

The Geometry of Ignorance: Detecting Suppressed Knowledge in LLMs through Activation Fragmentation

Adar Peleg*, Dvir Alsheich*, Tomer Ashuach, Avi Mendelson

¹Technion - Israel Institute of Technology

*Equal contribution

Correspondence: {a.p, dviralcheich, tomerashuach}@campus.technion.ac.il, mendlson@technion.ac.il

Abstract

Evaluating whether a large language model (LLM) genuinely knows a concept remains a fundamental problem, since behavioral metrics can be fooled by suppression: a model that refuses to answer may still harbor recoverable knowledge. The framework measures the geometric coherence of *anon-vectors*, the differences in hidden states between a concept-specific query and its anonymized paraphrase, and shows that genuine knowledge forms a tight, low-entropy cluster while ignorance forces geometric fragmentation. Across ten instruction-tuned LLMs from three model families, all fragmentation metrics reliably separate known from unknown concepts, and post-unlearning, fragmentation geometry correlates with adversarial recovery across four attack types and five unlearning algorithms, though the direction of that relationship is shown to depend on attack methodology. Furthermore, a layer-resolved analysis reveals that different adversarial mechanisms exploit this geometry at varying depths, with high attack-specific correlations that are invisible to behavioral evaluation. Together, these results constitute a single-phase, attack-free proxy for unlearning robustness that captures aspects of parametric knowledge inaccessible to output-level metrics.

A reference implementation is available at 

1 Introduction

The expansive pre-training of Large Language Models (LLMs) inevitably leads to the memorisation of vast amounts of factual associations (Petroni et al., 2019; Meng et al., 2023), raising critical legal, ethical, and safety concerns regarding privacy and the generation of harmful content (Carlini et al., 2021; Seyitoğlu et al., 2024). Traditionally, the boundaries of a model’s parametric knowledge have been evaluated through behavioral analysis, relying on surface-level metrics such as question-answering accuracy or token perplexity (Lynch

et al., 2024). Behavioral evaluations treat the model as a black box and fail to capture how factual associations are internally stored and retrieved during the forward pass (Geva et al., 2023). This gap is most consequential in the context of machine unlearning, which aims to selectively erase specific concepts without retraining (Li et al., 2024; Gandikota et al., 2025; Ashuach et al., 2026). Recent representation-level analyses reveal that unlearning interventions often result in the mere suppression or bypassing of knowledge, rather than genuine deletion (Xu et al., 2025; Lee et al., 2024; Schoepf et al., 2026). Because suppression mechanisms leave the knowledge structurally distinguishable in the parameters (Huu-Tien et al., 2025), it can be reliably recovered by malicious actors through activation steering (Seyitoğlu et al., 2024). Consequently, robust evaluation must shift to fine-grained, intrinsic metrics that can detect suppressed but recoverable knowledge.

Recent mechanistic interpretability work has introduced intrinsic probes that bypass generation and measure knowledge directly within intermediate model layers (Gottesman and Geva, 2024; Hong et al.). However, these metrics remain entangled with prompt syntax, conflating factual knowledge with the linguistic and syntactic structure of the query. Moreover, no single-phase metric reliably predicts whether suppressed knowledge is adversarially recoverable.

To address the limitations of both behavioral evaluations and syntax-entangled intrinsic metrics, we build upon the concept of the *anon-vector* (Seyitoğlu et al., 2024), defined as the directional shift between the hidden state of a concept-specific query and its identity-stripped paraphrase. We use this to introduce *activation fragmentation*, a geometric framework that isolates the precise footprint of parametric knowledge using two forward passes per query (no token generation, no iterative optimization).

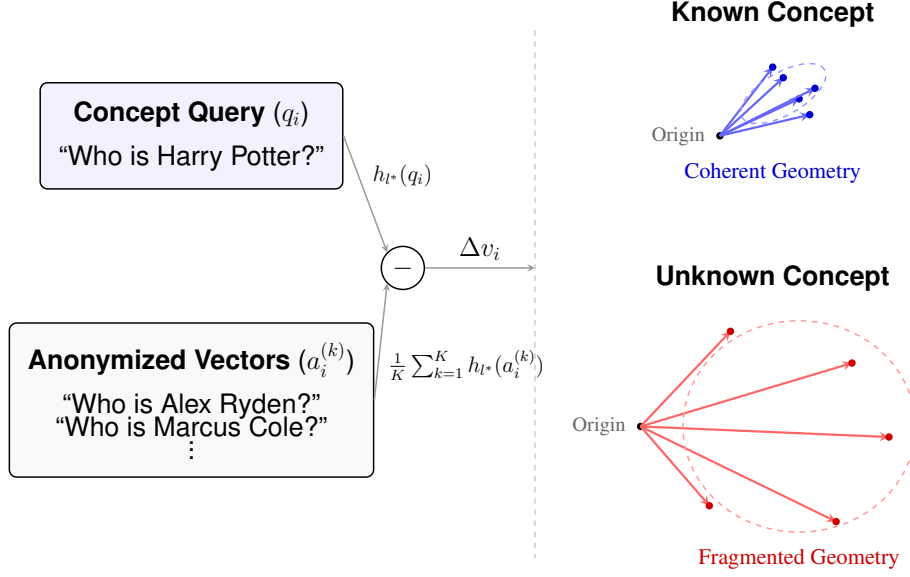


Figure 1: Activation fragmentation isolates conceptual knowledge. Subtracting the mean hidden state of a cloud of anonymized variants cancels shared linguistic syntax. Plotting the resulting anon-difference vectors (Δv_i) from the origin reveals that parametric knowledge manifests as tight, low-entropy covariance clusters of aligned vectors, whereas ignorance forces geometric fragmentation across high-variance directions.

We hypothesize that unknown concepts lack parametric structure: their anon-difference vectors scatter into high-variance, fragmented directions. By contrast, genuine parametric knowledge manifests as a tight, cohesive geometric signature in activation space.

Contributions. Our main contributions are as follows:

- We propose *activation fragmentation*, a fine-grained, representation-level metric that leverages anon-vectors to successfully isolate conceptual identity from syntax in intermediate model layers.
- We demonstrate empirically that fragmentation metrics distinguish known entities (e.g., Harry Potter) from completely unknown fictitious concepts (e.g., TOFU (Maini et al., 2024)) with robust statistical separability across ten instruction-tuned LLMs from three distinct model families.
- *Activation Uniformity Scatter Score (AUSS)*. We establish that post-unlearning fragmentation serves as a highly accurate, attack-free predictor of elements of unlearning robustness. (e.g. AUSS dispersion metrics predict in-context learning (ICL) recoverability).

2 Background

Knowledge in LLM activations. LLMs encode factual associations in their parameters, with feed-forward layers functioning as key-value memories that promote specific concepts into the residual stream (Geva et al., 2021, 2022). Intermediate layers carry geometrically richer knowledge representations than the final layer (Skean et al., 2025; Razzhigaev et al., 2023). Skean et al. (2025) measure representational compression layer-by-layer via the Rényi-2 Gram-matrix entropy: for any matrix $Z \in \mathbb{R}^{M \times d}$,

$$K_Z = \frac{ZZ^\top}{\text{tr}(ZZ^\top)} \quad S_2(Z) = -\log_2 \text{tr}(K_Z^2) \quad (1)$$

Lower S_2 indicates that the M vectors concentrate in a smaller subspace (higher representational compression), identifying where factual content is most geometrically concentrated in the hidden space. Applied to the raw query hidden states $\mathbf{H} = \{h_{l*}(q_i)\}$, this gives the baseline metric **Reg S_2** (Skean et al., 2025): $\text{Reg}_{S_2} = S_2(\mathbf{H})$.

Anon-vectors and concept-differential representations. A concept-specific signal can be isolated by contrasting the hidden state of a real query with that of a semantically equivalent but identity-stripped (anonymised) paraphrase. Seyitoğlu et al. (2024) exploit this to mount an *activation steering attack*: they compute a steering vector as the

raw difference between hidden states of a concept query and its anonymised counterpart, then inject it into the residual stream to re-activate suppressed knowledge in an already-unlearned model. The ConceptVectors dataset (Hong et al.) likewise demonstrates that named entities encoded in model parameters leave verifiable parametric traces at intermediate layers, and that jailbreak attacks ride along these very directions to bypass unlearning (Hong et al.). Our anon-difference vector Δv_i (Eq. 2) builds directly on the anonymisation technique of Seyitoğlu et al. (2024), averaging over K variants to stabilise the concept-differential signal. Neither prior work, however, treats the *population geometry* of a batch of such vectors as a diagnostic metric; that is a contribution of this paper.

3 Method

3.1 Metric Definitions

Anon-vectors and Entropy. Let $\{q_i\}_{i=1}^N$ be N queries about a target concept, and let $h_\ell(\cdot) \in \mathbb{R}^d$ be the mean-pooled hidden states at layer ℓ . For each q_i , we construct K anonymised variants $\{a_i^{(k)}\}_{k=1}^K$. The **anon-difference vector** at layer ℓ isolates conceptual identity by subtracting the mean anonymised state:

$$\Delta v_i(\ell) = h_\ell(q_i) - \frac{1}{K} \sum_{k=1}^K h_\ell(a_i^{(k)}) \quad (2)$$

This requires two forward passes per query in performed experiments (in general $K+1$ per query, with $K=1$ in experiment sections) and no token generation. All fragmentation metrics are computed from the matrix of anon-difference vectors at a chosen layer; how that layer is selected is stated in each experimental protocol (Section 4) and expended in (Appendix B.3). To measure representational rank concentration, we apply Gram-matrix entropy to these unnormalised vectors, defining our primary dispersion metric: $\text{AnonDiff}_{S_2} = S_2(\Delta \mathbf{V}(\ell))$. Applying S_2 to concept-differential vectors rather than raw activations is a novel contribution of this work. We also evaluate two variants: $\text{Reg}_{S_2} = S_2(\mathbf{H}^\ell)$ applies the same entropy to the raw hidden-state matrix (no anonymisation subtraction), serving as a baseline; and $\text{NormAnonDiff}_{S_2} = S_2(\hat{\Delta} \mathbf{V}(\ell))$ applies it to the L_2 -normalised anon-difference vectors.

AUSS Batch-mean metrics (Ours). To assess population-level geometric consistency, we parti-

tion the N pairs into $B = 10$ disjoint batches (Appendix B.3). For each batch b , we compute the unit-normalised mean anon-difference direction $\hat{d}_b$. We then measure geometric consistency across these B representative directions using two primary novel matrices that aim to measure *fragmentation*:

- **AUSS_{L2}**: Mean pairwise squared L_2 distance between batch-mean directions.
- **AUSS_{cos2}**: Complement of mean pairwise squared cosine between batch-mean directions.
- **AUSS_{Jac}**: Jaccard distance between the sets of the $k=32$ largest-magnitude coordinates of each pair of batch-mean directions, the fraction of top-activated dimensions that differ between batches.

Note that all three metrics are bounded: $\text{AUSS}_{L_2} \in [0, 2]$ and $\text{AUSS}_{\cos^2}, \text{AUSS}_{Jac} \in [0, 1]$. We expect that when a concept is known, batch-mean directions cluster tightly, driving all three metrics toward zero; when unknown, they are expected to scatter, driving them toward the upper bound. This bounded range means that Cohen’s $|d|$ can appear large even with modest raw gaps, because the pooled within-group standard deviation is bounded from above. A comprehensive list of additionally evaluated metric variations including base pairwise similarities, Jaccard incoherence, and covariance traces is detailed in Appendix C.

3.2 Entropy, Rank, and the Geometry of Parametric Knowledge

The central hypothesis of this work is that parametric knowledge leaves a geometric imprint. If a model *knows* a concept, diverse queries converge onto a shared subspace, producing a tight, low-entropy cluster of anon-difference vectors. Conversely, *ignorance* drives these vectors to scatter across the activation space, increasing entropy.

This connection between the Gram-matrix entropy S_2 and geometric dispersion is formalised by the following result (for proofs see Appendix A):

Theorem 1 (High entropy forces geometric fragmentation). *Let $A \in \mathbb{R}^{n \times d}$ be a matrix of L_2 -normalised activation vectors and let $G = AA^\top$ be the corresponding Gram matrix¹ with effective*

¹For a matrix A of unit-normalised vectors, the Gram matrix $G=AA^\top$ collects every pairwise cosine similarity ($G_{ij}=\cos(a_i, a_j)$), so its eigenvalue spectrum summarises how many effective directions the vectors collectively occupy.

rank $\text{EffRank}(G) = \text{Tr}(G)/\lambda_{\max}(G)$. Then the mean pairwise squared Euclidean distance satisfies

$$D_{L_2}^2 \geq 2 - \frac{2}{\text{EffRank}(G)}.$$

Since $2^{S_2/2} \leq \text{EffRank}(G) \leq 2^{S_2}$ (Proposition 1, Appendix A), high S_2 forces the effective rank upward, which structurally guarantees spatial dispersion. The complementary direction holds under a natural assumption: when all pairwise cosine similarities are non-negative (i.e., vectors share a common half-space). For anon-vectors this is geometrically motivated: if a model parametrically encodes a concept, every query about that concept shifts the hidden state *toward the same concept direction*, so the differences Δv_i collectively point away from the anonymised baseline in a consistent orientation. Under this assumption, a matching upper bound applies:

Theorem 2 (Low entropy forces geometric clustering). *Under the same setup as Theorem 1, if additionally $G_{ij} \geq 0$ for all $i \neq j$, then*

$$D_{L_2}^2 \leq \frac{2n}{n-1} \left(1 - 2^{-S_2(G)}\right).$$

In particular, $D_{L_2}^2 \rightarrow 0$ as $S_2(G) \rightarrow 0$.

Together, these bounds demonstrate that subspace-compression metrics directly reflect the presence or absence of a coherent concept representation, effectively isolating knowledge from mere behavioural suppression.

3.3 Semantic Testing

Anon-vectors over raw activations. Raw activations encode both conceptual content and prompt syntax. Subtracting the anonymised baseline (denoted as Δv_i) cancels shared syntactic components, isolating the conceptual signal. This produces a substantially stronger discriminative signal than working with raw activations directly (Table 1).

Semantic vs. generation artifacts. Behavioral evaluations measure accessible outputs, not latent storage. Since representational suppression mechanisms can block factual recall without deleting the underlying knowledge (Lee et al., 2024; Xu et al., 2025), evaluating hidden-state geometry provides a strictly more rigorous proxy for true parametric erasure. A model that refuses to answer, whether through RLHF-driven refusal or explicit unlearning, may still harbor recoverable knowledge in its

hidden states; generation-based metrics will report success while the knowledge persists, making them insufficient as a sole evaluation criteria.

The measurement problem: no true oracle.

The ideal ground truth for a knowledge-detection metric would be a model trained from scratch on a corpus that provably excludes the target concept, making its ignorance logically certain. In practice, this is prohibitively expensive and inaccessible: controlled pre-training runs require substantial compute, carefully curated multi-billion-token datasets, and institutional resources. There is, therefore, no true oracle for parametric knowledge. We cannot directly observe what a model knows, only what it outputs, and output-level proxies risk circularity when used to calibrate the very metrics designed to go beyond them.

