Pythia-6.9B passed the 0.90 reconstruction-cosine threshold at the configuration that matched the corresponding member-only SAE. Three models (GPT-Neo-2.7B, Mistral-7B, Llama-3-8B) failed the threshold across all attempted configurations. Qwen2-7B did not clear the threshold at the baseline L1 = 5e-4 in this single-seed evaluation.

The full GPT-Neo-2.7B L1 grid is omitted from the main table because all tested settings failed the quality criterion.

C. Additional Figures

This section collects supporting figures. Figure 8 shows the norm-versus-direction scatter and the parameter-scaling curve. Figure 9 reports membership AUROC across nine models for the original, reconstructed, and residual activations. Figure 10 shows the full member and non-member score distributions on Pythia-6.9B under the member-only and mixed-data SAEs. Figure 11 plots SAE reconstruction quality against the reconstruction-induced AUROC reduction. Figure 12 separates membership detection into norm and directional contributions by architecture family, and Figure 13 reports the per-layer membership concentration for each model.

Table 21. Fresh-probe permutation test for full-vector membership detectors. score_K is the AUROC of the detector trained on the projection onto the knowledge direction $\mathbf{d}_K$. Probe is the AUROC of an L2-regularised logistic regression trained on the full activation vector (five-fold CV, $C = 1.0$, $N = 2000$). Null Mean and Null Std are the mean and standard deviation of probe AUROC across $n = 50$ label shuffles. Signal is the difference between Probe and Null Mean. The Pythia-6.9B mixed-data condition uses a mixed-data SAE (reconstruction cosine 0.960). The two privacy-aware conditions use SAEs trained with the $\mathbf{d}_K$-projection penalty at $\lambda_{d_K} = 0.1$ (reconstruction cosine 0.943) and $\lambda_{d_K} = 1.0$ (reconstruction cosine 0.975). The GPT-Neo-2.7B condition uses a privacy-aware SAE at $\lambda_{d_K} = 1.0$ (reconstruction cosine 0.998).

Condition	Compartment	score_K	Probe	Null Mean	Null Std	Signal
Pythia-6.9B, mixed SAE	Original	0.803	0.819	0.501	0.017	+0.318
	Reconstructed	0.648	0.692	0.506	0.016	+0.186
	Residual	0.807	0.811	0.500	0.017	+0.311
Pythia-6.9B, λ_{d_K}=0.1	Original	0.803	0.815	0.501	0.017	+0.314
	Reconstructed	0.803	0.831	0.504	0.017	+0.327
	Residual	0.520	0.802	0.500	0.018	+0.302
Pythia-6.9B, λ_{d_K}=1.0	Original	0.803	0.815	0.501	0.017	+0.314
	Reconstructed	0.798	0.778	0.504	0.020	+0.274
	Residual	0.537	0.836	0.501	0.018	+0.335
GPT-Neo-2.7B, λ_{d_K}=1.0	Original	0.616	0.499	0.506	0.018	−0.007
	Reconstructed	0.612	0.583	0.501	0.013	+0.082
	Residual	0.565	0.549	0.506	0.019	+0.043

Table 22. Reconstruction-quality audit across attempted matched-hyperparameter mixed-data SAE configurations. Pass criteria are reconstruction cosine ≥ 0.90, mean L0 $< 0.6 \cdot d_{\text{sae}}$, and AUROC reduction ≥ 0.05. recon cos is reconstruction cosine on the corpus, L0 is mean active features per token, and AUROC reduction is the difference between original score_K AUROC and SAE-reconstructed score_K AUROC. Llama-3-8B is excluded under the criterion at every layer attempted, and the full GPT-Neo-2.7B L1 grid failed across five settings, only the primary configuration is shown.

Model	Config	recon cos	L0	AUROC reduction	Outcome
Pythia-6.9B	mult=4, L1=5e-4 (N=5 mean)	0.976	11.20	0.2094	PASS
Qwen2-7B	mult=4, L1=5e-4 (N=1)	0.889	4.93	0.0475	FAIL (recon)
Mistral-7B	mult=4, L1=2e-4	0.7571	2.39	0.0127	FAIL (recon, reduction)
Mistral-7B	mult=4, L1=3e-4	0.7236	1.95	0.0466	FAIL (recon, reduction)
GPT-Neo-2.7B	mult=4, L1=5e-4	—	515–2912	—	FAIL (L0)
Llama-3-8B	member-only checkpoint	−0.03	—	—	FAIL (recon)

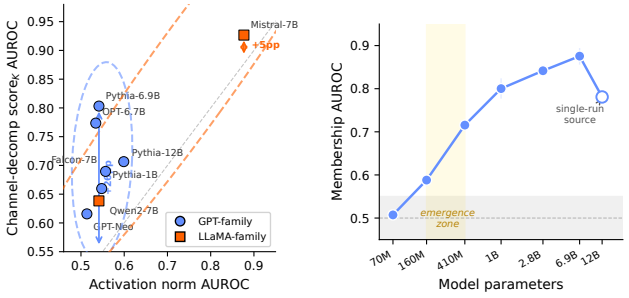

Figure 8. Norm versus directional contributions to membership detection (left), and the scaling curve across Pythia sizes (right). **Left panel** plots norm AUROC against score_K for nine models by architecture family. Norm AUROC uses the L2 norm of mean-pooled hidden activations as the detector input, and score_K uses the projection onto the knowledge direction $\mathbf{d}_K$. GPT-family models cluster at low norm AUROC with high directional signal, and Llama-style models reach high norm AUROC with smaller directional gains. **Right panel** shows best-layer membership AUROC across seven Pythia sizes from 70M to 12B parameters. The membership geometry emerges above approximately 400M parameters (Appendix A.5).

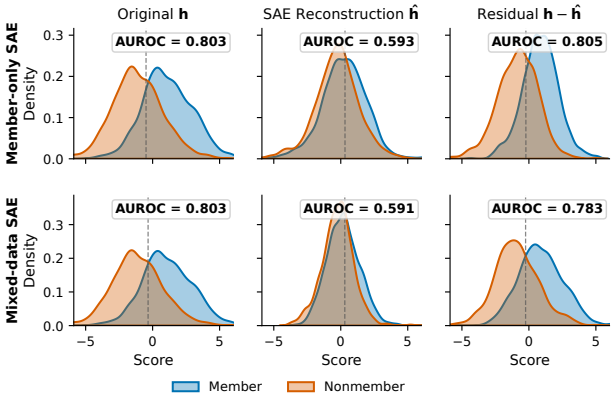

Figure 10. score_K distributions for members and non-members on Pythia-6.9B under two SAE types. **Top panels** report the member-only SAE on a single seed for visualisation, with the five-seed average reconstruction AUROC of 0.63 reported in Table 1. The top centre panel shows that member and non-member separation collapses on this seed, with AUROC moving from 0.80 to 0.59. **Bottom panels** report the mixed-data SAE control, where separation still collapses to AUROC 0.59 and argues against biased SAE training as the explanation. In both cases, probes recover the separation from the reconstruction residual (right panels).

