UNPACK: Unlearnability Prediction via Activation Characterization of Knowledge

Chih-Hao Hsu [1] **Teng-Yun Hsiao** [2]

Abstract

Machine unlearning aims to remove specific knowledge from trained models, yet verifying true removal is difficult because extractable information may persist. We study whether geometric features of activation space representations predict unlearning effectiveness and vulnerability to extraction attacks. Across five unlearning methods and two attack types, simple features, particularly *separability* and *centrality*, show moderate predictive power ($|r| \approx 0.4$ to 0.5) for certain method and attack pairs. Different attacks align with different predictors: centrality with steering attacks and separability with prompt based attacks, suggesting distinct mechanisms. Nonlinear models substantially outperform linear ones, with $R^2 > 0.7$, indicating strong nonlinear structure. In addition to the experimental results above, we also conduct a theoretical analysis focused on the *separability* to further clarify how geometric features influence machine unlearning.

1. Introduction

Machine unlearning has emerged as a critical capability for deployed language models, driven by privacy regulations, copyright concerns, and safety requirements (Liu et al., 2024). The goal is to selectively remove specific knowledge (personal information, copyrighted content, or dangerous capabilities) while preserving the general utility.

Despite these advancements, a critical gap remains: inability to retrieve information does not imply its erasure. Evidence shows that "unlearned" knowledge can often be recovered via activation steering (Seyitoglu et al., 2024) or adversarial prompting. This necessitates investigating *whether specific structural characteristics determine if knowledge is permanently eliminated or remains susceptible to recovery.*

Prior work offers qualitative insights. Seyitoglu et al. (2024) observed that broad, interconnected topics are easily extracted after unlearning, while isolated facts resist extraction. Li et al. (2025) show that skills form distinct clusters in activation space, and LUNAR (Liu et al., 2025) exploit this by steering activations into "refusal regions" for unlearning.

These observations motivate our central hypothesis: **the geometric properties of knowledge representations in activation space may quantitatively predict unlearning outcomes**. Specifically, we ask:

- **RQ1:** Can simple geometric features of activation space serve as indicators for unlearning and attack outcomes?
- **RQ2:** Are these relationships consistent across different unlearning/attack methods? What do differences reveal?

Contributions. To answer the above questions, our contributions are listed as follows:

- We present **UNPACK** (**Un**learnability **P**rediction via **A**ctivation **C**haracterization of **K**nowledge), a systematic framework for analyzing the relationship between activation geometry and unlearning vulnerability.[1] This method include: (1) a pipeline for extracting geometric features from model activations and correlating them with unlearning/extraction outcomes; (2) empirical analysis across 5 unlearning methods and 2 attack types; (3) evidence that different attacks are predicted by different geometric features, suggesting distinct mechanisms; and (4) discovery of strong non-linear (threshold) effects that explain why simple correlations underestimate predictability.
- Motivated by our experimental findings, we analyze separability from a theoretical perspective. We introduce a model independent of existing unlearning theories and use it to quantitatively demonstrate the effect of separability.

2. Related Work

Machine unlearning methods can be categorized by mechanism: *gradient-based* approaches like Gradient Ascent

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan [2]Department of Physics, National Taiwan University, Taipei, Taiwan. Correspondence to: Chih-Hao Hsu <b11902080@ntu.edu.tw>, Teng-Yun Hsiao <b10502058@ntu.edu.tw>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]Code available at https://github.com/HowardHsuuu/UNPACK

(Jang et al., 2023) and Gradient Difference (Liu et al., 2022); *distribution-based* methods like KL Divergence that match outputs to a reference distribution (Maini et al., 2024); and *behavioral* methods like IDK that train refusal responses (Maini et al., 2024). Eldan & Russinovich (2023) show targeted fine-tuning can remove specific knowledge.

Knowledge Extraction Attacks. Activation steering (Seyitoglu et al., 2024) compute directions in activation space, recovering "unlearned" information especially for interconnected knowledge. Prompt-based attacks use jailbreaks or role-play to bypass unlearning.

Activation Geometry. Li et al. (2025) show skills form separable clusters in FFN space, enabling targeted unlearning. LUNAR (Liu et al., 2025) steer activations into refusal regions. Probing studies (Belinkov, 2022) extract semantic information from representations. We extend this by correlating geometric properties with unlearning outcomes.

3. Method

Our pipeline consists of five stages: (1) query preparation, (2) activation extraction from base and unlearned models, (3) geometric feature computation, (4) attack testing, and (5) predictive modeling. We analyze features from both pre-unlearning (base) and post-unlearning model activations.

3.1. Geometric Features

For each query, we extract activations from multiple layers and compute six geometric features:

Feature	Description / Formula
Local Density	Inverse of average k-nearest neighbor distance ($k = 10$): $\text{den}(x) = \frac{1}{\frac{1}{k}\sum_{i=1}^{k} d_{\cos}(x, x_i^{(nn)})}$
Separability	Ratio of inter-class to intra-class distance (higher = more distinct): $\text{sep}(x) = \frac{\mathbb{E}_{x' \in \text{other}}[d_{\cos}(x, x')]}{\mathbb{E}_{x' \in \text{same}}[d_{\cos}(x, x')]}$ (1)
Centrality	Distance to class centroid (higher = less typical): $\text{cent}(x) = d_{\cos}(x, \mu_{\text{class}(x)})$
Isolation	Minimum distance to any other-class point.
Cluster Compactness	Variance of pairwise distances within the class.
Cross-layer Consistency	Average correlation of representations across layer pairs.