Bogus data as a proxy for not knowing.

We sidestep this measurement problem by grounding knowledge in dataset construction rather than model training. As the negative class, we draw from data. For example, the TOFU benchmark (Maini et al., 2024) consists entirely of fictitious authors invented for the benchmark and absent from any pre-training corpus. Because these entities were never written about anywhere, no model can have parametric knowledge of them, and there is no plausible mechanism by which such knowledge could exist. As the positive class, we can draw from Harry Potter (HP) QA pairs; HP knowledge is confirmed to persist in the parameters of most LLMs as a consequence of large-scale pre-training (Carlini et al., 2021; Seyitoğlu et al., 2024). A metric that separates HP from TOFU under this protocol has therefore demonstrated genuine knowledge detection, not behavioral correlation. The specific datasets and model counts used in our evaluation are described in Section 4.

4 Experiments

1. RQ1: How can we measure the binary existence of knowledge in LLMs? Can we suggest a metric that maximizes separation between known and unknown data?
2. RQ2: How can we measure fine-grained knowledge extraction ability from LLMs? Which metric has the highest correlation with Attack success rate (ASR)?

4.1 RQ1: Fragmentation as a Proxy for Parametric Knowledge

Setup. We evaluate all nine fragmentation metrics under the protocol presented in (Section 3.3) across ten instruction-tuned LLMs from three families: Qwen ($\times 4$), Gemma ($\times 2$), and LLaMA ($\times 4$), ranging from 7B to 14B parameters (Appendix B), using batch size $B=10$ (full sensitivity sweep over $B \in \{1, 2, 5, 10, 20, 50\}$ justifying this choice in Appendix B.2). The known-concept class consists of 96 QA pairs drawn from the Harry Potter novels; the unknown-concept class consists of 100 QA pairs from the TOFU benchmark (Maini et al., 2024). All ten models score 0% QA accuracy on TOFU, casually confirming that no knowledge of these fictitious entities exists. For each model, fragmentation is computed at the per-metric layer that maximizes HP–TOFU separation. Appendix B.3 also reports a fixed, deployable heuristic, the layer at $0.8\times$ relative model depth, requiring no per-model search, that matches the oracle sweep’s discriminability.

Results. All nine metrics correctly separate HP from TOFU across all 10 models and survive BH (Benjamini-Hochberg false discovery rate control at $\alpha = 0.05$) correction (Table 1; per-model raw values in Appendix Tables 5-6). AUSS_{cos²} achieves the largest separation ($|d|=21.51$, $p_{\text{adj}}=0.002$) and AUSS_{L2} is second ($|d|=18.91$), both dispersion higher for TOFU than HP in 10/10 models; the scale-invariant classifier check below (AUC= 1.0) confirms this separation holds independent of each metric’s raw scale.

Entropy-based dispersion metrics (NormAnonDiff S_2 , AnonDiff S_2 , Reg S_2 ; Section 3.1) likewise reach 10/10 with large effect sizes. Coherence metrics, i.e. centroid norm $\|\Delta\hat{v}\|$, Reg $\overline{\text{cos}}$, AnonDiff $\overline{\text{cos}}$, and AnonDiff $\bar{L}_2$, all show strictly higher values for HP (8/10 or 10/10 models), confirming that a known concept induces a tighter, more directionally consistent hidden-state geometry than a purely fictional one. Effect sizes reflect per-metric peak-layer selection; performance under a label-free layer protocol would be lower (Appendix B.3).

The separation generalizes beyond a single narrow contrast and is not a scale artifact. To rule out that this separation reflects one narrow contrast, we extend the comparison to Harry Potter (retain split) and Star Wars against the same TOFU baseline; all 24 BH-corrected comparisons across the 8 AUSS metrics are signif-

Table 1: **RQ1: Metric discriminability** (HP known vs. TOFU unknown) across 10 LLMs, evaluated at the per-metric layer maximising HP–TOFU separation, sorted by Cohen’s $|d|$. Gap: mean $|\text{HP} - \text{TOFU}|$ with 95% bootstrap CI. p_{adj} : Wilcoxon signed-rank, BH-corrected ($n=10$); $**p<0.01$. *: to concept-differential vectors. $\dagger$: raw-activation baseline, no anon-vector subtraction. Per-model values in Appendix (Tables 5–6).

Metric	$ \delta $	Gap [95% CI]	p_{adj}
<i>Dispersion ($\downarrow$ HP expected; 10/10 models)</i>			
AUSS _{cos²}	21.51	0.686 [0.667, 0.705]	0.002**
AUSS _{L2}	18.91	1.203 [1.166, 1.239]	0.002**
NormAnonDiff S_2^*	4.21	1.206 [1.022, 1.332]	0.002**
AnonDiff S_2^*	3.50	0.804 [0.671, 0.941]	0.002**
Reg $S_2^\dagger$	0.92	0.117 [0.054, 0.200]	0.002**
<i>Coherence ($\uparrow$ HP expected)</i>			
$\ \Delta\hat{v}\ $	1.69	0.182 [0.118, 0.244]	0.002**
AnonDiff $\overline{\text{cos}}$	1.29	0.128 [0.069, 0.185]	0.010**
Reg $\overline{\text{cos}}^\dagger$	1.25	0.044 [0.023, 0.065]	0.002**
AnonDiff $\bar{L}_2$	1.21	2.634 [1.412, 3.946]	0.002**

icant ($q \in [0.0063, 0.0260]$), with HP-retain effect sizes reaching $d=6.20$, larger than the original HP-forget contrast (Appendix Table 7). A scale-invariant leave-one-model-out classifier trained on the standardized 8-metric peak-layer feature vector, evaluated on the original HP-forget vs. TOFU pair, confirms the known/unknown separation is not an artifact of raw effect-size scale: Logistic Regression reaches 100% accuracy (AUC= 1.0, 9/9 models fully correct) on the fresh, direction-corrected extraction (Appendix Table 8).

4.2 RQ2: Fragmentation as a Predictor of Unlearning Robustness

Setup. RQ1 established that fragmentation geometry reflects whether knowledge exists. RQ2 asks whether it also reflects how *recoverable* that knowledge remains after unlearning. Fragmentation metrics for each model–concept–method triple are computed at the layer that minimises S_2 over the post-unlearning forget-set queries (label-free protocol; Appendix B.3). The layer-resolved sweep evaluates every transformer layer independently (Appendix D.5).

Models and unlearning methods. We apply five unlearning methods spanning output-level suppression to representation-level distortion: **GA** (Yao et al., 2024) negates the forget-set CE loss; **DPO** (Rafailov et al., 2024) treats forget pairs as rejected and retain pairs as chosen; **NPO** (Zhang et al., 2024) applies a negative preference loss against a frozen reference; **NPO+KL** (Zhang et al.,

2024) augments NPO with KL retain regularisation; and **RMU** (Huu-Tien et al., 2025) steers forget-set hidden states at a fixed intermediate layer toward a random target direction. GA and DPO act primarily on outputs, whereas RMU explicitly distorts internal representations. This distinction is reliably captured by fragmentation metrics even when the two approaches yield indistinguishable ROUGE-L scores (Appendix B.5). These methods are applied to three instruction-tuned models: LLaMA-2-7B-chat, Meta-LLaMA-3-8B-Instruct, and Qwen2.5-7B-Instruct, each also evaluated as an unmodified **Base** control, across three target concepts: *Harry Potter* (Eldan and Russinovich), *Star Wars*, and *William Shakespeare* (Hong et al.) ($3 \times 6 \times 3 = 54$ conditions; incomplete runs shown as “—”). A run is verified ($\checkmark$) when Forget ROUGE-L < 0.10 and Retain ROUGE-L $\geq 0.80 \times$ baseline; surface suppression alone does not guarantee robustness. Four attacks then probe recoverability: Activation Steering (Seyitoğlu et al., 2024), ICL (Brown et al., 2020), GCG (Zou et al., 2023), and MIA (Fu et al.) at geometric, contextual, gradient, and distributional levels, respectively; lower values indicate stronger unlearning (Appendix B.6). Correlation analysis excluding Base and WHP (Who’s Harry Potter; Eldan and Russinovich) runs ($n=44\text{--}45$ non-base runs with complete attack data).

Results. Figure 2 summarizes attack success by method across all three models; full per-run values are in Appendix D.7 (Tables 13–15). Table 2 shows every BH-significant geometry predictor per attack ($n=44\text{--}45$ non-Base, non-WHP runs); full Spearman table in Appendix Table 9.

AUSS fragmentation geometry predicts ICL recoverability; entropy-family metrics predict GCG and Steering. ICL is the most significantly predictable attack via the AUSS family: AUSS_{L2} ($\rho = -0.51^{**}$), centroid norm ($\rho = +0.51^{**}$), and AUSS_{cos²} ($\rho = -0.50^{**}$) all survive BH correction; methods that geometrically fragment hidden-state representations more thoroughly are harder to reach via in-context priming. GCG and Steering are predictable in aggregate via the Gram-entropy (Reg) and AnonDiff families rather than AUSS: Reg $\overline{\text{cos}}$ ($\rho = -0.51^{**}$) and Reg S_2 ($\rho = +0.45^{**}$) predict GCG, and AnonDiff S_2 ($\rho = +0.48^{**}$) predicts Steering. MIA is omitted: no metric survives BH correction (Appendix Table 9). Leave-one-run-out validation across verified unlearning runs ($n=26\text{--}27$) confirms ICL is the only attack

Table 2: **RQ2: BH-significant geometry predictors per attack** (Spearman ρ , $n=44\text{--}45$ non-Base non-WHP runs with complete attack data, BH-corrected within each attack’s 9-metric column). All metrics shown survive BH at $q < 0.05$; full 9-metric table (all attacks, including non-significant cells) in Appendix Table 9.

Metric	ICL	GCG	Steering
AUSS _{L2}	-0.51^{**}	—	—
AUSS _{cos²}	-0.50^{**}	—	—
$\ \Delta\hat{v}\ $	$+0.51^{**\dagger}$	—	—
Reg $\overline{\text{cos}}$	—	-0.51^{**}	—
Reg S_2	—	$+0.45^{**}$	—
AnonDiff S_2	—	—	$+0.48^{**}$

that is both leave-one-out stable (8/8 metrics) and BH-significant (7/8); Steering (8/8) and GCG (7/8) are largely stable but not significant, while MIA (2/8) is neither reliably stable nor significant (Appendix D.6).

$^\dagger \|\Delta\hat{v}\|$ is the exact algebraic dual of AUSS_{L2} (identical $|\rho|$, opposite sign by construction, since AUSS_{L2} $= 2 - 2\bar{c}$ for mean pairwise cosine $\bar{c}$ and $\|\Delta\hat{v}\|$ is a monotonic function of the same $\bar{c}$), not an independent confirmation.

Layer-resolved geometry concentrates late in the network for the two attacks it reliably predicts. The aggregate correlations in Table 2 reflect peak-layer values averaged over all metric-layer combinations. When we instead sweep Spearman ρ across *every* transformer layer for both AUSS and entropy metrics (Figure 3; full results in Appendix D.5), only GCG and ICL survive global BH correction, and both peak in the same late-network region:

GCG: AUSS_{Jac} at **L29** reaches $\rho = -0.853$ ($p = 0.0008$, $n = 11$ runs with complete layer-resolved data, smaller than the $n = 44\text{--}45$ aggregate pool), with NormAnonDiff S_2 close behind at L30 ($\rho = -0.832$, $p = 0.0015$); the gradient-based suffix attack’s signal is shaped by whichever layer the model’s final next-token decision is most sensitive to.

ICL: Batch_Cov_TopEig at **L28** reaches $\rho = -0.886$ ($p < 0.0001$, $n = 15$), one layer from GCG’s own peak; in-context priming must similarly override the model’s late-stage, generation-ready representation.

MIA and Steering show only weak, non-BH-significant layer-specific signal in the current dataset (best entropy-family points at L31–L32, $\rho \approx 0.55\text{--}0.65$, uncorrected $p < 0.05$ but $q_{\text{global}} \geq 0.22$); we report these descriptively rather

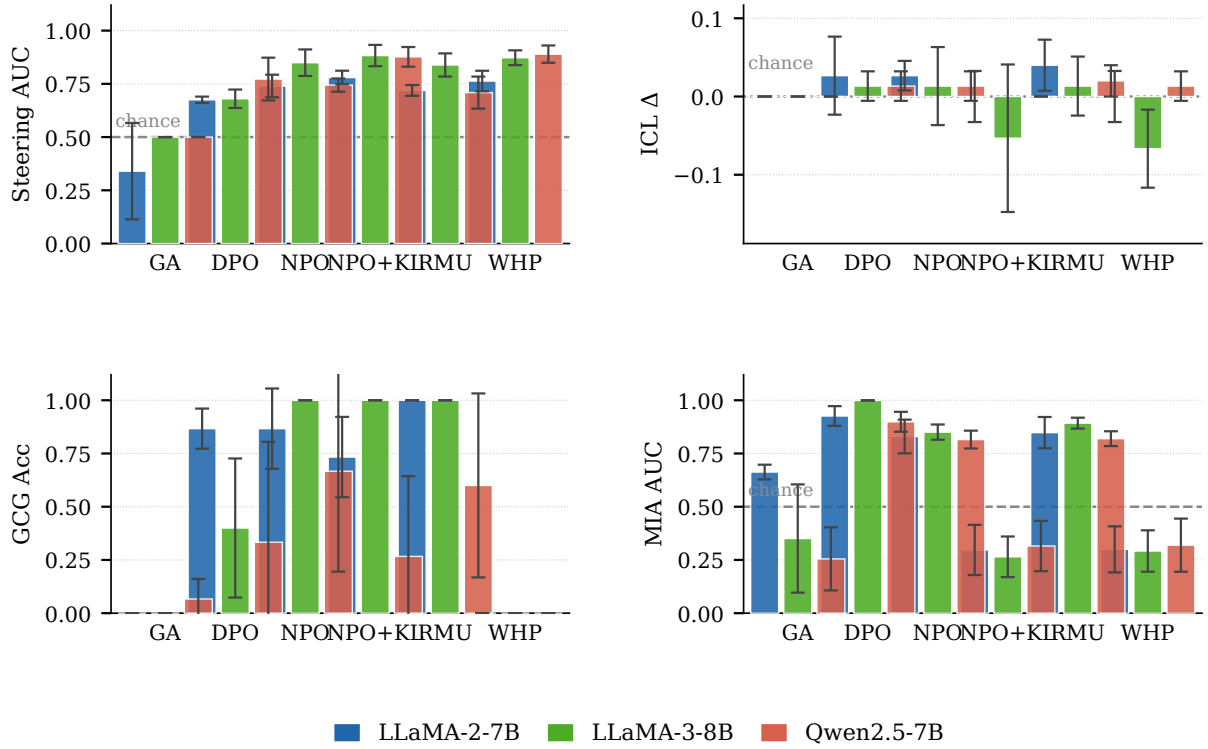


Figure 2: **RQ2: Mean attack success per unlearning method** across four attack types (2×2 grid), averaged over three concepts (error bars = std). Bars are grouped by unlearning method; colour indicates model (LLaMA-2-7B: blue, LLaMA-3-8B: green, Qwen2.5-7B: orange-red). Dashed lines mark chance level (0.5 for AUC-based attacks; 0 for ICL Δ). **GA** is the only method that consistently drives Steer-AUC near chance across all three models. **LLaMA-3** (green) reaches GCG Acc = 0 under every method, revealing strong model-family variation in gradient-attack vulnerability. ICL Δ is near zero across all methods and models; the geometric correlation with ICL recoverability emerges from the layer-resolved analysis (Table 11), not from this aggregate view. Full per-run values in Appendix D (Tables 13–15).

than as reliable predictors (Appendix Table 11).