	Original	SAE Recon.	Residual
Pythia-6.9B	0.803	0.633	0.833
Pythia-6.9B (mixed)	0.803	0.594	0.781
Pythia-1B	0.663	0.593	0.670
GPT-Neo	0.616	0.572	0.699
OPT-6.7B	0.774	0.557	0.866
Pythia-12B	0.764	0.612	0.723
Mistral-7B	0.927	0.579	0.926
Qwen2-7B	0.638	0.526	0.849
Falcon-7B†	0.690	0.694	0.827

Membership AUROC

Figure 9. Annotated heatmap of score_K membership AUROC across nine models and three activation compartments. Original is the raw mean-pooled hidden activation, Reconstructed is the SAE-reconstructed activation, and Residual is the difference (original minus reconstruction). The Pythia-6.9B mixed-data setting uses an SAE trained on both member and non-member activations as a control for training-set bias. The centre column reports reduced membership detection after SAE reconstruction across gate-passing settings. Pythia-1B, OPT-6.7B, Mistral-7B, and Falcon-7B are reported as diagnostic context and do not pass the reconstruction gate.

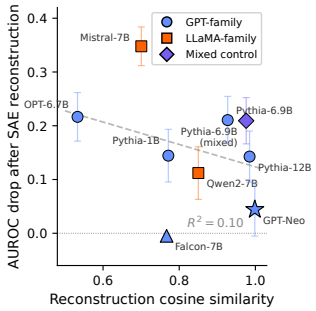

Figure 11. SAE reconstruction quality (reconstruction cosine) plotted against membership AUROC reduction. AUROC reduction is the difference between original score_K AUROC and SAE-reconstructed score_K AUROC. Each point is one model. There is no significant correlation ($R^2 = 0.10$), arguing against the reconstruction-induced reduction being an artefact of poor SAE reconstruction. GPT-Neo-2.7B reaches near-perfect reconstruction (cosine 0.998) yet still loses 4 percentage points of membership detection.

Table 23. L2-normalised residual membership AUROC. Raw Resid is the residual score_K AUROC computed on the unit-normalised residual vector, L2 Resid is the residual AUROC after L2-normalising each residual to unit length, and Norm Resid is the residual norm AUROC computed on the L2 norm alone. Δ(L2−Raw) is the difference between the L2-normalised and raw residual scores. The asterisk marks the three models with elevated residual norm AUROC, where L2 normalisation moves the directional score upward and indicates that the norm component partially masks directional signal. Per-model rows report the original single-seed evaluation, and five-seed averages for the controlled Pythia settings are in Table 1.

Model	Raw Resid	L2 Resid	Norm Resid	Δ(L2−Raw)
Pythia-1B	0.677	0.676	0.547	−0.001
GPT-Neo-2.7B	0.699	0.698	0.518	−0.001
Pythia-6.9B	0.805	0.806	0.511	+0.001
Mistral-7B	0.926	0.925	0.692	−0.001
OPT-6.7B*	0.866	0.873	0.958	+0.007
Pythia-12B*	0.814	0.831	0.995	+0.017
Qwen2-7B*	0.849	0.858	0.982	+0.009

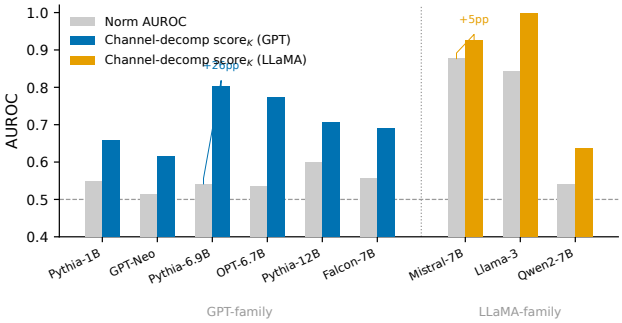

Figure 12. Norm versus directional contributions to membership detection by architecture family. Norm AUROC uses the L2 norm of mean-pooled hidden activations as the detector input, and score_K uses the projection onto the knowledge direction $\mathbf{d}_K$. Gap annotations report the directional contribution as score_K minus norm AUROC. GPT-family models carry membership signal almost entirely in directional structure with norm AUROC near chance (0.51 to 0.60), whereas Llama-style models encode membership in both magnitude and direction.

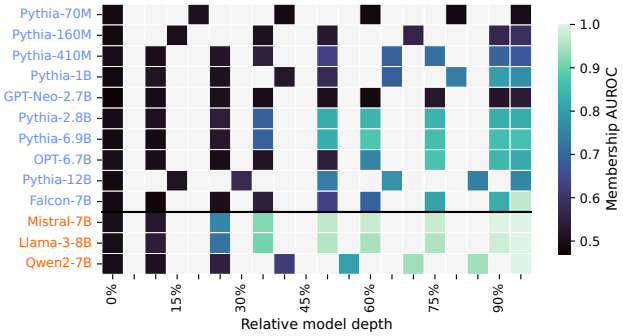

Figure 13. Membership detection AUROC across all layers for each model, with layer index normalised by model depth. Each row corresponds to one model. Darker shading indicates stronger membership signal. The layer-wise concentration pattern differs by architecture family. GPT-family models carry signal across a broad mid-layer band, whereas Llama-style models concentrate the signal in narrower peaks.

D. Robustness and Boundary Conditions

D.1. Held-Out Estimation of $\mathbf{d}_K$

The Pythia-6.9B in-partition mean reduction is 0.213, and the Pythia-12B in-partition mean reduction is 0.160 (held-out probe at the seed-49 SAE anchor, which differs from the 0.152 mean across the five SAE seeds 47–51 reported in Appendix D.11). Neither Pythia model meets the pre-specified magnitude criterion for the held-out reduction, although the qualitative residual-above-reconstruction ordering is preserved.

D.2. Paired Bootstrap for the Directional Sign-Test Settings

Every setting in this group has residual AUROC above reconstruction AUROC with the 95 percent CI excluding zero, consistent with the residual-above-reconstruction ordering reported in the main text.

Binomial sign-test arithmetic. The directional sign test uses the settings that pass the validity gate in Table 1 as its denominator, with each setting counted once. The non-Pythia subset (GPT-Neo-2.7B, GPT-Neo-2.7B mixed-data, Qwen2-7B, Gemma-2-2B) yields one-sided $p \approx 0.0625$ under the chance baseline. The full gate-passing set yields $p \approx 0.008$.