The highest-performing layer is highly predictable across models, consistently residing within the 0.80–0.88 range of total model depth. A single fixed relative depth of $0.8 \times$ model depth, with no per-model information at all, already gives 8/8 metrics correct-direction and significant at $n=9$; leave-one-model-out transfer of this depth achieves 72/72 (model, metric) pairs correct-direction, identical to an oracle per-model search (Appendix B.3).

4.3 Benchmarking Against Prior Intrinsic Probes

Trivial raw-similarity baselines. Table 9 already includes head-to-head comparisons against trivial intrinsic baselines: the Reg S_2 (adapted from Skea et al. 2025) and Reg $\overline{\text{cos}}$ rows measure raw, non-anonymised hidden states with no anon-vector subtraction.

Head-to-head against published method. We benchmark AUSS directly against the representa-

tional metrics of Xu et al. (2025) (PCA-similarity, PCA-shift, CKA), which, unlike our method, require both pre- and post-unlearning model checkpoints. Across all 54 RQ2 configurations, BH correction over the full 44-test pool shows a mixed, attack-dependent picture: AUSS’s best metric outperforms Xu et al.’s best metric on ICL ($\rho = -0.572$, $q = 0.0001$ vs. $\rho = +0.381$, $q = 0.0212$) and Steering ($\rho = -0.452$, $q = 0.0042$ vs. $\rho = +0.428$, $q = 0.0076$), while Xu et al.’s CKA outperforms AUSS’s best metric on GCG ($\rho = +0.374$, $q = 0.0212$ vs. $\rho = -0.315$, $q = 0.068$, not BH-significant) and MIA (neither survives correction: $\rho = +0.336$ vs. $\rho = +0.085$). 11 of 44 tests survive BH correction; 8 of those 11 are AUSS metrics (full 44-test table in Appendix Table 10).

5 Discussion

The geometry of parametric knowledge. RQ1 establishes that activation fragmentation reliably distinguishes known from unknown concepts, and

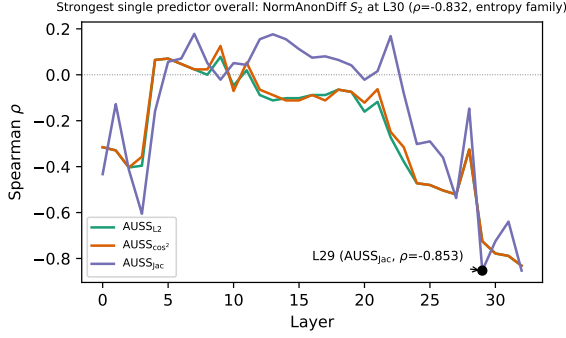


Figure 3: **Spearman ρ by layer, GCG attack.** Each line is one AUSS fragmentation metric; the peak at L29 (AUSS_{Jac} , $\rho=-0.853$) is marked. The strongest single predictor overall is NormAnonDiff S_2 at L30 ($\rho=-0.832$, entropy family; Appendix D.5). GCG recoverability is governed by late-layer geometry, close to where ICL’s own peak falls (Appendix D.5). (Full four-attack sweep in Appendix D.5.)

the geometry explains why. The anon-vector construction subtracts the mean hidden state of identity-stripped paraphrases, canceling the shared linguistic and syntactic signal and leaving a residual that reflects only the concept-specific content encoded at that layer. When a model parametrically knows a concept, its internal representations across distinct queries about that concept agree on a direction the residual vectors form a tight, low-entropy cluster. When the concept is absent from the parameters, no such agreement exists: the residual after syntax cancellation is noise, and the vectors scatter into high-variance fragmented directions. This geometric signature is what our metrics quantify, and it is entirely invisible to behavioral evaluation.

Architecture-agnostic universality. The coherence signature described above is not an artifact of any particular architecture. The 10/10 correct-direction result for the strongest metrics ($\text{AUSS}_{\cos^2}$, AUSS_{L_2} , AnonDiff S_2) in RQ1 spans three model families (Qwen, Gemma, and LLaMA) across different architectures, training recipes, and parameter counts from 2B to 14B. This cross-architecture consistency suggests that the geometric signature of parametric knowledge is not idiosyncratic to any particular model design. Fragmentation metrics qualify as model-family-neutral auditing tools, requiring no architecture-specific calibration.

Suppression versus deletion. A consistent RQ2 pattern is that reducing Forget ROUGE-L does not reliably reduce fragmentation: GA applied to LLaMA-3-8B reaches Forget ROUGE-L = 0.000

yet leaves AUSS_{L_2} and $\|\bar{\Delta}v\|$ nearly unchanged from the Base model. GA and DPO act primarily on the output distribution, leaving internal concept geometry largely intact, whereas RMU, which steers representations directly, shifts fragmentation more noticeably but does not erase it. This matches the redirection-rather-than-deletion picture reported by Xu et al. (2025) and Schoepf et al. (2026): representation-level interventions come closest to genuine deletion, and fragmentation metrics measure how far even they fall short.

Layer-resolved geometry concentrates late in the residual stream for the attacks it reliably predicts. Of the four attacks, GCG and ICL are the two whose layer-specific signal survives BH correction at the global scope (Appendix D.5), and both concentrate their strongest signal in the same late-network region: GCG peaks at L29 (AUSS_{Jac} , $\rho=-0.85$) and ICL at L28 (Batch_Cov_TopEig, $\rho=-0.89$), one layer apart. This is mechanistically coherent rather than coincidental. GCG’s adversarial suffix is optimised end-to-end by back-propagating through the entire network, so its gradient signal is shaped by whichever layer the model’s final next-token decision is most sensitive to. ICL must similarly override the model’s late-stage, generation-ready representation to redirect its output. Both attacks are therefore best diagnosed by measuring fragmentation deep in the residual stream rather than at an early or intermediate depth. MIA and Steering show only weak, non-BH-significant layer-specific signal in the current dataset (best entropy-family points at L31–L32, $\rho \approx 0.55$ – 0.65 , uncorrected $p < 0.05$ but not surviving global correction); we treat these descriptively rather than as reliable single-layer predictors. Appendix B.3 reports a practical, label-free layer-selection rule that performs close to this attack-specific oracle search without per-attack calibration.

Attack-dependent sign phenomena. The correlation’s sign is not uniform across attacks (Table 9): AUSS_{L_2} predicts ICL negatively ($\rho=-0.51$), while Reg S_2 and AnonDiff S_2 predict GCG and Steering positively ($\rho=+0.45$, $\rho=+0.48$). Based on the different mechanisms by which these attacks recover knowledge, we expect their correlation with fragmentation to differ in signal. ICL recovers "knowledge" through the model’s own generation, which plausibly requires a single coherent direction, empirically it seems that fragmentation tracks

the absence of one. GCG and Steering instead work by searching or intervening on the representation directly, where possibly a more dispersed representation offers more directions to exploit, which would explain the opposite sign. We discuss in Limitations how this account could be further validated in future work.

Practical implications for unlearning evaluation.

Activation fragmentation can serve as a computationally cheap, pre-attack screening step in unlearning evaluation pipelines. If fragmentation metrics confirm that the concept geometry has been genuinely disrupted, practitioners can have higher confidence that the unlearning is not merely suppression; if fragmentation remains intact, that is a strong signal that adversarial recovery will succeed without running expensive attacks. Its forward-pass-only design makes it applicable even where generation is prohibited, such as API-access audits subject to output restrictions.

6 Related Work

Parametric Knowledge and Intrinsic Evaluation.

Factual associations are geometrically structured within intermediate transformer layers (Petroni et al., 2019; Geva et al., 2021; Meng et al., 2023). While recent intrinsic probes aim to evaluate this knowledge without token generation (Gottesman and Geva, 2024; Hong et al.), they frequently conflate conceptual content with query syntax or require a reference model. We build upon the Gram-matrix entropy framework (Skean et al., 2025) and the geometric grounding of intermediate layers (Razzhigaev et al., 2023), but depart from prior work by measuring the *population geometry* of anon-vectors. This isolates conceptual identity from structural artifacts without relying on pre-unlearning baselines.

Machine Unlearning and Adversarial Robustness.

Current unlearning interventions—spanning gradient ascent, preference optimisation, and representation steering (Yao et al., 2024; Rafailov et al., 2024; Huu-Tien et al., 2025)—frequently produce *suppression* rather than genuine deletion (Xu et al., 2025; Lee et al., 2024; Schoepf et al., 2026). Because knowledge remains structurally intact, adversarial techniques such as activation steering (Seyitoğlu et al., 2024), in-context learning (Brown et al., 2020), and GCG (Zou et al., 2023) can bypass behavioral safeguards to recover

suppressed facts. Prior representation-level audits either require pre- and post-unlearning model comparisons (Xu et al., 2025) or the active execution of these costly attacks (Lynch et al., 2024). In contrast, our fragmentation framework offers a *single-phase*, generation-free proxy designed to *predict* vulnerability to these exact recovery channels using a single geometric forward pass.

7 Limitations

Our evaluation has several limitations. First, deterministic string substitution for anonymisation risks tokenizer-driven lexical artifacts. We mitigate this using morphologically similar names and multi-variant averaging ($K = 5$) in RQ2 (supported by consistent ablations in Appendix B.2), though future work should explore tokenization-matched replacements, LLM paraphrasing, or systematically benchmarking alternative anonymisation pipelines beyond the scheme used here. Second, relying on a single anonymised variant ($K = 1$) per query in RQ1 may retain idiosyncratic phrasing, which we offset by averaging across $N = 96\text{--}100$ queries per concept. Third, our batch-based construction requires $B \geq 2$ forget-set samples to define a valid direction vector, which may not suit single-sample or very small-sample unlearning settings; a promising extension augments a single target sample with neutral or synthetic data to compute a pseudo-batch direction, which we leave to future work. Fourth, the explanation offered in Section 5 for why fragmentation predicts ICL and GCG/Steering with opposite sign is inferred from the correlational pattern alone; we do not validate it against the models’ actual generations or intermediate reasoning traces, and confirming it would require inspecting chain-of-thought or generation-level outputs alongside the geometry, which we leave to future work. Finally, our concept coverage in RQ1, while now validated across four conditions (Harry Potter forget/retain, Star Wars, TOFU; Appendix D.2), remains limited to English-language fictional/factual domains. Expanding to graded knowledge levels and non-fiction technical domains is necessary to fully confirm the generality of our findings.

References

Tomer Ashuach, Dana Arad, Aaron Mueller, Martin Tutek, and Yonatan Belinkov. 2026. **CRISP: Persistent Concept Unlearning via Sparse Autoencoders.** *arXiv preprint*. ArXiv:2508.13650 [cs.CL].

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language Models are Few-Shot Learners](#). *arXiv preprint*. ArXiv:2005.14165 [cs.CL].
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. [Extracting Training Data from Large Language Models](#). *arXiv preprint*. ArXiv:2012.07805 [cs.CR].
- Ronen Eldan and Mark Russinovich. Who’s Harry Potter? Approximate Unlearning in LLMs.
- Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration.
- Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. 2025. [Erasing Conceptual Knowledge from Language Models](#). *arXiv preprint*. ArXiv:2410.02760 [cs.CL].
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. [Dissecting Recall of Factual Associations in Auto-Regressive Language Models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.
- Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. [Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer Feed-Forward Layers Are Key-Value Memories](#). *arXiv preprint*. ArXiv:2012.14913 [cs].
- Daniela Gottesman and Mor Geva. 2024. [Estimating knowledge in large language models without generating a single token](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4019.
- Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic Test of Unlearning Using Parametric Knowledge Traces.
- Dang Huu-Tien, Trung-Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. 2025. [On Effects of Steering Latent Representation for Large Language Model Unlearning](#). *arXiv preprint*. ArXiv:2408.06223 [cs].
- Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. 2024. [A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity](#). *arXiv preprint*. ArXiv:2401.01967 [cs.CL].
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, and 38 others. 2024. [The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning](#). *arXiv preprint*. ArXiv:2403.03218 [cs.LG].
- Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. [Eight Methods to Evaluate Robust Unlearning in LLMs](#). ArXiv:2402.16835 [cs].
- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. [TOFU: A Task of Fictitious Unlearning for LLMs](#). *arXiv preprint*. ArXiv:2401.06121 [cs.LG].
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. [Locating and Editing Factual Associations in GPT](#). *arXiv preprint*. ArXiv:2202.05262 [cs.CL].
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. [Language Models as Knowledge Bases?](#) *arXiv preprint*. ArXiv:1909.01066 [cs.CL].
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. [Direct Preference Optimization: Your Language Model is Secretly a Reward Model](#). *arXiv preprint*. ArXiv:2305.18290 [cs.LG].
- Anton Razzhigaev, Matvey Mikhalechuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2023. [The Shape of Learning: Anisotropy and Intrinsic Dimensions in Transformer-Based Models](#).
- Stefan Schoepf, Michael Curtis Mozer, Nicole Elyse Mitchell, Alexandra Brintrup, Georgios Kaissis, Peter Kairouz, and Eleni Triantafillou. 2026. [Redirection for Erasing Memory \(REM\): Towards a universal unlearning method for corrupted data](#). *arXiv preprint*. ArXiv:2505.17730 [cs].
- Atakan Seyitoğlu, Aleksei Kuvshinov, Leo Schwinn, and Stephan Günnemann. 2024. [Extracting Unlearned Information from LLMs with Activation Steering](#). *arXiv preprint*. ArXiv:2411.02631 [cs].
- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Mal-ladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. MUSE: Machine Unlearning Six-Way Evaluation for Language Models.

Oscar Skea, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. [Layer by Layer: Uncovering Hidden Representations in Language Models](#). *arXiv preprint. ArXiv:2502.02013* [cs].

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. 2025. [Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs](#). *arXiv preprint. ArXiv:2505.16831* [cs].

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. [Large Language Model Unlearning](#). *arXiv preprint. ArXiv:2310.10683* [cs.CL].

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. [Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning](#). *arXiv preprint. ArXiv:2404.05868* [cs.LG].

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. [Universal and Transferable Adversarial Attacks on Aligned Language Models](#). *arXiv preprint. ArXiv:2307.15043* [cs.CL].

A Mathematical Proofs

This section provides complete proofs for the theoretical results stated in Section 3.2.

A.1 Proof of Theorem 1

Theorem (High entropy forces geometric fragmentation). *Let $A \in \mathbb{R}^{n \times d}$ be a matrix of L_2 -normalised activation vectors and let $G = AA^\top$ be the corresponding Gram matrix with $\text{EffRank}(G) = \text{Tr}(G)/\lambda_{\max}(G)$. Then the mean pairwise squared Euclidean distance satisfies*

$$D_{L_2}^2 \geq 2 - \frac{2}{\text{EffRank}(G)}.$$

Proof. Let $v_1, \dots, v_n \in \mathbb{R}^d$ denote the rows of A , with $\|v_i\|_2 = 1$ for all i .

Step 1 (Gram matrix structure). Because the rows are unit vectors, $G_{ij} = v_i^\top v_j$ and the diagonal satisfies $G_{ii} = 1$, so $\text{Tr}(G) = n$. Hence $\text{EffRank}(G) = n/\lambda_{\max}(G)$.

Step 2 (Distance-cosine identity). For any pair $i \neq j$,

$$\|v_i - v_j\|_2^2 = \|v_i\|_2^2 - 2v_i^\top v_j + \|v_j\|_2^2 = 2 - 2G_{ij}.$$

Step 3 (Express $D_{L_2}^2$ via the Gram matrix). Define the mean off-diagonal cosine similarity

$$\bar{c} = \frac{1}{n(n-1)} \sum_{i \neq j} G_{ij},$$

so that $D_{L_2}^2 = 2 - 2\bar{c}$. Summing all entries of G and separating diagonal from off-diagonal,

$$\mathbf{1}^\top G \mathbf{1} = \text{Tr}(G) + \sum_{i \neq j} G_{ij} = n + n(n-1)\bar{c}.$$

Solving for $\bar{c}$ and substituting:

$$\begin{aligned} D_{L_2}^2 &= 2 - 2\bar{c} \\ &= \frac{2n - 2\mathbf{1}^\top G \mathbf{1}/n}{n-1}. \end{aligned}$$

Step 4 (Rayleigh quotient bound). For any unit vector x , $\lambda_{\max}(G) \geq x^\top G x$. Taking $x = \mathbf{1}/\sqrt{n}$ (which satisfies $\|x\|_2 = 1$):

$$\lambda_{\max}(G) \geq \frac{\mathbf{1}^\top G \mathbf{1}}{n}.$$

Step 5 (Conclusion). From Step 4, $\mathbf{1}^\top G \mathbf{1}/n \leq \lambda_{\max}(G)$, so $2n - 2\mathbf{1}^\top G \mathbf{1}/n \geq 2n - 2\lambda_{\max}(G)$. Dividing by $n-1 > 0$:

$$\begin{aligned} D_{L_2}^2 &= \frac{2n - 2\mathbf{1}^\top G \mathbf{1}/n}{n-1} \geq \frac{2(n - \lambda_{\max}(G))}{n-1} \\ &= \frac{2n}{n-1} \left(1 - \frac{1}{\text{EffRank}(G)} \right). \end{aligned}$$

Since $n/(n-1) \geq 1$ for all $n \geq 2$,

$$\begin{aligned} D_{L_2}^2 &\geq 2 \left(1 - \frac{1}{\text{EffRank}(G)} \right) \\ &= 2 - \frac{2}{\text{EffRank}(G)}. \end{aligned}$$

□

Remark. The proof actually establishes the tighter bound $D_{L_2}^2 \geq \frac{2n}{n-1} (1 - 1/\text{EffRank}(G))$; the theorem states the slightly weaker form $2 - 2/\text{EffRank}(G)$, which is independent of n and easier to interpret. The bound is tight (up to the $n/(n-1)$ factor) when all n vectors are mutually orthogonal: then $G = I_n$, $\lambda_{\max}(G) = 1$, $\text{EffRank}(G) = n$, and $D_{L_2}^2 = 2 = 2 - 2/n + 2/n$.

A.2 Entropy-EffRank Relationship

The following proposition formalises the claim in Section 3.2 that a high Rényi-2 Gram-matrix entropy S_2 forces a large effective rank, completing the implication chain $\text{high } S_2 \Rightarrow \text{high EffRank} \Rightarrow \text{large } D_{L_2}^2$.

Proposition 1 (Entropy–EffRank sandwich). *Let G be a positive semi-definite matrix with normalised spectral distribution $p_i = \lambda_i(G)/\text{Tr}(G) \geq 0$, $\sum_i p_i = 1$, ordered $p_1 \geq p_2 \geq \dots \geq 0$. Define $S_2(G) = -\log_2(\sum_i p_i^2)$ and $\text{EffRank}(G) = 1/p_1$. Then*

$$2^{S_2(G)/2} \leq \text{EffRank}(G) \leq 2^{S_2(G)}.$$

Proof. Upper bound. Because $p_1 \geq p_i$ for every i ,

$$\sum_i p_i^2 \leq p_1 \sum_i p_i = p_1.$$

Taking reciprocals (both sides positive) reverses the inequality: $1/p_1 \leq 1/\sum_i p_i^2$, i.e., $\text{EffRank}(G) \leq 2^{S_2(G)}$.

Lower bound. Trivially, $p_1^2 \leq \sum_i p_i^2$ (one non-negative term cannot exceed their sum). Taking square roots: $p_1 \leq \sqrt{\sum_i p_i^2} = 2^{-S_2(G)/2}$. Hence $\text{EffRank}(G) = 1/p_1 \geq 2^{S_2(G)/2}$. $\square$

Corollary 1 (High S_2 forces geometric fragmentation). *For any matrix $A \in \mathbb{R}^{n \times d}$ of L_2 -normalised vectors with Gram matrix $G = AA^\top$,*

$$D_{L_2}^2 \geq 2 - 2 \cdot 2^{-S_2(G)/2} = 2(1 - 2^{-S_2(G)/2}).$$

In particular, $D_{L_2}^2 \rightarrow 2$ as $S_2(G) \rightarrow \infty$.

Proof. From Proposition 1, $\text{EffRank}(G) \geq 2^{S_2(G)/2}$, so $1/\text{EffRank}(G) \leq 2^{-S_2(G)/2}$. Substituting into Theorem 1: $D_{L_2}^2 \geq 2 - 2/\text{EffRank}(G) \geq 2 - 2 \cdot 2^{-S_2(G)/2}$. $\square$

Remark. Proposition 1 shows that S_2 and EffRank are tightly coupled: both grow together as the spectral mass spreads across more eigenvalue directions. The upper bound $\text{EffRank} \leq 2^{S_2}$ also shows that the two quantities agree to within a factor of two on the log-scale, justifying the shorthand “ S_2 is an increasing function of EffRank ” used in the main text.

A.3 Proof of Theorem 2 and the Exact Entropy Identity

We first establish a key lemma that is an exact algebraic identity—no inequalities, no extra assumptions. It shows that S_2 and the mean squared pairwise cosine are two encodings of the same spectral quantity.

Lemma 1 (Gram entropy equals mean squared cosine). *Let $v_1, \dots, v_n \in \mathbb{R}^d$ be L_2 -normalised vectors with Gram matrix $G = AA^\top$ and normalised Gram $K_G = G/\text{Tr}(G) = G/n$. Let $\overline{\cos^2} = \frac{1}{n(n-1)} \sum_{i \neq j} (v_i^\top v_j)^2$ be the mean pairwise squared cosine similarity. Then*

$$\text{Tr}(K_G^2) = \frac{1 + (n-1)\overline{\cos^2}}{n} \quad (3)$$

equivalently $S_2(G) = -\log_2\left(\frac{1+(n-1)\overline{\cos^2}}{n}\right)$. In particular, S_2 is a strictly decreasing bijection of $\overline{\cos^2}$: $S_2 = 0 \Leftrightarrow \overline{\cos^2} = 1$ (all pairs parallel or anti-parallel), and $S_2 = \log_2 n \Leftrightarrow \overline{\cos^2} = 0$ (all pairs orthogonal).

Proof. Expanding the Frobenius norm of G :

$$\begin{aligned} \text{Tr}(G^2) &= \sum_{i,j=1}^n G_{ij}^2 \\ &= \underbrace{\sum_i G_{ii}^2}_{=n \text{ (diagonal}=1)} + \sum_{i \neq j} G_{ij}^2 \\ &= n + n(n-1)\overline{\cos^2}. \end{aligned}$$

Dividing by $\text{Tr}(G)^2 = n^2$:

$$\begin{aligned} \text{Tr}(K_G^2) &= \frac{\text{Tr}(G^2)}{n^2} = \frac{n + n(n-1)\overline{\cos^2}}{n^2} \\ &= \frac{1 + (n-1)\overline{\cos^2}}{n}. \end{aligned}$$

Since $\overline{\cos^2} \in [0, 1]$ (each $G_{ij} \in [-1, 1]$), the right-hand side is continuous and strictly increasing in $\overline{\cos^2}$, making $S_2 = -\log_2(\cdot)$ a strictly decreasing bijection. Boundary values: $\overline{\cos^2} = 1$ gives $\text{Tr}(K_G^2) = 1$, $S_2 = 0$; $\overline{\cos^2} = 0$ gives $\text{Tr}(K_G^2) = 1/n$, $S_2 = \log_2 n$. $\square$

Remark. Lemma 1 gives the converse of Theorem 1 for the squared-cosine metric: $S_2 \rightarrow 0$ if and only if $\overline{\cos^2} \rightarrow 1$, which means every pair of vectors satisfies $|v_i^\top v_j| \rightarrow 1$, i.e. they are either perfectly aligned or perfectly anti-aligned. Crucially, a low S_2 alone does not guarantee that vectors point in the same direction (positive cosines); two anti-parallel unit vectors achieve $S_2 = 0$ yet $D_{L_2}^2 = 4$. Theorem 2 closes this gap by adding the non-negativity condition, which exactly rules out anti-alignment.

We are now ready to prove Theorem 2.

Proof of Theorem 2. The key step is the following inequality, which holds under the assumption $G_{ij} \geq 0$ for $i \neq j$: since $\cos \theta_{ij} = G_{ij} \in [0, 1]$, we have $\cos^2 \theta_{ij} \leq \cos \theta_{ij}$, and therefore

$$\begin{aligned} \overline{\cos^2} &= \frac{1}{n(n-1)} \sum_{i \neq j} \cos^2 \theta_{ij} \\ &\leq \frac{1}{n(n-1)} \sum_{i \neq j} \cos \theta_{ij} = \bar{c} \end{aligned}$$

where $\bar{c}$ is the mean off-diagonal cosine. Rearranging, $\bar{c} \geq \overline{\cos^2}$.

From Lemma 1, $\overline{\cos^2} = \frac{n \operatorname{Tr}(K_G^2) - 1}{n-1} = \frac{n \cdot 2^{-S_2} - 1}{n-1}$, so

$$\bar{c} \geq \frac{n \cdot 2^{-S_2} - 1}{n-1}.$$

Using $D_{L_2}^2 = 2 - 2\bar{c}$ (Step 2 and Step 3 of the proof of Theorem 1):

$$\begin{aligned} D_{L_2}^2 &= 2 - 2\bar{c} \leq 2 - \frac{2(n \cdot 2^{-S_2} - 1)}{n-1} \\ &= \frac{2(n-1) - 2n \cdot 2^{-S_2} + 2}{n-1} \\ &= \frac{2n(1 - 2^{-S_2})}{n-1}. \end{aligned}$$

As $S_2(G) \rightarrow 0$, the right-hand side $\rightarrow 0$, so $D_{L_2}^2 \rightarrow 0$. $\square$

Remark. The bound is tight at both extremes. When all n vectors are identical ($v_i = v$ for all i): $G = \mathbf{1}\mathbf{1}^\top$, $\operatorname{Tr}(K_G^2) = 1$, $S_2 = 0$, and the bound gives $D_{L_2}^2 \leq 0$, matching the exact value $D_{L_2}^2 = 0$. When all pairs are orthogonal ($G_{ij} = 0$ for $i \neq j$, so $G = I_n$): $S_2 = \log_2 n$ and the bound gives $\frac{2n}{n-1}(1 - 1/n) = 2$, matching the exact value $D_{L_2}^2 = 2$.

B Implementation Details

This appendix provides the technical implementation details for all experiments. Code and data will be released upon acceptance.

All datasets and models utilized in this study, including MUSE-Books, TOFU, and ConceptVecs, alongside the evaluated language models, were accessed and employed strictly in accordance with their original licenses and intended use cases. These existing artifacts were explicitly designed for advancing machine learning research, which aligns directly with our objective of evaluating knowledge suppression and unlearning mechanisms. We

use these datasets and metrics under the MIT license, and therefore publish our code under the same license. Our reference implementation and framework are intended solely for academic and research contexts to audit language model safety; any derivative applications or data distributions must remain compatible with these research-oriented, non-commercial access conditions.

B.1 Metric Computation

Input format. For every concept query q_i we construct a *regular* text and an *anonymised* text. Regular text concatenates the natural question and its full answer: "{question}\nAnswer: {answer}", and the anonymised text applies a concept-specific substitution map to both fields (see Section B.2 for the three distinct anonymisation procedures). For anon-vector computation, the anon hidden state at layer ℓ is the mean over K anonymised variants:

$$h_\ell^{\text{anon}}(q_i) = \frac{1}{K} \sum_{k=1}^K h_\ell(a_i^{(k)}).$$

$K=1$ for RQ1 hidden-state extraction and RQ2 forget pairs; $K=5$ for the steering-attack SV computation in RQ2 (see Section B.6).

B.2 Anonymisation Procedures

All anonymised texts are produced by deterministic string substitution — no language model is used at any stage of the anonymisation pipeline. Three distinct procedures are used, depending on the dataset and downstream task.

Procedure 1: HP pairs for RQ1 (single-variant, fictional-name substitution). The MUSE-Books HP pairs are anonymised using the shared entity table from `data/generate_anon_qa.py` (HP_REPLACEMENTS), which contains 83 named-entity rules applied in order of decreasing string length to prevent partial matches (e.g. “Hermione Granger” is replaced before “Hermione”), and created using Seyitoğlu et al. (2024) and expanded with Claude-Opus-4.6. Replacements are plausible fictional names that carry no HP-specific signal; variant index 0 is selected, yielding a single deterministic anonymised pair per original pair ($K=1$). Key substitutions: *Harry Potter* $\rightarrow$ *Alex Ryden*, *Hermione Granger* $\rightarrow$ *Vera Ashwood*, *Ron Weasley* $\rightarrow$ *Sam Coldridge*, *Lord Voldemort* $\rightarrow$ *Dark Emperor Mordax*, *Albus Dumbledore* $\rightarrow$ *Headmaster*

Voss, Hogwarts → *Crystal Academy*, *Quidditch* → *Broomball*. The substitution is applied to both the question and the answer using a case-insensitive regex. This is the same table used by Procedure 3 for RQ2; using variant 0 ensures methodological consistency across tasks. 85 of 96 pairs are changed; the remaining 11 contain no HP-specific entity in the table.