D.3. Bag-of-Words Vocabulary Baseline

A natural concern is that the residual detector might be detecting surface vocabulary rather than internal model structure. We test that explanation directly with a bag-of-words classifier on the same 1000 member and 1000 non-member documents used for the controlled Pythia-6.9B reconstruction/residual evaluation. The classifier is a logistic regression trained on TF-IDF features with unigrams and bigrams under five-fold stratified cross-validation, and AUROC is computed after pooling the out-of-fold predictions.

The pooled AUROC is 0.4566, rounded to 0.457, so this token-level baseline does not distinguish members from non-members on the controlled split. The result argues against a simple unigram vocabulary explanation of the residual signal. It does not rule out higher-order surface artefacts, since the bag-of-words representation discards word order and context.

D.4. Length-Feature Baseline

We compute three length-derived membership classifiers on the controlled OpenWebText split. The classifiers use character length, word count, and vocabulary richness (unique-token fraction). On $n_{\text{member}} = n_{\text{nonmember}} = 1000$, the AUROCs are 0.502 (character length), 0.509 (word count),

Table 24. Per-split held-out reductions from the partition-fit probe for the knowledge direction $\mathbf{d}_K$ (§3.7). Each split estimates $\mathbf{d}_K$ on a random 70 percent of the canonical $N = 2000$ labelled pool and evaluates the score_K reduction on the disjoint 30 percent held-out partition. Held-out reduction is the difference between original score_K AUROC and SAE-reconstructed score_K AUROC on the held-out partition. Bootstrap 95% CIs use $n_{\mathrm{boot}} = 10\,000$ paired resamples. $\Delta_{\mathrm{HO-IP}}$ reports the per-split held-out reduction minus the in-partition mean reduction on the same model. recon_cos is the SAE reconstruction cosine on the corpus.

Model	split_seed	Held-out reduction	95% CI	$\Delta_{\mathrm{HO-IP}}$	recon_cos
Pythia-6.9B layer 16, seed 43, mult=4 L1=5e-4					
Pythia-6.9B	0	0.145	[0.106, 0.184]	−0.068	0.979
Pythia-6.9B	1	0.171	[0.123, 0.220]	−0.042	0.979
Pythia-6.9B	2	0.166	[0.123, 0.210]	−0.047	0.979
Pythia-6.9B	3	0.199	[0.155, 0.243]	−0.014	0.979
Pythia-6.9B	4	0.135	[0.088, 0.181]	−0.078	0.979
Pythia-6.9B	5	0.158	[0.117, 0.201]	−0.055	0.979
Pythia-6.9B	6	0.155	[0.113, 0.198]	−0.058	0.979
Pythia-6.9B	7	0.149	[0.111, 0.187]	−0.064	0.979
Pythia-6.9B	8	0.054	[0.012, 0.097]	−0.159	0.979
Pythia-6.9B	9	0.157	[0.117, 0.197]	−0.056	0.979
Pythia-12B layer 18, seed 49, mult=4 L1=5e-4					
Pythia-12B	0	0.106	[0.064, 0.148]	−0.054	0.992
Pythia-12B	1	0.122	[0.074, 0.171]	−0.038	0.992
Pythia-12B	2	0.129	[0.089, 0.171]	−0.031	0.992
Pythia-12B	3	0.155	[0.112, 0.198]	−0.005	0.992
Pythia-12B	4	0.104	[0.054, 0.153]	−0.057	0.992
Pythia-12B	5	0.084	[0.040, 0.128]	−0.077	0.992
Pythia-12B	6	0.110	[0.064, 0.157]	−0.050	0.992
Pythia-12B	7	0.082	[0.043, 0.120]	−0.078	0.992
Pythia-12B	8	0.042	[−0.004, 0.088]	−0.119	0.992
Pythia-12B	9	0.120	[0.079, 0.161]	−0.040	0.992

Table 25. Per-setting paired bootstrap for the residual-above-reconstruction directional sign test. Original is score_K AUROC on raw activations, Recon is score_K AUROC on the SAE-reconstructed activation, and Residual is score_K AUROC on the residual. Margin is Residual minus Recon. The 95% CI is a paired bootstrap with $n_{\mathrm{boot}} = 10\,000$ resamples and bootstrap seed 0, and CI excl. 0 reports whether the interval excludes zero. Two further Qwen2-7B fine-tune seeds (43 and 44) at multiplier 4 test whether the residual-above-reconstruction ordering is specific to the original fine-tuning seed. The two headline binomial sign tests are computed on the validity-gate-passing settings of Table 1, with the non-Pythia subset of four ($p \approx 0.0625$) and the full set of seven ($p \approx 0.008$). The per-row audit reported here covers nine settings (five inverting cells, the two Pythia anchors, and two Qwen2 secondary seeds) and overlaps with but is not identical to the gate-passing set. All nine settings show positive Residual−Recon margins (one-sided $p \approx 1/2^9 \approx 0.002$). This is reported as a redundant robustness check, not as a substitute for the headline denominators, which remain pre-specified by the validity gate of Table 1.

Setting	Original	Recon	Residual	Margin	95% CI	CI excl. 0
Controlled Pythia experiments						
Pythia-6.9B (mixed), seed 42 anchor	0.803	0.591	0.783	+0.193	[+0.160, +0.227]	yes
Pythia-12B, seed 47	0.764	0.596	0.726	+0.130	[+0.094, +0.166]	yes
Cross-architecture experiments						
Pythia-1B, mult=4, layer 14 (cross-arch diagnostic, not the SAE layer 8 of Table 2), seed 42	0.663	0.592	0.671	+0.079	[+0.060, +0.099]	yes
Pythia-2.8B, mult=4, layer 20, seed 42	0.799	0.683	0.803	+0.120	[+0.098, +0.142]	yes
Qwen2-7B, mult=4, seed 42	0.638	0.553	0.673	+0.120	[+0.086, +0.153]	yes
Qwen2-7B, mult=8, seed 42	0.638	0.552	0.674	+0.122	[+0.090, +0.153]	yes
GPT-Neo-2.7B, mult=8, seed 42	0.616	0.580	0.652	+0.072	[+0.049, +0.095]	yes
Qwen2-7B mult=4 across additional fine-tuning seeds						
Qwen2-7B, mult=4, seed 43 (additional seed)	0.638	0.585	0.658	+0.073	[+0.050, +0.096]	yes
Qwen2-7B, mult=4, seed 44 (additional seed)	0.638	0.578	0.659	+0.081	[+0.056, +0.106]	yes

and 0.507 (vocabulary richness), all at chance. Mean character lengths are 889.6 (member) and 889.1 (non-member), confirming that the controlled split is length-matched by construction. Surface text length cannot account for the membership separability we observe at the activation level.