Procedure 2: TOFU pairs for RQ1 (single-variant, author-table substitution). TOFU pairs reference entirely fictitious authors. We pre-scan all 100 pairs to collect every unique author full name, then build a global fictional-name map: each author is assigned `first_name[i] + last_name[i]` from the name pools of Seyitoğlu et al. (`tofu_anonymize.json`), indexed deterministically in order of first appearance (with explicit overrides from the dataset’s `replacements` field taking precedence). The map is applied consistently across all pairs (longest-first, case-insensitive), so cross-pair author identity is preserved throughout the dataset. 83 of 100 pairs are changed; the remaining 17 contain only generic phrasing with no author-name mention that the regex can match.

Procedure 3: ConceptVectors QA pairs for RQ2 (five-variant, entity-table substitution). For the steering-attack SV computation in RQ2, each evaluation question requires multiple anonymised variants so that the steering vector averages over several “identity-free” phrasings of the same question. We pre-generate $K=5$ such variants per question at data-build time using `data/generate_anon_qa.py`, which maintains a separate replacement table per concept. Each table lists 5 fictional substitutes for every named entity in the concept:

- **Harry Potter** → *Alex Ryden, Marcus Cole, Torin Drake, Daven Ash, Kael Stone* (and analogous per-character tables for Hermione Granger, Voldemort, Hogwarts, Quidditch, J.K. Rowling, etc.).
- **Star Wars** → *Kai Starborn, Zane Lightfield, Taran Skyes, Riven Starfall, Dax Brightstar* (for Luke Skywalker; equivalent tables for Darth Vader, Han Solo, Tatooine, lightsaber, etc.).
- **William Shakespeare** → equivalent five-name tables for Shakespeare himself, Hamlet,

Macbeth, the Globe Theatre, Stratford-upon-Avon, etc.

Substitution is again longest-first and literal. Variant k is constructed by replacing every matched entity with its k -th substitute (same index k for all entities within a single variant, so the five variants each represent a self-consistent fictional universe). The five anonymised questions are stored in the `anon_questions` list field of each `qa_eval_pairs` entry in the pre-built concept JSON files.

Note on RQ2 forget pairs. The Wikipedia-derived forget pairs used for *training* unlearning (not for the steering SV) use bracket placeholder tokens instead of fictional names (e.g. *Harry Potter* → [THE PROTAGONIST], *Lord Voldemort* → [THE ANTAGONIST], *Hogwarts* → [THE ACADEMY]). This produces a single anonymised pair per training example; multiple variants are not needed for the CE-loss computation.

B.3 Extraction, Sampling, and Layer Selection

Hidden-state extraction. All hidden states are extracted in a single forward pass with no gradient computation (`torch.no_grad()`). We register output hooks on every transformer block to collect the full stack of $L+1$ hidden states (embedding layer included) simultaneously for a single tokenised input. The per-token hidden state at position t and layer ℓ is mean-pooled across the sequence dimension using an attention-mask-weighted average, yielding a single vector $h_{\ell}(q_i) \in \mathbb{R}^d$ per input. All models are loaded in `bf16` precision with `device_map="auto"` to distribute parameters across available GPUs.

Batch sizes and sample counts.

- **RQ1 (knowledge detection):** $N=96$ Harry Potter query pairs (MUSE-Books) and $N=100$ TOFU pairs, each split into $B=10$ random batches of approximately 9–10 pairs for computing per-batch direction vectors. All ten models are evaluated without repetition (embeddings are deterministic).
- **RQ2 (unlearning):** Concept-specific forget sets contain approximately 50–80 Wikipedia-derived text pairs (see Section B.4). Fragmentation metrics are computed on the full forget set at the AUSS_{L2} -peak layer identified individually for each model–concept pair; $B=10$ batches.

Rationale for hyperparameter choices. Several choices in our pipeline are not uniquely determined by the experimental design; we document them here for transparency. **$B=10$ batches:** from $B=2$ upward every metric separates HP from TOFU in the hypothesized direction, with effect size growing monotonically with B (Table 3); the paired Wilcoxon test’s resolution is bounded by $n=9$ models rather than B , so $B=10$ is a practical middle ground, not an arbitrary or fragile choice. $B=1$ is structurally undefined (no batch-to-batch variance with a single sample); a single-item reformulation that avoids this constraint entirely is discussed in Section 7.

Table 3: Batch-size sensitivity: Cohen’s d (Wilcoxon p) for HP vs. TOFU separation at $B \in \{2, 5, 10, 20, 50\}$, $n=9$ models, direction-corrected fresh extraction. All cells are in the hypothesized direction.

Metric	B	$d(p)$
$\ \Delta\hat{v}\ $ / Global_Coh.	2	0.14 (.82)
	5	0.64 (.13)
	10	0.72 (.13)
	20	0.92 (.04)
	50	1.13 (.02)
Batch_SecondMoment_TopEig	2	0.19 (.73)
	5	0.35 (.36)
	10	0.42 (.36)
	20	0.43 (.36)
	50	0.37 (.57)
AUSS _{L2} / Batch_Cov_Trace	2	0.17 (.73)
	5	0.62 (.13)
	10	0.74 (.10)
	20	0.93 (.04)
	50	1.10 (.02)
AUSS _{cos²}	2	0.31 (.43)
	5	0.40 (.36)
	10	0.49 (.30)
	20	0.47 (.30)
	50	0.43 (.20)
AUSS _{Jac}	2	1.25 (.01)
	5	1.18 (.02)
	10	1.18 (.01)
	20	0.97 (.02)
	50	0.54 (.16)
Batch_Cov_TopEig	2	0.17 (.73)
	5	0.42 (.30)
	10	0.18 (.65)
	20	0.28 (.43)
	50	0.33 (.43)

Seed 42: standard reproducibility seed; sensitivity to the partition seed is a stated limitation (Section 7). **Peak layer** = $\arg \max_{\ell} \text{AUSS}_{L2}$: AUSS_{L2} had the strongest RQ1 signal in preliminary runs; the same layer is used for all metrics within a run to avoid per-metric search costs. Per-layer correlation analysis confirms that the final layers carry the strongest ASR signal for most attacks. $K=1$ for RQ1 and forget pairs, $K=5$ for

steering SV: a single anonymised variant is sufficient for metric computation and minimises hidden-state extraction time; five variants are used for the steering vector to average over phrasings and reduce direction noise (Seyitoğlu et al., 2024). **ICL:** $n_{\text{shots}}=5$, **ROUGE-L threshold 0.30:** five-shot is the standard few-shot setting of Brown et al. (2020); the 0.30 threshold is loose enough to accept paraphrases and matches the TOFU benchmark convention (Maini et al., 2024). **GCG:** $n_{\text{eval}}=5$ (**shortest answers**), **500 steps, suffix length 20:** shortest-answer pairs converge fastest under gradient-based optimisation; 500 steps and a 20-token suffix are the nanogcg defaults from Zou et al. (2023). **MIA:** **50 forget / 50 retain:** equal-class balance for ROC AUC computation; standard in membership-inference literature. **RMU:** $\ell_r=7$, $c=20$, $\alpha_r=1200$: adopted directly from the defaults of Huu-Tien et al. (2025) for 7–8B models; $\ell_r=7$ is approximately the middle of the network. **WHP excluded from correlation analysis:** WHP redirects concept queries to anonymised answers rather than erasing the representation; its geometry is qualitatively different from the suppression-based methods (GA/DPO/NPO/NPO+KL/RMU), and including it would confound the correlation analysis.

Peak-layer selection. For RQ1, each metric selects its own peak layer ℓ^* independently as the layer that maximises the HP–TOFU separation for that metric (i.e. $\arg \max_{\ell} [\bar{m}_{\ell}^{\text{TOFU}} - \bar{m}_{\ell}^{\text{HP}}]$ for dispersion metrics, reversed for coherence metrics where HP should be higher). This per-metric peak is computed across all ten models simultaneously. For RQ2, ℓ^* is selected per run as the layer that maximises AUSS_{L2} on the post-unlearning checkpoint; the same layer is used for all other fragmentation metrics in that run.

A practical, label-free layer-selection heuristic. The oracle per-model peak layer is not idiosyncratic: relative depth (ℓ^*/L) clusters tightly across all 9 models (0.79–0.98), with the highest-performing layer consistently falling in the 0.80–0.88 range. A single fixed relative depth of $0.8 \times$ model depth, using no per-model information at all, already gives 8/8 metrics correct-direction and statistically significant at $n=9$. Leave-one-model-out transfer of the median peak relative depth (learned from the other 8 models, applied blind to the held-out model) achieves 72/72 (model, metric) pairs correct-direction, identical to the oracle per-model search (Table 4). For practical deployment, tak-

ing the layer at $0.8 \times$ depth is close enough to the oracle-optimal result and requires no per-model calibration.

Table 4: Fixed relative-depth sweep (no per-model search), $n=9$ models, Cohen’s d (p shown only at $r=0.8$, all $p \leq 0.012$ there); sign shown per metric’s own coherence/dispersion convention, all cells correct-direction at $r=0.8$. LOMO transfer: leave-one-model-out transfer of the median peak depth from the other 8 models, all 8 metrics, all 9 held-out models (72/72 pairs), identical to the 72/72 oracle per-model search.

Metric	r	d
$\ \overline{\Delta\hat{v}}\ $	0.2	0.42
	0.3	0.67
	0.4	1.34
	0.5	1.51
	0.6	1.72
	0.7	2.21
	0.8	3.30 (.0039)
AUSS _{L2}	0.2	0.48
	0.3	0.70
	0.4	1.30
	0.5	1.42
	0.6	1.62
	0.7	2.04
	0.8	3.01 (.0039)
AUSS _{cos²}	0.2	0.04
	0.3	0.12
	0.4	0.22
	0.5	0.31
	0.6	0.60
	0.7	0.96
	0.8	1.65 (.0078)
AUSS _{Jac}	0.2	0.34
	0.3	0.83
	0.4	1.21
	0.5	1.44
	0.6	1.81
	0.7	2.15
	0.8	2.60 (.0039)
Batch_Cov_TopEig	0.2	0.09
	0.3	0.18
	0.4	0.84
	0.5	1.42
	0.6	2.10
	0.7	2.60
	0.8	6.31 (.0039)

Gram-entropy computation. Eigenvalues of $K_Z = ZZ^\top / \text{Tr}(ZZ^\top)$ are computed with `torch.linalg.eigvalsh` (symmetric eigendecomposition, stable for positive-semidefinite matrices). Negative eigenvalues arising from numerical error are clamped to zero and the spectrum is renormalised to sum to one before computing entropy. For S_1 we mask eigenvalues $\lambda_i < 10^{-12}$ to avoid $\log(0)$. Batches with $\text{Tr}(ZZ^\top) < 10^{-10}$ or $N < 2$ are skipped and flagged as NaN.

B.4 Dataset Details

Harry Potter (HP) — RQ1. We use the **MUSE-Books** benchmark (Shi et al.), accessed via HuggingFace (muse-bench/MUSE-Books, configuration `knowmem`, split `forget_qa`). The split contains 300 open-domain question–answer pairs drawn from the seven Harry Potter novels. We retain $N=96$ pairs after filtering to those with non-empty answers; these cover characters, spells, locations, and plot events.

Each pair is anonymised using Procedure 1 (Section B.2): the 83-entry fictional-name table from HP_REPLACEMENTS, variant index 0, yielding a single anonymised pair per original pair ($K=1$). For example, *Harry Potter* $\rightarrow$ *Alex Ryden* and *Hogwarts* $\rightarrow$ *Crystal Academy*.

TOFU — RQ1. We use the **TOFU** dataset (Maini et al., 2024) (locuslab/TOFU, HuggingFace), which contains 4,000 question–answer pairs about 200 entirely fictitious authors generated by GPT-4 and never present in any public training corpus. We sample $N=100$ pairs uniformly at random (seed 42) from the full-retain split. Because the authors are fabricated, all ten models score 0.00 exact-match QA accuracy on TOFU, making it an ideal null baseline for the knowledge-detection experiment. Procedure 2 (Section B.2) is applied: a global fictional-name map over all TOFU fictitious authors, keeping author identity consistent across pairs. 83 of 100 pairs are changed; 17 pairs containing only generic phrasing are unchanged by design—the anon-vector should carry minimal concept-specific signal for unknown concepts.

ConceptVectors — RQ2. For the unlearning experiments we use three target concepts: **Harry Potter**, **Star Wars**, and **William Shakespeare**. Forget pairs and evaluation QA pairs are constructed from the **ConceptVectors** dataset (Hong et al.) (YihuaiHong/ConceptVectors, HuggingFace), which provides Wikipedia-sourced text completions and question–answer pairs for 293 named entities across multiple model families.

For each concept we use the test and development splits for both llama2-7b and olmo-7b entries, deduplicating by question text. Wikipedia article text is additionally segmented into sliding-window (prefix, continuation) pairs of approximately 128 tokens to increase the effective forget-set size. The final forget set contains 50–80 pairs

per concept. Retain pairs are drawn from the `unrelated_QA` field of *other* concepts in `ConceptVectors`, ensuring the retain set covers facts the model should continue to answer correctly throughout unlearning; typically 50 retain pairs are used.

Important implementation note: The `ConceptVectors` HuggingFace repository contains JSON files with heterogeneous schemas across model families; we bypass the `datasets.load_dataset()` API and load the JSON files directly from the local cache to avoid schema-mismatch errors.

B.5 Unlearning Details

Hardware. All unlearning experiments are conducted on two **NVIDIA L40S GPUs** (48 GB VRAM each, 96 GB total). Models are loaded with `device_map="auto"` in `bf16` precision, causing the HuggingFace accelerate library to automatically shard larger models (LLaMA-2-7B, LLaMA-3-8B, Qwen2.5-7B) across both GPUs. `PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True` is set at process start to eliminate CUDA memory fragmentation across sequential unlearning runs. Total GPU-hours were not directly logged, so we estimate a total from process timestamps recorded in our run logs. RQ1 hidden-state extraction averages ≈ 31 minutes per model across the ten evaluated models; the RQ2 unlearning-and-attack pipeline averages $\approx 10\text{--}20$ minutes per model-method-concept run across the 63-run grid (Appendix D.7), with additional passes for the layer-resolved sweep (Section D.5), the CKA/ L_2 baseline comparison (Section 4), and the per-sample analyses in this appendix. Both GPUs are occupied for the duration of every run (`device_map="auto"` shards each model across both). Extrapolating these per-run durations across the full set of runs logged for this project, we estimate total compute at approximately 120–150 GPU-hours on the hardware described above.

Common training configuration. Unless noted otherwise, all methods share the following hyperparameters: 10 training epochs, learning rate $\eta = 10^{-5}$, batch size $B = 4$, maximum sequence length 256 tokens, gradient clipping to norm 1.0, and the AdamW optimiser (8-bit variant via `bitsandbytes` when available, otherwise standard PyTorch AdamW). Gradient checkpointing is enabled throughout to reduce peak VRAM usage. A frozen reference-model copy is loaded onto

CPU memory (pinned) and moved to GPU only when needed, which is required by DPO, NPO, NPO+KL, and RMU.