D.5. Label-Shuffled Permutation Test on the Cosine Alignment

To check that the cosine alignment between the knowledge direction $\mathbf{d}_K$ and the recall direction $\mathbf{d}_R$ is not an artefact of label assignment, we shuffle the (member, non-member) and (high-recall, low-recall) labels independently within the controlled split and recompute the corresponding directions. Across $n = 1000$ random label permutations per model, the observed $|\cos(\mathbf{d}_K, \mathbf{d}_R)|$ is consistent with the permutation distribution on 8 of the 9 models (per-model p-values range from 0.162 to 1.000, no significant deviation at $\alpha = 0.05$). On Mistral-7B the observed cosine falls below the permutation distribution at $p = 0.037$, which supports the near-orthogonality claim. The cosine alignment is therefore not driven by label structure on any model.

D.6. Per-Model SAE Hyperparameter Detail

Dictionary form. We use ReLU plus L1 dictionaries to match the prior SAE-based memorisation work we audit (Wang et al., 2025b; Frikha et al., 2025; Muhamed et al., 2025; Suri et al., 2025). The reconstruction/residual decomposition itself requires only a learned decoder dictionary W_{dec}, but all main interpretations are conditioned on adequate reconstruction fidelity, following the SAEBench emphasis on reconstruction quality as a prerequisite for feature-level interpretation (Karvonen et al., 2025). JumpReLU, BatchTopK, and Matryoshka SAE replications (Rajamanoharan et al., 2024; Bussmann et al., 2024; 2025) are future work.

Table 26 records the SAE settings used for the main decomposition. Hyperparameters are chosen to obtain a reliable reconstruction under the reconstruction-quality criterion in §3.7, not to maximise the reconstruction-induced reduction. For each model we tune the dictionary expansion ratio and the L1 sparsity coefficient, requiring reconstruction cosine at least 0.90 and mean L0 below 60% of dictionary size for the strict tier. The Pythia-6.9B mixed-data SAE uses the five-seed mixed-data evaluation at mult=4 and L1=5e-4.

Pythia-6.9B and Pythia-12B both use mult=4 and L1=5e-4. GPT-Neo-2.7B uses mult=8, clearing the strict reconstruction criterion (cosine 0.998), while Qwen2-7B uses mult=8 at the permissive boundary (cosine 0.850). OPT-6.7B clears at mult=8 with L1=2e-4 and is included only for residual norm decomposition, since its residual carries norm signal independently of direction (Table 23). Mistral-7B does not reconstruct activations reliably enough under the plain-L1 configurations we evaluated. Per-seed configurations are released with the artefact, and wider retuning and unsuccessful attempts are documented in Appendix B.18.

The hyperparameter retuning is per-model. The reconstruction/residual result is therefore conditional on each model's reliable-reconstruction configuration, not a fixed-hyperparameter comparison. Cross-model replication across JumpReLU, BatchTopK, and Matryoshka SAEs remains future work.

D.7. L1-regularised SAE Training Instability

Standard L1-regularised ReLU SAE training is sensitive to random initialisation near the sparsity phase transition (Bricken et al., 2023). Below the transition, training collapses into a dense basin where most features fire on most tokens. Above the transition, training reaches a sparse basin with the target mean L0. The transition L1 depends on activation magnitude statistics that vary across model and layer, so a single global L1 cannot guarantee sparse-basin convergence on every model.

We observe this empirically in supplementary five-seed retraining analyses on the cross-architecture SAEs at L1 $= 5 \times 10^{-4}$. The controlled Pythia experiments at this same L1 train into the sparse basin on every seed (mean L0 below 2 percent of dictionary size across the three configurations, reconstruction cosine above 0.92). On GPT-Neo-2.7B with the same hyperparameter, two of five retrained seeds reach the sparse basin (mean L0 near 700 of 20480 features, reconstruction cosine 0.998). The remaining seeds collapse into the dense basin (mean L0 near 17000 of 20480, reconstruction cosine between 0.16 and 0.60). The two outcomes are not noisy estimates of a single result. They are distinct training basins, consistent with the documented L1 phase transition and with the SAE-quality conventions reviewed by Karvonen et al. (2025).

The validity gate handles this failure mode by construction. Dense-basin training runs fall well below both the 0.85 permissive and the 0.90 strict reconstruction-cosine thresholds and are excluded from quantitative comparison. Newer SAE variants such as TopK (Gao et al., 2024) and BatchTopK (Bussmann et al., 2024) avoid the L1 transition by replacing the sparsity penalty with a hard top-k selection, but this paper deliberately scopes its main decomposition to the standard L1-regularised ReLU configuration used by the membership-related interventions we audit. A bounded TopK scope test is reported in Appendix D.8.

D.8. TopK SAEs Do Not Reproduce the Residual Ordering

This section reports a scope test that bounds the residual-above-reconstruction ordering to L1-regularised ReLU

Table 26. Per-model sparse autoencoder hyperparameters under the reconstruction-quality criterion. Ratio is the dictionary expansion factor $d_{\text{sae}}/d_{\text{model}}$, and L1 is the L1 sparsity coefficient. Status records whether the configuration cleared the strict reconstruction gate (recon cos ≥ 0.90) or appears in a wider sparsity band as cross-architecture diagnostic only. The Pythia-6.9B entry reports the mixed-data five-seed configuration. The Mistral-7B entry reports the L1 range tested across plain-L1 attempts, none of which cleared the strict gate. The elastic-net grid referenced in the limitations is documented in Appendix B.18.

Model	Status	Ratio	L1
Pythia-6.9B	passes threshold	4	5e−4
Pythia-12B	passes threshold (replication)	4	5e−4
Qwen2-7B	passes threshold at mult 8	8	5e−4
GPT-Neo-2.7B	passes threshold at mult 8	8	5e−4
Pythia-1B	wider band (cross-arch only)	4	5e−4
Pythia-2.8B	wider band (cross-arch only)	4	5e−4
Falcon-7B	sparsity boundary (cross-arch)	4	5e−4
OPT-6.7B	residual norm confound	8	2e−4
Mistral-7B	no setting cleared the strict gate	4	1e−4 to 5e−4
Llama-3-8B	failed reconstruction criterion	4	5e−4

SAEs at the hyperparameters we use. A reader may ask whether the ordering is a property of dictionary autoencoders in general or specific to the L1-regularised ReLU configuration we audit. To test this, we trained TopK SAEs (Gao et al., 2024) at the headline Pythia-6.9B layer 16 multiplier-4 setting, with $K \in \{32, 64, 128\}$ and five seeds each, holding every other hyperparameter fixed to the L1-regularised ReLU configuration.

The ordering reverses under TopK. Across all 15 settings the score_K AUROC of the reconstruction exceeds that of the residual, so the residual lies above the reconstruction in none of the 15. The reconstruction-cosine gate passes on all 15, with mean cosine 0.983, so the TopK SAEs reconstruct the activation well. The reconstruction itself retains most of the membership signal, with the mean reduction from original to reconstruction at about 0.04, against about 0.21 for the L1-regularised ReLU SAE at the same model and layer. Under TopK the membership signal therefore stays in the part of the activation that the top-k most active features reconstruct, not in the residual.

This is one alternative dictionary mechanism, not a survey, so it reinforces rather than retracts the headline finding. The residual-above-reconstruction ordering reported in the main results is a property of the L1-regularised ReLU SAEs at our hyperparameters, and the multi-seed and multi-architecture evidence inside that family is unaffected. We have not tested JumpReLU, gated, or BatchTopK dictionaries, so we make no wider architectural claim from this one mechanism.

D.9. Feature Selection Does Not Reach the Residual Membership Signal

This section reports a feature-selection audit on Pythia-6.9B under the L1 SAE, following the privacy-aware feature-selection approach of Frikha et al. (2025). It corroborates the sparse-code adequacy result of Appendix B.16, that the SAE encoding does not preserve the membership signal in the raw hidden state. We test whether selecting and suppressing SAE features removes the residual membership signal.

A logistic detector trained on the full SAE code on Pythia-6.9B under the L1 SAE reaches mean membership AUROC 0.526 (95% CI [0.518, 0.534]), close to chance. The sparse code therefore carries little membership signal on its own, consistent with the sparse-code adequacy test.

We then apply three feature-selection criteria from the privacy-aware approach, top-k magnitude, mean difference, and the combined criterion, and measure their effect on verbatim extraction and on the residual membership detector. On Pythia-6.9B under the L1 SAE, the three criteria reduce verbatim extraction, with the largest extraction reductions reaching about -0.19 for top-k magnitude and about -0.15 for mean difference. The residual membership detector stays inert across the same interventions. Its AUROC moves by at most 0.0218 across the 60 feature-selection settings we evaluate.

The two outcomes separate feature-level extraction control from membership detection on Pythia-6.9B under the L1 SAE. Suppressing the selected SAE features reduces verbatim extraction but leaves the residual membership detector at its original level. The membership signal is therefore not located in the SAE feature dictionary, which matches the sparse-code adequacy finding in Appendix B.16.

D.10. Held-Out Estimation Preserves Ordering but Reduces Magnitude

We test whether the single score_K direction depends on the partition used to estimate it by refitting $\mathbf{d}_K$ on one labelled split and evaluating it on the held-out split (§3.7).

A separate concern is whether the membership direction $\mathbf{d}_K$ is tied to the labelled partition used to fit it. If the direction shifts across partitions, the precise score_K reduction in Table 1 should be read as partition-conditional even if the residual-above-reconstruction ordering still holds.

We test this by splitting the canonical 2000-document labelled pool into a 70% subset for fitting $\mathbf{d}_K$ and a disjoint 30% subset for evaluation. Pythia-6.9B uses layer 16 from the seed 43 mult=4 L1=5e-4 SAE, and Pythia-12B uses

layer 18 from the seed 49 mult=4 L1=5e-4 SAE. For each of ten random splits, we refit $\mathbf{d}_K$ on the fitting subset, evaluate score_K AUROC on the held-out subset, and compute the held-out reconstruction reduction with 10000 bootstrap resamples. The stability threshold compares the held-out reduction against the in-partition reduction minus two cross-split standard deviations.

The magnitude claim does not pass this stability analysis. On Pythia-6.9B, the in-partition mean reduction is 0.213 with cross-split standard deviation 0.012, while the mean held-out reduction is 0.149. On Pythia-12B, the in-partition mean reduction is 0.160 with cross-split standard deviation 0.012, while the mean held-out reduction is 0.105. In both cases, the held-out reduction falls below the corresponding threshold. The qualitative residual-above-reconstruction ordering nevertheless survives held-out estimation on both Pythia models. This probe therefore narrows the quantitative claim without overturning the qualitative decomposition result. Per-split numerical detail is in Table 24.

D.11. Pythia-12B Replication Detail

Pythia-12B asks whether the controlled Pythia-6.9B finding recurs at larger scale within the same model family. We train five independent mixed-data SAEs on Pythia-12B layer 18 at mult=4 and L1=5e-4. The five SAE seeds are 47 through 51.

All five retained SAEs clear the reconstruction criterion, with reconstruction cosine in $[0.991, 0.992]$. The reconstruction reductions are 0.169, 0.139, 0.160, 0.131, and 0.161, giving mean 0.152 and standard deviation 0.016. Each reduction is above the 0.10 strong-effect threshold. The residual AUROC sits below the original on all five evaluations, but it is more stable across seeds than the dictionary projection. Residual AUROC standard deviation is 0.005, while dictionary-projection AUROC standard deviation is 0.016. Since the original AUROC is fixed at 0.764, the reduction variance is the dictionary-projection variance by construction.

This replication varies SAE initialisation while holding the fine-tuned Pythia-12B model fixed. It therefore supports a within-family scale replication of the residual-above-reconstruction pattern, but it does not measure fine-tuning-seed variance at 12B and does not establish cross-architecture controlled replication.

D.12. Gemma-2-2B Preserves the Qualitative Reconstruction Ordering

Gemma-2-2B adds a Google-family model to the qualitative cross-architecture set. We train one mixed-data SAE at layer 16 with multiplier 4 ($d_{\text{sae}} = 9{,}216$) and L1 = 5e-4 on the union of 1000 member and 1000 non-member OpenWebText activations, then evaluate the same reconstruction/residual decomposition on the canonical evaluation pool. The run uses one fine-tuning seed and one SAE seed.

The reconstruction cosine is 0.911, and the mean number of active features is 1408.6 out of 9216. Original score_K AUROC is 0.801, SAE-reconstruction AUROC is 0.769, and residual AUROC is 0.820. The reconstruction reduction is only 0.032, below the 0.10 strong-effect threshold used for the controlled Pythia claim, but the qualitative ordering residual $>$ original $>$ reconstruction holds. The residual norm AUROC is 0.612, below the 0.65 norm-confound flag. Gemma also satisfies the residual-above-reconstruction ordering at this single seed, consistent with the binomial sign test reported in the main results section.

Gemma-2-2B therefore contributes to qualitative generality, not to the within-model strong-reduction evidence. The model is smaller than the controlled Pythia experiments, and neither fine-tuning-seed variance nor variance across SAE seeds is measured here.