Gradient Ascent (GA). GA negates the standard language-modelling cross-entropy loss on the forget pairs:

$$\mathcal{L}_{\text{GA}} = -\mathcal{L}_{\text{CE}}(\theta; \mathcal{D}_{\text{forget}}).$$

No retain regularisation is applied. This follows the original gradient-ascent unlearning baseline (Yao et al., 2024).

DPO. Direct Preference Optimisation (Rafailov et al., 2024) treats forget pairs as *rejected* continuations and retain pairs as *chosen* continuations. The DPO objective is:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right],$$

with $\beta = 0.1$. Our usage of DPO for unlearning follows the open-unlearning framework of Maini et al. (2024).

NPO. Negative Preference Optimisation (Zhang et al., 2024) pushes the policy away from the reference model on the forget set only:

$$\mathcal{L}_{\text{NPO}} = \frac{2}{\beta} \mathbb{E}_{\mathcal{D}_{\text{forget}}} \left[\log \left(1 + \left(\frac{\pi_{\theta}}{\pi_{\text{ref}}} \right)^{-\beta} \right) \right],$$

with $\beta = 0.1$.

NPO+KL. Extends NPO with a KL-divergence regularisation term on the retain set to preserve general model capability (Zhang et al., 2024):

$$\mathcal{L}_{\text{NPO+KL}} = \mathcal{L}_{\text{NPO}} + \alpha D_{\text{KL}}(\pi_{\theta}(\cdot|x_r) \parallel \pi_{\text{ref}}(\cdot|x_r)),$$

with $\alpha = 1.0$ and $\beta = 0.1$.

RMU. Representation Misdirection for Unlearning (Huu-Tien et al., 2025) steers the hidden states of forget inputs toward a fixed random direction $\mathbf{u}$ at a chosen intermediate layer ℓ_r , while simultaneously anchoring retain hidden states close to the frozen reference:

$$\mathcal{L}_{\text{RMU}} = \|h_{\ell_r}^{\theta}(x_f) - c\mathbf{u}\|_2^2 + \alpha_r \|h_{\ell_r}^{\theta}(x_r) - h_{\ell_r}^{\text{ref}}(x_r)\|_2^2$$

where $\mathbf{u} \sim \mathcal{N}(0, I)$ is sampled once and fixed, $c = 20$, $\ell_r = 7$ (approximately the middle of the network for 7–8B models), and $\alpha_r = 1200$. The random vector $\mathbf{u}$ is created in `bf16` to match the model’s parameter dtype.

WHP (WhoIsHarryPotter). Following Eldan and Russinovich, WHP trains the model to *redirect* concept-specific queries to anonymised answers, rather than maximising the forget loss. Forget pairs use the original question as input and the *anonymised* answer as the training target; the retain pairs use the standard question–answer format. The combined loss is:

$$\mathcal{L}_{\text{WHP}} = \mathcal{L}_{\text{CE}}(\theta; q_f, a_f^{\text{anon}}) + \alpha_r \mathcal{L}_{\text{CE}}(\theta; \mathcal{D}_{\text{retain}}),$$

with $\alpha_r = 1.0$. No frozen reference model is required (unlike the KL-based retain regularisers in NPO+KL and the original WHP paper), which avoids the GPU memory fragmentation caused by pinning a second full model copy to a fixed GPU. Because WHP *redirects* rather than erases the concept representation, it does not satisfy the standard forget ROUGE-L verification criterion (post-unlearn forget ROUGE-L rises rather than falls); we annotate WHP runs with a $\leftrightarrow$ symbol.

B.6 Attack Implementation Details

We evaluate four adversarial recovery attacks after unlearning. All attacks use seed 42 and operate on the post-unlearning checkpoint held in GPU memory immediately after training (the model is not reloaded between attacks). Attacks are run in the order Steering $\rightarrow$ ICL $\rightarrow$ MIA $\rightarrow$ GCG; GCG runs last because the nanogcg optimiser can corrupt the CUDA context on some GPU configurations, after which the model is freed.

Activation Steering. Following Seyitoğlu et al. (2024), we compute a *query-specific* steering vector for each evaluation question Q_i :

$$S_L(Q_i) = h_L(Q_i) - \frac{1}{3} \sum_{k=1}^3 h_L(Q_i^{*(k)}),$$

where $h_L(\cdot)$ is the hidden state at layer $L = n_{\text{layers}} - 2$ (the second-to-last transformer layer, consistent with the paper’s “just before the final layer” specification), extracted at the first token position. The steering vector is the raw difference (not unit-normalised), so the scale carries magnitude information.

The vector $S_L(Q_i)$ is injected **at the first decode step only** (when the key-value cache has been filled by the prompt and the model generates the first new token, i.e. the hidden-state tensor has sequence length 1), scaled by coefficient $c = 2.0$. Crucially, the vector is *not* applied during the prefill

phase (sequence length > 1), which would modify the KV cache and corrupt attention rather than steering generation. A stateful hook flag ensures the perturbation fires exactly once per generation call.

For each of 25 evaluation pairs, we generate $n = 500$ stochastic completions (temperature = 2.0, top- $k = 40$, max_new_tokens = 10) both with and without the steering hook. The score for each pair is the maximum word-frequency of any answer token across the 500 samples (word-frequency score, following the paper’s scoring protocol). The overall metric is the ROC AUC computed by treating steered scores as the positive class and unsteered scores as the negative class: AUC > 0.5 indicates that steering recovers more correct-answer tokens than no-steering, i.e. the concept geometry is partially intact.

In-Context Learning (ICL). ICL tests whether presenting the model with $n_{\text{shots}} = 5$ concept-specific examples as context is sufficient to recover the forgotten knowledge via in-context prompting.

From the available QA pairs, 25 pairs are held out for evaluation and the remainder form the demonstration pool. For each evaluation question, 5 demonstrations are sampled uniformly (excluding the current question) and prepended in the format “{Q}\nAnswer: {A}\n\n”. Responses are generated greedily (max_new_tokens = 50). A response is marked correct when ROUGE-L with the reference answer is ≥ 0.30 (soft-match threshold to account for answer paraphrasing). The attack score is the ICL accuracy; a higher score indicates that the model’s knowledge is still recoverable via context.

GCG Adversarial Suffix. We use the Greedy Coordinate Gradient (GCG) method of Zou et al. (2023) via the nanogcg Python library. To reduce computational cost while retaining sensitivity, we evaluate on the $n_{\text{eval}} = 5$ pairs with the shortest reference answers (easiest targets for gradient-based suffix optimisation).

Each suffix is initialised as 20 repeated “! ” tokens and optimised for 500 gradient steps against the first 8 tokens of the reference answer (a short target prefix converges substantially faster than the full answer string and corresponds to the affirmative-prefix strategy used in the original paper). The model’s chat template is applied automatically by nanogcg before generation. After optimisation, the adversarial input is constructed

as {question} + " " + best_suffix and decoded greedily (max_new_tokens = 256). A pair is marked successful if the generated text contains a keyword from the reference answer or achieves ROUGE-L ≥ 0.25 . The attack score is the attack success rate (ASR) across the 5 pairs.

Membership Inference Attack (MIA). We employ a standard likelihood-ratio MIA: the model’s per-token conditional log-likelihood $s(q, a) = -\log P_\theta(a | q)$ is used as a membership score.

We score $n_{\text{forget}} = 50$ forget-concept QA pairs and $n_{\text{retain}} = 50$ retain-set pairs (sampled from pairs about unrelated topics). QA pairs with factual short-answer format are preferred over Wikipedia-text pairs for the forget set, as gradient-ascent unlearning on Wikipedia-style text can produce degenerate NLL values exceeding 200 nats for date-prefixed completions. The ROC AUC is computed with forget-set pairs as the positive class (higher $-\log P$ = lower likelihood = model has successfully forgotten) and retain-set pairs as the negative class. AUC > 0.5 indicates the model still assigns higher probability to the forget set than the retain set, i.e. unlearning is incomplete at the distribution level.

C Additional Evaluated Metrics

This section details the supplementary geometric and baseline metrics evaluated in our study, expanding upon the primary metrics introduced in Section 3.1.

Baseline pair-wise metrics. These metrics apply standard geometric measures directly to the raw hidden states and serve as comparison baselines for the anon-difference metrics.

Reg S_2 — Gram-matrix entropy of the regular hidden states:

$$\text{Reg } S_2 = S_2(\mathbf{H}).$$

Reg $\bar{L}_2$ — Mean L_2 norm of regular hidden states:

$$\text{Reg } \bar{L}_2 = \frac{1}{N} \sum_{i=1}^N \|h_{\ell^*}(q_i)\|_2.$$

Reg $\overline{\text{cos}}$ — Mean pairwise cosine similarity of regular hidden states:

$$\text{Reg } \overline{\text{cos}} = \frac{1}{N(N-1)} \sum_{i \neq j} \cos(h_{\ell^*}(q_i), h_{\ell^*}(q_j)).$$

Additional anon-difference pair-wise metrics.

AnonDiff $\bar{L}_2$ — Mean L_2 norm of anon-difference vectors:

$$\text{AnonDiff } \bar{L}_2 = \frac{1}{N} \sum_{i=1}^N \|\Delta v_i\|_2.$$

AnonDiff $\overline{\text{cos}}$ — Mean pairwise cosine similarity of anon-difference vectors:

$$\text{AnonDiff } \overline{\text{cos}} = \frac{1}{N(N-1)} \sum_{i \neq j} \cos(\Delta v_i, \Delta v_j).$$

Additional batch-mean (AUSS) metrics.

AUSS $_{\text{cos}}$ $\uparrow$ — Complement of the mean pairwise cosine similarity between batch-mean directions (the linear analogue of AUSS $_{\text{cos}^2}$):

$$\text{AUSS}_{\text{cos}} = 1 - \frac{1}{B(B-1)} \sum_{b \neq b'} \hat{d}_b \cdot \hat{d}_{b'}.$$

Unlike AUSS $_{\text{cos}^2}$, this is sensitive to the *sign* of alignment; it equals zero only when all B directions are perfectly parallel (fully consistent concept direction).

AUSS $_{\text{Jac}}$ $\uparrow$ — Mean pairwise soft Jaccard incoherence of the positive-component supports of the batch-mean directions. For unit vector $\hat{d}_b$, define its positive-activation profile $p_b \in [0, 1]^d$ with $p_{b,j} = \max(\hat{d}_{b,j}, 0)$. The pairwise soft Jaccard distance is

$$J(\hat{d}_b, \hat{d}_{b'}) = 1 - \frac{\sum_j \min(p_{b,j}, p_{b',j})}{\sum_j \max(p_{b,j}, p_{b',j})},$$

and

$$\text{AUSS}_{\text{Jac}} = \frac{1}{B(B-1)} \sum_{b \neq b'} J(\hat{d}_b, \hat{d}_{b'}).$$

This measures whether the *sparsity patterns* of the batch directions agree, independent of sign or magnitude effects captured by cosine metrics.

Covariance-based metrics. Let $\overline{\Delta v} = \frac{1}{N} \sum_i \Delta v_i$ be the mean anon-difference vector and $\hat{\Sigma} = \frac{1}{N-1} \sum_i (\Delta v_i - \overline{\Delta v})(\Delta v_i - \overline{\Delta v})^\top \in \mathbb{R}^{d \times d}$ be the sample covariance matrix.

Cov $_{\text{trace}}$ $\downarrow$ — Trace of the sample covariance of the anon-difference vectors:

$$\text{Cov}_{\text{trace}} = \text{Tr}(\hat{\Sigma}).$$

A large trace indicates concept signal is dispersed across many directions (fragmented); a small value indicates compact, concentrated representations.

$\text{Cov}_{\text{top}} \uparrow$ — Largest eigenvalue of the sample covariance:

$$\text{Cov}_{\text{top}} = \lambda_{\max}(\hat{\Sigma}).$$

This captures how much variance concentrates along the dominant concept direction. After unlearning, a large Cov_{top} indicates the primary concept direction is intact in the representation space, predicting susceptibility to context-based recovery such as ICL.

$\text{SecMoment}_{\text{top}} \uparrow$ — Largest eigenvalue of the (un-centred) second moment matrix:

$$\text{SecMoment}_{\text{top}} = \lambda_{\max} \left(\frac{1}{N} \sum_{i=1}^N \Delta v_i \Delta v_i^{\top} \right).$$

Unlike Cov_{top} , this is sensitive to both the mean direction $\overline{\Delta v}$ and the residual spread, making it a joint measure of magnitude and directionality.

D Full Experimental Results

D.1 RQ1: Per-Model Fragmentation and Entropy Values

See the full results in Table 5 and Table 6.

D.2 RQ1: Multi-Domain Validation and Classifier Discriminability

Four-way domain comparison. To test whether the HP-vs-TOFU separation reflects a structural known/unknown difference rather than an artifact of one narrow contrast, we extend the comparison to Harry Potter (retain split) and Star Wars, each tested against the same TOFU baseline across all 9 models (fresh, direction-corrected extraction), with two-sided Wilcoxon signed-rank tests and BH/FDR correction applied jointly across all 24 (metric, pair) tests.

Both additional “known” domains (HP-retain and Star Wars) separate from TOFU on every metric, in the same sign as the HP-forget contrast; HP-retain effect sizes are, if anything, larger (up to $d=6.20$) than the HP-forget contrast itself.

Classifier-based discriminability. The discriminability analysis in Table 1 reports Δ/CI , which depends on the scale of the underlying metric. To confirm the separation is not an artifact of that scale, we build one sample per (model, domain) pair from the 8-dimensional standardized AUSS peak-layer feature vector and train Logistic Regression and a small MLP (8 hidden units) under leave-one-model-out cross-validation, so both domains of a held-out

model are always scored by a classifier that never saw that model.

Both classifiers achieve near-perfect discriminability, with zero per-model misclassification for Logistic Regression, confirming the known/unknown separation is a robust structural feature of the representations rather than an artifact of the metrics’ raw scale.

D.3 RQ2: Full Geometry vs ASR Correlation Table

See the full results at Tables 13–15 and Table 9.

D.4 Head-to-Head Against Xu et al. (2026)

Full 44-test comparison (8 AUSS metrics + 3 Xu et al. metrics $\times$ 4 attacks) underlying the main-text summary in Section 4, BH-corrected globally across all 44 tests, $n=53$ –54. AUSS values come from checkpoints retrained specifically for this comparison (the originals were already deleted at the time this analysis was run), paired with the original checkpoints’ attack scores; this table should be read as an independent second data point rather than checkpoint-consistent with Table 9.