Sparsity scope of the geometric and feature-alignment analyses. The full channel-decomposition, FSC, norm-direction, and layer-trajectory analyses were extended to Gemma-2-2B in a later run. The Gemma SAE used for these analyses operates at mean L_0 of 1408 of 9216 features (15.3 percent), well above the comparable range across the other models in Table 1 (Pythia-12B at 0.05 percent, Qwen2-7B with multiplier 8 at 0.01 percent, Pythia-1B at 1.29 percent, Pythia-6.9B at 0.07 percent, and GPT-Neo-2.7B with multiplier 8 at 3.77 percent). The reconstruction cosine remains within the validity gate at 0.911, so the residual signal in Table 1 is comparable in that gate's terms. The disparity arises because the same L1 coefficient (5×10^{-4}) does not equally constrain activation magnitudes across model families, and Gemma's activations require a stronger penalty to reach the standard sparsity range.

We group the Gemma full-analyses into four findings by their dependence on SAE feature sparsity. The residual-above-reconstruction ordering (Table 1) depends only on reconstruction cosine, which Gemma satisfies, so this finding stands without sparsity scoping. The channel-decomposition minimum principal angle of 38.33 degrees, against the typical 75 degrees and above on the other models, refers to the cross-model alignment between $\mathbf{S}_K$ subspaces of different models defined in Eq. 4 (a coordinate-level subspace-overlap metric distinct from the within-model $\mathbf{S}_K \leftrightarrow \mathbf{S}_R$ mean principal angle reported in the per-layer tables) and is computed from raw activations through contrastive principal component analysis without any SAE involvement, so this finding is also independent of the sparsity disparity. The norm-direction membership AUROC of

0.600 at the best layer is a pure activation-magnitude analysis with no SAE dependence and is also independent of sparsity. The feature-alignment asymmetry $\mathrm{FSC}_K^{\mathrm{cf}}$ approximately equal to $\mathrm{FSC}_R^{\mathrm{cf}}$ (0.414 against 0.425) on Gemma, however, depends directly on which SAE features are classified as discriminative and how selective those features are. At a near-dense L_0 of 15.3 percent, feature selectivity is materially weaker than at below 4 percent L_0, and the asymmetry observed on the other settings requires sparse selective features to manifest. We therefore restrict the analyses in Tables 17, 18, and 16 and Figure 4 to the SAE settings within the comparable sparsity range. Gemma is retained in Table 1 where the validity gate is satisfied. The channel-decomposition and norm-direction measurements for Gemma are reported in this appendix for completeness, and are not integrated into the main geometric tables. A stronger-L1 Gemma retraining evaluation is outside the submitted scope.

D.13. Standard Probe Replication and Output-Layer Scope

The main text scopes the reconstruction/residual result to a layer-local measurement. The following analysis explains this scope restriction. A natural alternative reading is that the effect reflects the score_K probe rather than the internal representation. To test this, we repeat the SAE patching decomposition on Pythia-6.9B fine-tuned for 5 epochs with three standard membership inference probes. The probes are the loss attack, MIN-K%=20, and zlib, following the canonical implementations referenced in Yeom et al. (2018), Shi et al. (2023), and Carlini et al. (2021).

For each probe we compare original AUROC on raw activations, AUROC after SAE reconstruction, and AUROC on the residual. The replication criterion requires the reconstruction AUROC to fall by at least 0.05 on at least two of the three probes. The replication criterion is not met. The original AUROC saturates on all three probes, and the SAE-patched reconstruction AUROC stays within rounding of the original. This result suggests that downstream layers and the LM head compensate for the SAE-patched activation. The reconstruction reduction is therefore a property of the layer-local internal measurement, not an output-level improvement to standard membership attacks.

Direction estimation and bootstrap conventions. The membership direction $\mathbf{d}_K$ used for score_K is fitted by five-fold cross-validation on the OpenWebText evaluation partition, with the projection then computed on held-out folds. All AUROC and ΔAUROC confidence intervals reported in the body and appendix tables use $n_{\mathrm{boot}} = 10{,}000$ paired bootstrap resamples, with bootstrap pairs at the document level so that the same labelled documents enter the resample on the original and the comparison condition. Standard deviation across SAE seeds is reported separately from the bootstrap interval where independently trained SAEs are compared, since seed variance and within-fit variance answer different questions.

D.14. Detailed Limitations and Boundary Conditions

The main text states the claim boundaries compactly. This entry expands them so that each limitation can be read against the claim it bounds.

Held-out partition sensitivity of the magnitude claim. Held-out estimation of $\mathbf{d}_K$ supports the qualitative residual-above-reconstruction ordering but not the original magnitude claim. Refitting $\mathbf{d}_K$ on a 70 percent partition and evaluating on a disjoint 30 percent partition produces a smaller mean reconstruction-induced reduction than the in-partition value (Pythia-6.9B in-partition 0.213 against held-out 0.149, and Pythia-12B in-partition 0.160 against held-out 0.105). The qualitative residual $\geq$ original $>$ reconstruction ordering nevertheless survives held-out estimation on both Pythia models. Held-out estimation therefore weakens the magnitude claim while preserving the qualitative ordering. $\mathbf{d}_K$ as a single direction is sensitive to the partition used for fitting, and the qualitative ordering is not driven by evaluating and fitting that direction on the same documents.

SAE variant scope. The SAE conclusion is limited to standard L1-regularised ReLU SAEs as evaluated here. Newer variants such as JumpReLU, BatchTopK, and Matryoshka SAEs alter the reconstruction and sparsity tradeoff, so they should be audited before generalising the result to all sparse dictionary parameterisations (Rajamanoharan et al., 2024; Bussmann et al., 2024; 2025). A TopK scope test at the headline setting is reported in Appendix D.8, where the ordering reverses.

Practical implication. The extraction-detection separation and the paraphrase-orientation diagnostic together support a single practical conclusion. SAE feature-level editing should not be treated as a privacy guarantee unless the target model is audited for membership signal recoverable from the reconstruction residual.

The evidence has two sources. The controlled Pythia experiments carry the primary causal evidence, because both the member-only and mixed-data dictionaries reproduce the same residual-above-reconstruction ordering at reconstruction quality that passes the validity gate. The cross-architecture experiments across non-Pythia models provide corroborating evidence where reconstruction quality is sufficient, and are reported as a diagnostic where it is not. These experiments support qualitative generality of the residual-above-reconstruction ordering, not a universal architecture-level law. The norm-direction distinction across architecture

families is descriptive. The residual principal component ablation suggests partial low-dimensional structure, yet it does not satisfy the stronger falsification criteria needed for causal localisation. These findings concern layer-local membership recoverability under controlled fine-tuning with IID member and non-member splits, not individual-sentence pretraining MIA, which recent work has shown to be confounded by distribution shift (Maini et al., 2024; Duan et al., 2024). The fine-tuning regime amplifies per-sample membership signal beyond what is observable at pretraining scale.

The strongest causal evidence is within Pythia. The matched-hyperparameter mixed-data SAE control succeeds on Pythia-6.9B, and Pythia-12B provides a within-family replication at smaller magnitude. Replication attempts on GPT-Neo-2.7B across five L1 settings and Mistral-7B across two L1 settings did not reconstruct activations reliably enough for controlled decomposition evidence (Appendix B.18). The broader cross-architecture pattern supports qualitative generality, but it is not the same kind of controlled replication as the Pythia experiments.

All fine-tuned models share the same fine-tuning protocol and the same OpenWebText member partition. This design controls the document unit and distribution matching, but it also means that any artefact of the shared fine-tuning setup could affect both the controlled and descriptive results. The Pythia-6.9B corpus-disjointness check argues against SAE train-test corpus overlap on that model (§3.7), but cross-corpus replication for the remaining models is future work.

Several model-specific outcomes define the scope of the cross-architecture claim. Llama-3-8B did not reconstruct activations reliably enough under the attempted configuration. The Qwen2-7B mixed-data evaluation fell below the strict reconstruction-cosine threshold at multiplier 4, while the multiplier-8 retune clears the threshold and enters the cross-architecture set. A Mistral-7B elastic-net SAE evaluation across three λ_{L_2} settings did not yield a reliable reconstruction, so Mistral is excluded from the cross-model set. The family pattern reported descriptively in §4.5 therefore comprises Pythia-6.9B, Pythia-12B, GPT-Neo-2.7B, and Qwen2-7B.

The single score_K direction is partition-sensitive. The held-out estimation probe (§4.6) shows that the magnitude of the score_K reduction is smaller when $\mathbf{d}_K$ is fit and evaluated on disjoint label partitions. The qualitative residual-above-reconstruction ordering is preserved on held-out partitions at both Pythia models, so the residual-recoverability finding is independent of the partition fit at the qualitative level. The precise reduction magnitudes are not.

The privacy-aware SAE experiment covers two models, and post-unlearning persistence is not tested. The limitations should be read together. A reviewer evaluating any one boundary should read it together with the others, since the claim depends on the access setting, corpus, SAE family, and evidence tier.

D.15. Additional Controls

The controls paragraph in §4.6 groups several control analyses by the alternative explanations they address. The following paragraphs give the method and outcome for each control.

The before fine-tuning negative control asks whether the membership direction exists before fine-tuning. Rerunning the channel detector on the not fine-tuned base Pythia-6.9B model across the same five layers leaves the detector at chance, so the membership signal requires fine-tuning. The layer sensitivity test asks whether layer 16 was chosen because it was unusually favourable. Rerunning the membership detector on layers 12, 14, 16, 18, and 20 shows that AUROC increases monotonically with depth, while $\mathbf{d}_K$ and $\mathbf{d}_R$ remain near-orthogonal across the layers tested.

The random-direction baseline asks whether arbitrary residual-stream directions can separate members from non-members. For each model, we draw 100 random unit directions and compute membership AUROC. The fixed-orientation random-direction mean sits near chance on every model, and the 99th percentile remains below the fitted membership direction. The random-init SAE control asks whether the result follows from any reconstruction-shaped operator. Replacing the trained SAE on Pythia-6.9B layer 16 with a randomly initialised SAE collapses reconstruction cosine to near zero and reduces reconstructed score_K AUROC to chance. The trained SAE reaches near-unity reconstruction cosine yet still reduces reconstructed AUROC, indicating that the finding requires a learned high-fidelity dictionary.

The GPT-Neo-2.7B multiplier-8 evaluation and Qwen2-7B multiplier-8 retune test whether denser dictionaries change the qualitative ordering. GPT-Neo-2.7B reconstructs almost perfectly and reproduces the inversion in which residual exceeds original at lower magnitude, with $\mathbf{d}_K$ coverage cosine near zero. Qwen2-7B clears the strict reconstruction-cosine threshold at multiplier 8. Its reduction magnitude increases relative to multiplier 4 but remains below the strong-effect threshold, while $\mathbf{d}_K$ coverage cosine moves further from zero. The Qwen2 retune is therefore reported as an SAE-quality improvement for Qwen2, not as a non-Pythia strong-reduction result. The random-init SAE control remains Pythia-6.9B only.

The Pythia-12B $K = 10$ ablation in Table 27 uses the member-only SAE for continuity with the Pythia-6.9B configuration. We also rerun the same ablation on the Pythia-

12B mixed-data SAE used for the reconstruction/residual evaluation, using seed 47 and leaving the remaining settings unchanged.

The mixed-data rerun reduces the ablation effect from $+0.176$ in the main-text estimate to $+0.018$ (95% CI $[0.015, 0.022]$). The matched Gaussian-noise control produces a larger reduction than the targeted subspace, and the random 5% column-mask control reproduces the ablated AUROC. Reconstruction cosine is 0.991. The per-text ratio of residual norm to original norm is 0.308, a diagnostic outside the reconstruction and sparsity criteria specified in §3.7.

This outcome is why the subspace ablation is illustrative, not the reference point. It is consistent with residual membership signal being partially concentrated, but it does not pass the same-SAE and random-mask control analyses needed for a strong causal localisation claim. The Pythia-12B mixed-data reconstruction/residual decomposition itself is unaffected, since that evaluation always used the mixed-data SAE evaluation and reports a mean reconstruction reduction of 0.152.

Table 27. Top-K residual principal component ablation on the four models that pass the reconstruction-quality criterion. $K = 10$ is the primary setting, and $K = 5$ is included as a sparser-support control. AUROC reduction is the paired difference between the unablated and ablated probe AUROC, with bootstrap 95% CI from $n_{\text{boot}} = 10\,000$ resamples. The Decisive gate column reports whether the primary intervention beats both random-rotation and matched-Gaussian noise controls at the 95% bootstrap margin. The decisive gate is not met on any of the eight model and rank combinations, with Pythia-12B at $K = 10$ producing the largest subspace-level reduction.

Model	K	AUROC reduction	95% CI	Decisive gate
Pythia-12B	10	+0.176	[+0.165, +0.187]	not met
Pythia-6.9B	10	+0.025	[+0.021, +0.029]	not met
Qwen2-7B	10	+0.003	[+0.001, +0.005]	not met
GPT-Neo-2.7B	10	−0.001	[−0.004, +0.002]	not met
Pythia-12B	5	+0.103	[+0.094, +0.112]	not met
Pythia-6.9B	5	+0.013	—	not met
Qwen2-7B	5	+0.002	—	not met
GPT-Neo-2.7B	5	−0.001	—	not met

A same-SAE rerun on the Pythia-12B mixed-data SAE reduces the $K = 10$ estimate to $+0.018$. See Appendix D.15.