D.5 Layerwise Geometry Analysis

Methodology. Following established practice in mechanistic interpretability (e.g., causal tracing; Meng et al. 2023), we treat this layer-resolved sweep as an exploratory diagnostic map of *where* different adversarial mechanisms exploit the geometry, rather than a multiplicity-corrected hypothesis test for a new global predictor; the corrected significance counts below quantify how much of the raw signal survives once treated as the latter. We extend the correlation analysis of Section 4 to every transformer layer. For each run we extract hidden states at *all* L layers in a single forward pass with `output_hidden_states=True` (mean-pooled over the response tokens), giving vectors $\mathbf{h}_{\text{reg}}^{(\ell)}$ and $\mathbf{h}_{\text{anon}}^{(\ell)}$ at each layer $\ell \in \{0, \dots, L-1\}$. We compute Spearman ρ between each (metric, layer) pair and each attack-success score across all runs with complete data. Two metric families are evaluated in parallel:

- **AUSS metrics** (8 metrics, defined in Section 3): Centroid_Norm, AUSS_{L2}, AUSS_{cos²}, AUSS_{Jac}, Global_Coherence, Batch Covariance, etc.
- **Entropy / AnonDiff metrics** (14 metrics): Gram-matrix Rényi- α entropy of the regular

Table 5: Fragmentation metrics (HP known vs. TOFU unknown) for all ten LLMs at peak layer $\ell^*=18$. QA_{HP}: QA accuracy on HP; TOFU QA is 0.00 for every model. $\|\Delta\hat{v}\|_{\uparrow}$: centroid norm (coherence; $\uparrow$ = model knows). AUSS_{L2} $\downarrow$ / AUSS_{cos²} $\downarrow$: dispersion ($\downarrow$ = model knows). **Bold**: HP entry in the theoretically expected direction.

Model	Harry Potter (known)				TOFU (unknown, QA=0.00)		
	QA _{HP}	$\ \Delta\hat{v}\ _{\uparrow}$	AUSS _{L2} $\downarrow$	AUSS _{cos²} $\downarrow$	$\ \Delta\hat{v}\ $	AUSS _{L2}	AUSS _{cos²}
<i>Qwen family</i>							
Qwen-7B-Chat	0.083	0.752	0.952	0.577	0.639	1.309	0.683
Qwen2.5-7B	0.250	0.684	1.150	0.501	0.601	1.402	0.408
Qwen-14B-Chat	0.135	0.804	0.780	0.553	0.688	1.164	0.696
Qwen2.5-14B	0.323	0.697	1.116	0.522	0.689	1.151	0.440
<i>Gemma family</i>							
Gemma-2B	0.042	0.949	0.223	0.209	0.638	1.316	0.835
Gemma-7B	0.083	0.882	0.491	0.417	0.773	0.891	0.608
<i>LLaMA family</i>							
LLaMA-2-7B	0.125	0.800	0.800	0.549	0.553	1.543	0.792
LLaMA-2-13B	0.219	0.903	0.411	0.363	0.595	1.437	0.891
LLaMA-3-8B	0.312	0.900	0.423	0.370	0.655	1.268	0.826
LLaMA-3.1-8B	0.219	0.903	0.409	0.360	0.624	1.356	0.850

Table 6: Gram-matrix entropy (S_2 , Rényi order 2) and companion statistics at per-metric peak layer ℓ^* for ten LLMs. **Reg**: raw hidden states; **AnonDiff**: difference vectors $\Delta V = h_{\text{reg}} - h_{\text{anon}}$; applying Gram entropy to ΔV is our adaptation of Skean et al. (2025). $S_2\downarrow$: lower = more compressed = model knows concept. $\bar{L}_2\uparrow$, $\overline{\text{cos}}\uparrow$: higher = stronger/more coherent representation. **Bold**: HP entry consistent across all 10 models. Reg $\bar{L}_2$ omitted ($p_{\text{adj}}=0.43$, mixed direction).

Model	Harry Potter (known)						TOFU (unknown, QA=0.00)					
	Reg			AnonDiff			Reg			AnonDiff		
	S_2	$\overline{\text{cos}}$		S_2	L_2	$\overline{\text{cos}}$	S_2	$\overline{\text{cos}}$		S_2	L_2	$\overline{\text{cos}}$
<i>Qwen family</i>												
Qwen-7B-Chat	0.007	0.995		3.808	20.4	0.138	0.043	0.969		4.769	16.9	0.098
Qwen2.5-7B	0.002	0.999		3.715	29.5	0.118	0.009	0.996		4.614	24.7	0.121
Qwen-14B-Chat	0.018	0.990		3.544	21.5	0.166	0.090	0.954		4.795	16.7	0.088
Qwen2.5-14B	0.002	0.999		3.825	37.2	0.119	0.007	0.997		4.743	30.8	0.132
<i>Gemma family</i>												
Gemma-2B	0.327	0.867		3.584	15.4	0.328	0.757	0.845		4.098	12.7	0.057
Gemma-7B	0.053	0.973		3.758	2.43	0.254	0.200	0.895		4.397	1.79	0.108
<i>LLaMA family</i>												
LLaMA-2-7B	0.007	0.996		3.813	6.33	0.183	0.032	0.975		4.543	5.36	0.048
LLaMA-2-13B	0.055	0.967		1.635	6.90	0.275	0.221	0.868		2.106	5.16	0.044
LLaMA-3-8B	0.045	0.973		3.755	1.95	0.263	0.197	0.893		4.580	1.57	0.072
LLaMA-3.1-8B	0.038	0.977		3.714	2.41	0.265	0.168	0.907		4.543	2.06	0.064

(Reg), concept-differential (AnonDiff), and row-normalised concept-differential (NormAnonDiff) hidden-state matrices, plus mean cosine and mean L_2 statistics.

Figure 4 shows the full ρ -vs.-layer sweep for all four attacks using the six leader metrics (three AUSS, three entropy). The GCG panel reproduces and extends Figure 3 from the main text.

Best layer-metric predictors per attack. Refer to Table 11.

Key findings.

1. **GCG and ICL are the only two attacks whose layer-specific signal survives correction.** Both peaks fall late in the network and one layer apart (GCG at L29, ICL at L28);

GCG survives global correction ($q=0.046$) but not the stricter per-attack-column correction ($q=0.057$), while ICL survives both. See the main-text Discussion for the mechanistic reading.

2. **MIA and Steering do not survive correction at either scope.** Their best (layer, metric) cells are only nominally significant at the raw $p<0.05$ level ($q_{\text{global}}\geq 0.22$ for both); we report them descriptively rather than as reliable predictors.
3. **The aggregate and layer-resolved analyses find GCG signal via different metric families.** The aggregate correlations in Table 9 find GCG significant via the Reg (Gram-entropy) family (Reg $\overline{\text{cos}}$, $\rho=-0.51$);

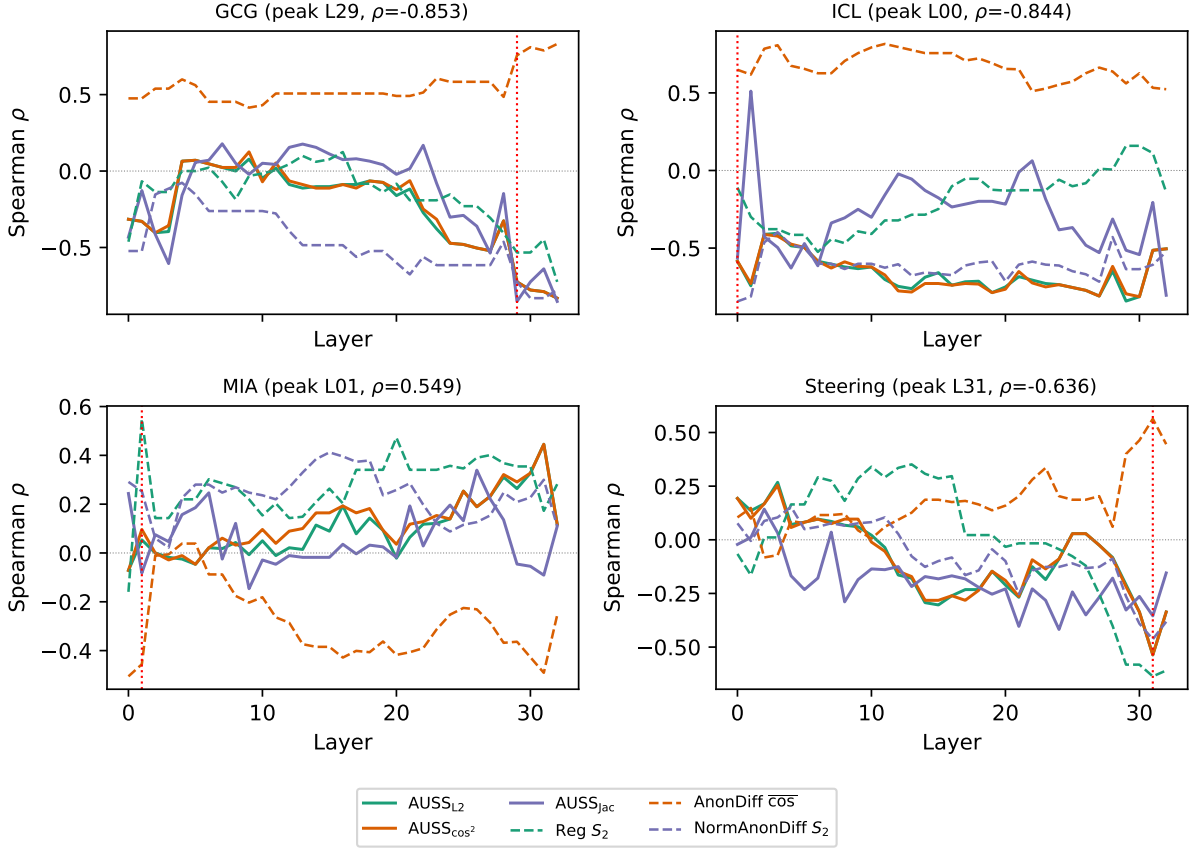


Figure 4: **Spearman ρ by layer for all four attacks, leader metrics.** Solid lines: AUSS family (AUSS_{L_2} , $\text{AUSS}_{\cos^2}$, AUSS_{Jac}). Dashed lines: Entropy family ($\text{Reg } S_2$, $\text{AnonDiff } \cos^2$, $\text{NormAnonDiff } S_2$). Dotted red vertical line: attack-specific peak layer among the six plotted metrics (GCG $\rightarrow$ L29, ICL $\rightarrow$ L00 for this restricted set; see Table 11 for the full-metric-universe peaks used in the main text, GCG $\rightarrow$ L29, ICL $\rightarrow$ L28). GCG and ICL both concentrate their strongest signal late in the network; MIA and Steering show weaker, non-BH-significant layer-specific signal in this dataset. $n=10\text{--}15$ runs per attack (complete data).

the layer-resolved sweep instead finds its strongest GCG signal via AUSS_{Jac} at L29 ($\rho=-0.85$). Both are real but reflect different aspects of the geometry: the peak-layer AUSS metric (argmax over AUSS_{L_2}) does not coincide with GCG’s own predictive layer.

4. **Positive ρ for entropy metrics on MIA and Steering.** Where entropy-family metrics show their strongest (non-significant) signal, higher entropy correlates with *higher* attack success, directionally opposite to the AUSS findings on GCG/ICL, reflecting that entropy measures spread of the concept-differential subspace while coherence measures its alignment.

D.6 Ranking Consistency of the Aggregate Correlations

To test whether the aggregate correlations in Table 9 are driven by one influential run, we compute leave-one-run-out (LOO) Spearman ρ for each of

the 8 AUSS metrics against each of the 4 attacks, restricted to verified unlearning runs ($n=26\text{--}27$).

Only ICL is both leave-one-out stable and statistically significant after BH correction (8/8 metrics stable, 7/8 significant). Steering and GCG are largely stable (8/8 and 7/8) but not significant; MIA is neither reliably stable nor significant. We therefore present the correlation findings within the context of the ICL attack, rather than as a generalized cross-attack ranking claim.

D.7 RQ2: Full Per-Run Results

Table 7: Four-way domain comparison: Cohen’s d (p , BH-corrected) for each condition vs. TOFU, $n=9$ models. All 24 cells are BH-significant ($q \in [0.0063, 0.0260]$).

Metric	vs. TOFU	d (p)
$\ \bar{\Delta\hat{v}}\ $ / Global_Coh.	HP-forget	+1.58 (.0039)
	HP-retain	+3.13 (.0039)
	Star Wars	+1.69 (.0039)
Batch_SecondMoment_TopEig	HP-forget	+2.57 (.0039)
	HP-retain	+6.20 (.0039)
	Star Wars	+4.62 (.0039)
AUSS _{L2} / Batch_Cov_Trace	HP-forget	−1.55 (.0039)
	HP-retain	−3.35 (.0039)
	Star Wars	−1.43 (.0117)
AUSS _{cos²}	HP-forget	−0.92 (.0195)
	HP-retain	−1.59 (.0039)
	Star Wars	−2.00 (.0039)
AUSS _{Jac}	HP-forget	−1.87 (.0039)
	HP-retain	−2.09 (.0039)
	Star Wars	−2.82 (.0039)
Batch_Cov_TopEig	HP-forget	−1.34 (.0078)
	HP-retain	−2.00 (.0039)
	Star Wars	−1.36 (.0039)

Table 8: Classification performance on known (HP) vs. unknown (TOFU) domains using standardized 8-dimensional AUSS features, LOMO-CV, $n=9$ models (18 samples).

	Logistic Regr.	MLP (8 hidden)
Accuracy	100%	94.4%
AUC	1.00	1.00
Correct	18/18	17/18
Models fully correct	9/9	8/9

Table 9: Spearman ρ between all fragmentation metrics and attack-success rate across $n=44\text{--}45$ post-unlearning runs (excluding Base and WHP). $**/*/\dagger$: BH-adjusted $p < 0.01/0.05/0.10$ within each attack column (9 metrics). *: adaptation of Gram-matrix entropy to concept-differential vectors (Skean et al., 2025). **Bold**: survives BH at $p < 0.05$. Only bold, starred cells ($q < 0.05$) represent statistically robust findings; unmarked values are shown for completeness so the full 9-metric $\times$ 4-attack grid is auditable, and should not be read as evidence of a relationship. None of MIA’s AUSS/AnonDiff correlations survive BH correction; see Section 4.

Metric	Steering	ICL	GCG	MIA
<i>AUSS dispersion</i>				
AUSS _{L2}	+0.00	−0.51**	−0.06	+0.04
AUSS _{cos²}	+0.06	−0.50**	−0.01	+0.10
$\ \overline{\Delta\hat{v}}\ $	−0.00	+0.51**	+0.06	−0.04
<i>AnonDiff</i>				
AnonDiff _{cos}	−0.24	+0.27	−0.09	−0.15
AnonDiff _{S₂[*]}	+0.48**	−0.22	+0.22	+0.20
AnonDiff _{L₂}	+0.15	+0.04	−0.05	−0.07
<i>Reg (raw-activation baseline)</i>				
Reg _{S₂[*]}	+0.27	+0.05	+0.45**	+0.29
Reg _{cos}	−0.31	+0.01	−0.51**	−0.27
Reg _{L₂}	−0.07	−0.09	−0.24	−0.16

Table 10: Full Xu et al. head-to-head comparison, sorted by q_{global} . $**/*$: $q < 0.01/0.05$.

Family	Metric	Attack	ρ	p	q_{global}
AUSS	Centroid_Norm	ICL	+0.551	0.00002	0.0001**
AUSS	AUSS _{cos²}	ICL	−0.572	0.00001	0.0001**
AUSS	AUSS _{L₂}	ICL	−0.551	0.00002	0.0001**
AUSS	Global_Coherence	ICL	+0.551	0.00002	0.0001**
AUSS	Batch_Cov_Trace	ICL	−0.551	0.00002	0.0001**
AUSS	Batch_SecMom_TopEig	ICL	+0.557	0.00001	0.0001**
AUSS	AUSS _{Jac}	Steering	−0.452	0.00067	0.0042**
Xu et al.	cka_mean	Steering	+0.428	0.00137	0.0076**
Xu et al.	pca_sim_mean	ICL	+0.381	0.00485	0.0212*
AUSS	Batch_Cov_TopEig	ICL	−0.385	0.00440	0.0212*
Xu et al.	cka_mean	GCG	+0.374	0.00529	0.0212*
Xu et al.	cka_mean	MIA	+0.336	0.01378	0.0505
AUSS	Batch_Cov_TopEig	GCG	−0.315	0.02020	0.0684
Xu et al.	cka_mean	ICL	+0.310	0.02409	0.0757
AUSS	Batch_Cov_TopEig	Steering	−0.287	0.03693	0.1083
Xu et al.	pca_sim_mean	MIA	+0.269	0.05126	0.1410
Xu et al.	pca_sim_mean	GCG	+0.263	0.05497	0.1423
AUSS	Centroid_Norm	GCG	+0.218	0.11245	0.2262
AUSS	Batch_Cov_Trace	GCG	−0.218	0.11245	0.2262
AUSS	Global_Coherence	GCG	+0.218	0.11245	0.2262
AUSS	AUSS _{Jac}	GCG	−0.218	0.11308	0.2262
AUSS	AUSS _{L₂}	GCG	−0.218	0.11245	0.2262
Xu et al.	pca_shift_mean	MIA	+0.213	0.12546	0.2300
AUSS	Batch_SecMom_TopEig	GCG	+0.212	0.12336	0.2300
Xu et al.	pca_shift_mean	GCG	−0.205	0.13778	0.2425
AUSS	AUSS _{cos²}	GCG	−0.189	0.17119	0.2897
AUSS	Batch_Cov_Trace	Steering	−0.110	0.43311	0.5775
AUSS	Global_Coherence	Steering	+0.110	0.43311	0.5775
Xu et al.	pca_sim_mean	Steering	−0.114	0.41497	0.5775
AUSS	AUSS _{L₂}	Steering	−0.110	0.43311	0.5775
Xu et al.	pca_shift_mean	ICL	+0.122	0.38399	0.5775
AUSS	Centroid_Norm	Steering	+0.110	0.43311	0.5775
AUSS	Batch_SecMom_TopEig	Steering	+0.111	0.42803	0.5775
AUSS	Batch_SecMom_TopEig	MIA	−0.079	0.57261	0.6460
AUSS	Batch_Cov_Trace	MIA	+0.080	0.56753	0.6460
AUSS	AUSS _{cos²}	MIA	+0.085	0.54492	0.6460
AUSS	Global_Coherence	MIA	−0.080	0.56753	0.6460
AUSS	AUSS _{L₂}	MIA	+0.080	0.56753	0.6460
AUSS	Centroid_Norm	MIA	−0.080	0.56753	0.6460
AUSS	Batch_Cov_TopEig	MIA	+0.066	0.63949	0.7034
AUSS	AUSS _{cos²}	Steering	−0.054	0.69991	0.7511
AUSS	AUSS _{Jac}	ICL	−0.051	0.71931	0.7536
AUSS	AUSS _{Jac}	MIA	+0.032	0.81909	0.8381
Xu et al.	pca_shift_mean	Steering	−0.011	0.93846	0.9385

Table 11: Best (layer, metric) predictor per attack in the layer-resolved analysis. n : number of runs with complete data at that layer. AUSS and entropy families are shown separately; the overall best metric across both is marked $\star$. q_{attack} : BH correction within the attack’s own column (8 AUSS + entropy tests); q_{global} : BH correction across all 4,224 (layer, metric, attack) tests at once. $*q < 0.05$.

Attack	Family	Best layer	Best metric	ρ	p	$q_{\text{attack}} / q_{\text{global}}$
GCG	AUSS	L29	AUSS _{Jac}	−0.853	0.0008	0.057 / 0.046*
	Entropy	L30	NormAnonDiff S_2 $\star$	−0.832	0.0015	0.057 / 0.046*
ICL	AUSS	L28	Batch_Cov_TopEig $\star$	−0.886	<0.0001	0.006* / 0.009*
	Entropy	L00	AnonDiff S_2	−0.861	0.0002	0.011* / 0.037*
MIA	AUSS	L31	Centroid_Norm	−0.446	0.170	0.999 / 0.581
	Entropy	L32	Reg $\bar{L}_2$ $\star$	+0.627	0.039	0.999 / 0.254
Steering	AUSS	L31	$\ \Delta \hat{v}\ $	+0.536	0.089	0.987 / 0.401
	Entropy	L31	Reg $\cos$ $\star$	+0.645	0.032	0.987 / 0.224

Table 12: Leave-one-out sign-flip stability, 8 AUSS metrics per attack, $n=26$ –27 verified runs.

Attack	Metrics with 0 LOO sign flips	BH-significant ($q < 0.05$)
ICL	8/8	7/8
Steering	8/8	0/8
GCG	7/8	0/8
MIA	2/8	0/8

Table 13: **RQ2: Full post-unlearning results, LLaMA-2-7B-chat.** ROUGE-L↓: forget-set ROUGE-L after unlearning. ✓: verified (ROUGE-L < 0.10 & retain ≥ 0.80× base). $\|\Delta\hat{v}\|\uparrow$: centroid norm; AUSS_{L2}↓, AUSS_{cos²}↓: dispersion (per-method peak layer ℓ^*). Steer-AUC↓: steering AUC; Steer-WF↓: mean steered word-frequency score. ICL $\Delta\uparrow$: ICL accuracy minus baseline. MIA↓: membership inference AUC; GCG↓: GCG ASR. Blank cells: checkpoint unavailable.

Method	Cpt	✓	Forget	Geometry ℓ^*			Attack Success ↓				
			ROUGE-L↓	$\ \Delta\hat{v}\ \uparrow$	AUSS _{L2} ↓	AUSS _{cos²} ↓	Steer-AUC	Steer-WF	ICL Δ	MIA	GCG
Base	HP		0.131	0.878	0.507	0.439	0.428	0.265	+0.000	0.827	0.000
	SW		0.152	0.876	0.516	0.447	0.449	0.296	+0.120	0.955	0.000
	WS		0.090	0.911	0.379	0.341	0.503	0.274	+0.040	0.837	0.000
GA	HP	✓	0.000	0.823	0.717	0.553	0.500	0.000	+0.000	0.678	0.000
	SW	✓	0.000	0.875	0.521	0.442	0.500	0.000	+0.000	0.695	0.000
	WS	✓	0.000	0.855	0.598	0.475	0.020	0.000	+0.000	0.615	0.000
DPO	HP	✓	0.082	0.868	0.549	0.469	0.694	0.283	−0.040	0.873	0.800
	SW	✓	0.098	0.866	0.554	0.474	0.673	0.281	+0.040	0.986	0.800
	WS	✓	0.092	0.913	0.371	0.335	0.658	0.287	+0.080	0.920	1.000
NPO	HP		0.127	0.860	0.580	0.492	0.756	0.286	+0.000	0.805	0.600
	SW		0.108	0.877	0.513	0.446	0.668	0.281	+0.040	0.937	1.000
	WS	✓	0.092	0.922	0.333	0.303	0.795	0.295	+0.040	0.748	1.000
NPO+KL	HP	✓	0.098	0.854	0.603	0.507	0.769	0.259	−0.040	0.449	0.600
	SW		0.105	0.867	0.550	0.473	0.746	0.272	+0.000	0.281	0.600
	WS	✓	0.091	0.917	0.354	0.320	0.823	0.286	+0.040	0.162	1.000
RMU	HP		0.131	0.878	0.509	0.441	0.710	0.264	+0.000	0.806	1.000
	SW		0.143	0.875	0.519	0.449	0.694	0.300	+0.040	0.951	1.000
	WS	✓	0.100	0.908	0.388	0.348	0.754	0.267	+0.080	0.787	1.000
WHP	HP		0.255	0.858	0.585	0.494	0.756	0.248	−0.040	0.442	0.000
	SW		0.175	0.856	0.595	0.505	0.708	0.269	+0.000	0.276	0.000
	WS		0.135	0.917	0.353	0.319	0.826	0.281	+0.040	0.181	0.000

Table 14: **RQ2: Full post-unlearning results, Meta-LLaMA-3-8B-Instruct.** ROUGE-L↓: forget-set ROUGE-L after unlearning. ✓: verified (ROUGE-L < 0.10 & retain ≥ 0.80× base). $\|\Delta\hat{v}\|\uparrow$: centroid norm; AUSS_{L2}↓, AUSS_{cos²}↓: dispersion (per-method peak layer ℓ^*). Steer-AUC↓: steering AUC; Steer-WF↓: mean steered word-frequency score. ICL $\Delta\uparrow$: ICL accuracy minus baseline. MIA↓: membership inference AUC; GCG↓: GCG ASR. Blank cells: checkpoint unavailable.

Method	Cpt	✓	Forget	Geometry ℓ^*			Attack Success ↓				
			ROUGE-L↓	$\ \Delta\hat{v}\ \uparrow$	AUSS _{L2} ↓	AUSS _{cos²} ↓	Steer-AUC	Steer-WF	ICL Δ	MIA	GCG
Base	HP		0.062	0.856	0.594	0.493	0.423	0.437	+0.000	0.915	0.000
	SW		0.113	0.873	0.529	0.455	0.424	0.379	+0.040	0.987	0.000
	WS		0.068	0.901	0.416	0.370	0.425	0.347	+0.040	0.952	0.000
GA	HP	✓	0.000	0.872	0.534	0.439	0.500	0.000	+0.000	0.700	0.000
	SW	✓	0.000	0.801	0.797	0.591	0.500	0.000	+0.000	0.102	0.000
	WS	✓	0.000	0.814	0.750	0.436	0.500	0.000	+0.000	0.250	0.000
DPO	HP	✓	0.000	0.867	0.551	0.466	0.720	0.076	+0.040	1.000	0.400
	SW	✓	0.017	0.842	0.646	0.519	0.620	0.017	+0.000	0.998	0.800
	WS	✓	0.000	0.923	0.328	0.298	0.700	0.037	+0.000	1.000	0.000
NPO	HP		0.112	0.863	0.566	0.482	0.930	0.493	+0.000	0.817	1.000
	SW		0.167	0.871	0.535	0.462	0.838	0.404	−0.040	0.900	1.000
	WS	✓	0.096	0.923	0.328	0.298	0.779	0.356	+0.080	0.834	1.000
NPO+KL	HP		0.178	0.867	0.552	0.472	0.945	0.488	−0.120	0.389	1.000
	SW		0.208	0.847	0.627	0.525	0.882	0.393	−0.120	0.249	1.000
	WS		0.121	0.921	0.335	0.305	0.822	0.313	+0.080	0.157	1.000
RMU	HP	✓	0.063	0.856	0.595	0.493	0.896	0.444	−0.040	0.862	1.000
	SW		0.123	0.871	0.538	0.460	0.854	0.394	+0.040	0.925	1.000
	WS	✓	0.072	0.902	0.416	0.370	0.766	0.350	+0.040	0.890	1.000
WHP	HP		0.174	0.859	0.583	0.493	0.922	0.483	−0.080	0.415	0.000
	SW		0.208	0.860	0.579	0.494	0.846	0.382	−0.120	0.283	0.000
	WS		0.103	0.926	0.316	0.288	0.850	0.334	+0.000	0.178	0.000

Table 15: **RQ2: Full post-unlearning results, Qwen2.5-7B-Instruct.** ROUGE-L↓: forget-set ROUGE-L after unlearning. ✓: verified (ROUGE-L < 0.10 & retain ≥ 0.80× base). $\|\overline{\Delta\hat{v}}\|↑$: centroid norm; AUSS_{L2}↓, AUSS_{cos²}↓: dispersion (per-method peak layer ℓ^*). Steer-AUC↓: steering AUC; Steer-WF↓: mean steered word-frequency score. ICL $\Delta↑$: ICL accuracy minus baseline. MIA↓: membership inference AUC; GCG↓: GCG ASR. Blank cells: checkpoint unavailable.

Method	Cpt	✓	Forget	Geometry ℓ^*			Attack Success ↓				
			ROUGE-L↓	$\ \overline{\Delta\hat{v}}\ ↑$	AUSS _{L2} ↓	AUSS _{cos²} ↓	Steer-AUC	Steer-WF	ICL Δ	MIA	GCG
Base	HP		0.106	0.872	0.534	0.439	0.433	0.373	+0.040	0.982	0.000
	SW		0.098	0.866	0.557	0.473	0.482	0.261	+0.080	0.982	0.000
	WS		0.080	0.907	0.395	0.352	0.495	0.373	+0.000	0.938	0.000
GA	HP	✓	0.000	0.936	0.275	0.253	0.500	0.000	+0.000	0.119	0.000
	SW	✓	0.000	0.901	0.417	0.370	0.500	0.000	+0.000	0.461	0.000
	WS	✓	0.000	0.934	0.284	0.255	0.500	0.000	+0.000	0.185	0.200
DPO	HP	✓	0.004	0.884	0.484	0.422	0.730	0.194	+0.000	0.861	0.000
	SW		0.117	0.834	0.676	0.548	0.677	0.343	+0.000	0.965	1.000
	WS	✓	0.020	0.895	0.441	0.389	0.911	0.238	+0.040	0.871	0.000
NPO	HP		0.198	0.863	0.569	0.478	0.744	0.374	+0.000	0.805	1.000
	SW		0.189	0.846	0.632	0.524	0.782	0.324	+0.000	0.871	0.000
	WS		0.105	0.913	0.370	0.331	0.706	0.433	+0.040	0.770	1.000
NPO+KL	HP		0.349	0.865	0.559	0.474	0.934	0.458	+0.000	0.482	0.000
	SW		0.167	0.838	0.662	0.542	0.822	0.210	+0.000	0.246	0.800
	WS		0.145	0.908	0.390	0.348	0.874	0.449	+0.000	0.218	0.000
RMU	HP										
	SW	✓	0.093	0.783	0.860	0.354	0.784	0.263	+0.000	0.854	1.000
	WS	✓	0.082	0.875	0.522	0.439	0.634	0.364	+0.040	0.785	0.800
WHP	HP		0.258	0.864	0.564	0.477	0.943	0.310	+0.000	0.495	0.000
	SW		0.172	0.843	0.644	0.533	0.845	0.205	+0.000	0.218	0.000
	WS		0.207	0.900	0.423	0.373	0.881	0.405	+0.040	0.245	0.000