D.16. Corpus-Disjoint Dictionary Control

The body controls paragraph uses this check to address train-test SAE corpus overlap on Pythia-6.9B. The standard mixed-data SAE training pool overlaps with the canonical evaluation pool by construction, so the reconstruction reduction could in principle reflect the SAE having seen evaluation documents during dictionary training.

To test that explanation, we retrain the Pythia-6.9B mixed-data SAE at the analysis layer on an OpenWebText partition selected to have zero document overlap with the reconstruction/residual evaluation pool. The multiplier, L1 coefficient, corpus-cycling pipeline, fine-tuned model, and evaluation pool are held constant relative to the standard mixed-data run.

The disjoint-corpus dictionary clears the strict reconstruction criterion and yields a reconstruction reduction of 0.278 in the favourable direction relative to the in-corpus baseline of 0.209 mean across the five mixed-data seeds. The qualitative residual-above-reconstruction ordering is preserved. This argues against train-test SAE corpus contamination as the explanation on Pythia-6.9B. Cross-architecture replication of corpus-disjoint dictionary training remains future work.

D.17. TPR at Low FPR and Paraphrase Diagnostic

A bidirectional audit that picks the orientation maximising AUROC on each evaluation condition is no longer a robustness test, because it lets the detector refit at evaluation time. A bidirectional report would conceal the orientation flip and inflate apparent paraphrase robustness, which is why the finding in §4.6 requires fixed-orientation AUROC.

Table 28 reports bootstrap-resampled true positive rate (TPR) at the 0.1% FPR cutoff for the residual $\mathbf{d}_K$ membership channel across four models. The 0.1% FPR cutoff is the operating point used by recent fine-tuning MIA work (Chen et al., 2026; Ilić et al., 2026). We use $n_{\text{boot}} = 10{,}000$ bootstrap resamples drawn with replacement from the per-text score level, with seed 12345 and $n_{\text{member}} = n_{\text{nonmember}} = 1000$ per model.

Table 28. True positive rate (TPR) at 0.1% false positive rate (FPR) for the residual $\mathbf{d}_K$ membership detector across four models. TPR (point) is the point estimate, Bootstrap mean is the resampled mean across $n_{\text{boot}} = 10\,000$ paired bootstrap resamples, and 95% CI is the bootstrap interval. Pythia-12B has the highest point estimate at low FPR, with overlapping bootstrap CIs across the three non-Qwen2 models. Qwen2-7B is not statistically distinguishable from chance at this sample size, consistent with Qwen2's weak architectural replication of the dark-subspace effect documented in Appendix B.18.

Model	TPR (point)	Bootstrap mean	95% CI
Pythia-6.9B	0.015	0.018	[0.005, 0.050]
Pythia-12B	0.032	0.035	[0.020, 0.071]
GPT-Neo-2.7B	0.025	0.025	[0.014, 0.042]
Qwen2-7B	0.010	0.007	[0.000, 0.023]

Three of the four models (Pythia-6.9B, Pythia-12B, GPT-Neo-2.7B) have CI lower bounds excluding the chance baseline of 0.001 at FPR = 0.001. Qwen2-7B's CI lower bound at 0.000 means its TPR at 0.1% FPR is not statistically distinguishable from chance at this sample size. The bootstrap captures sampling variance over the 1000 plus 1000 evaluation tuples but not SAE training seed variance.

The controls paragraph routes two privacy-audit details here. The first is whether the residual channel adds signal at low

false-positive rates, where false accusations of training-data use are costly. The second is whether bidirectional AUROC reporting can obscure detector-direction sensitivity under paraphrase.

For the low-FPR audit, we measure the standard loss attack, zlib, and MIN-K% baselines plus the residual $\mathbf{d}_K$ channel at the 1% FPR and 0.1% FPR thresholds on Pythia-6.9B, Qwen2-7B, GPT-Neo-2.7B, and Pythia-12B at epoch 5. Controlled fine-tuning saturates per-example loss on three of the four models. On Pythia-6.9B, Qwen2-7B, and Pythia-12B, the standard loss, zlib, and Min-K-style baselines all reach the TPR ceiling at 1% FPR. GPT-Neo-2.7B is the one model where loss does not saturate, and there the residual $\mathbf{d}_K$ channel exceeds the best standard baseline with paired bootstrap CI above zero. At the stricter 0.1% FPR threshold preferred for privacy-auditing evaluations (Chen et al., 2026; Ilić et al., 2026), the residual $\mathbf{d}_K$ channel TPR is statistically distinguishable from chance on three of four models (Pythia-6.9B, Pythia-12B, GPT-Neo-2.7B), with Qwen2-7B not distinguishable from chance.

For the paraphrase diagnostic, we evaluate the SAE residual membership detector on paraphrases with shuffled word order of member documents. We replicate the diagnostic on Pythia-6.9B seed 42, Qwen2-7B, and Pythia-12B. Bidirectional AUROC reporting gives a lift on paraphrases relative to originals, but the fixed-orientation audit shows that the detector direction flips between conditions. Under the orientation learned on originals, paraphrased member scores shift in the opposite direction relative to non-members. The paraphrased member and non-member distributions remain linearly separable, but in the reversed direction. We therefore treat the paraphrase outcome as a methodological warning about bidirectional AUROC reporting, not as evidence that paraphrasing increases memorisation.

Bidirectional AUROC and the reconstruction-residual comparison. The reconstruction-residual comparison in Table 1 was recomputed under the same bidirectional AUROC convention. Because that convention reports the larger of the two possible sign conventions on each view, a reader could reasonably ask whether the residual column scores high because the AUROC computation is allowed to flip the detector direction view by view. We checked this directly on every setting with the per-text scores. The fitted membership direction separates members from non-members in the same way on the original activation, on the reconstruction, and on the residual. Members score higher than non-members on all three views, and no setting ever requires a sign swap. The reported values therefore match what a fixed-direction audit would have produced, so the ordering in which residual AUROC exceeds reconstruction AUROC is a property of the activations themselves, not an artefact of giving the AUROC computation freedom to flip its sign. This is the opposite pattern from the paraphrase case above, where the orientation genuinely reverses and the bidirectional helper would hide that reversal.

Code Availability

Code, data, and analysis scripts are available at https://github.com/NiloyPurkait/DarkSubspace. The repository structure and per-claim artefact mapping are documented in the artefact README at the repository root, and the experimental protocol is documented in the accompanying protocol-disclosure document.