For each feature, we aggregate across layers using mean, std, max, and min, yielding 24 features per query. We apply PCA (50 components) before computing features.

3.2. Unlearning and Attack Methods

We evaluate four unlearning methods from the TOFU benchmark (Maini et al., 2024) plus one pre-trained model. **Gradient Ascent** maximizes loss on the forget set, effectively reversing learning, though it risks catastrophic collapse where model utility degrades rapidly (Zhang et al., 2024). **Gradient Difference** balances this by subtracting forget-set gradients from retain-set gradients, preserving utility while inducing forgetting. **KL Divergence** minimizes divergence between the current model and a reference "never-learned" distribution on retain data, while maximizing loss on forget data. **IDK Response** is purely behavioral: it fine-tunes the model to output "I don't know" for target queries without removing the underlying knowledge. Finally, **WhoIsHarryPotter** (Eldan & Russinovich, 2023) uses targeted fine-tuning on reinforced examples to remove Harry Potter knowledge from Llama-2-7b.

For extraction attacks, we implement **Activation Steering** following the ANONACT method (Seyitoglu et al., 2024). This approach exploits the observation that replacing entity names in queries (e.g., "Harry Potter" $\rightarrow$ "[PERSON]") induces systematic shifts in hidden activations that encode entity-specific information. Steering vectors are computed as $v = h_{\text{orig}} - \bar{h}_{\text{anon}}$ and injected during generation with strength $\alpha = 2.0$, biasing the model toward reproducing the "unlearned" content. We also evaluate **prompt-based attacks** based on role-play and indirect elicitation, which operate semantically rather than geometrically.

3.3. Evaluation

We measure **Correct Answer Frequency (CAF)**: the fraction of 30 generations (temperature 2.0) containing the ground truth. **Unlearning Effectiveness** is the CAF reduction from base to unlearned model.

4. Experiment

Datasets and Models. We use **TOFU** (Maini et al., 2024) (100 queries about fictional authors, forget10 subset) and **Harry Potter** from MUSE-Books (Shi et al., 2024) (100 queries). For TOFU, we evaluate four unlearning methods on phi-1.5 using official benchmark checkpoints. For Harry Potter, we use the WhoIsHarryPotter model (Eldan & Russinovich, 2023) based on Llama-2-7b. We extract activations from layers 8–23 (phi-1.5) and 8–31 (Llama-2-7b), and train predictive models with 5-fold cross-validation.

We first present unlearning and attack effectiveness across methods, then analyze whether geometric features can predict these outcomes (RQ1), and finally examine consistency

Table 1. Unlearning and attack results. All values are Correct Answer Frequency (CAF) as percentages, except Eff which shows unlearning effectiveness.

Method	Base	Retain	Eff	Steer	Prompt
HP	11.9	4.3	95.7	4.5	9.0
Grad Diff	74.7	11.5	88.5	11.0	32.0
Grad Asc	73.7	0.8	99.2	0.8	1.0
KL	73.9	0.9	99.1	0.9	1.0
IDK	73.3	21.4	78.6	21.0	55.2

across experiments (RQ2).

4.1. Experimental Results

4.1.1. Unlearning and Attack Effectiveness

As shown in Table 1, Gradient Ascent and KL are most robust, with <1% extraction for both attacks. IDK is most vulnerable: 55.2% prompt CAF despite 78.6% unlearning effectiveness shows the model learned to refuse, not forget. Overall, prompt attacks outperform steering (19.6% vs 7.6% average CAF).

4.1.2. Predictive Power of Geometric Features

Linear Correlations. Only 10% of feature-target pairs show significant correlations ($|r| > 0.25$, $p < 0.05$). The strongest appear in Gradient Difference:

Table 2. Best single-feature predictors by target (Pearson r). All from TOFU_grad_diff experiment with base model geometry.

Target	Best Predictor	r
Unlearning	centrality_min	+0.471
Steering	centrality_min	−0.460
Prompt	separability_mean	−0.515

Key finding: Different attacks have different best predictors. Centrality predicts both unlearning success (+) and steering vulnerability (−): atypical knowledge (far from cluster center) is easier to unlearn but harder to extract via steering. Separability predicts prompt vulnerability (−): knowledge less separated from other classes is more vulnerable to prompt-based extraction.

Linear vs Non-linear Models. Table 3 reveals that non-linear models dramatically outperform linear models for most experiments, suggesting threshold effects rather than continuous relationships.

The gap between linear and non-linear performance (e.g., $-0.01 \rightarrow 0.79$ for Gradient Ascent) indicates threshold effects: geometric features predict outcomes well only when they cross certain values, not through continuous relationships. IDK and Harry Potter show poor predictability even with non-linear models, discussed in Section 6.

Table 3. R^2 comparison: Linear regression vs best tree-based model. Results for base model geometry predicting various targets. Negative R^2 indicates worse than mean prediction.

	Unlearning		Prompt CAF	
Experiment	Linear	NonLin	Linear	NonLin
Grad Asc	−0.01	**0.79**	−0.01	**0.79**
KL	−0.01	**0.59**	−0.01	**0.59**
Grad Diff	+0.18	**0.46**	+0.03	**0.24**
IDK	−0.42	−0.14	−0.32	−0.04
HP	−0.20	−0.14	−3.26	−4.62

Feature Importance and Decision Tree Rules. We conduct experiment to analyze the mean importance, and utilize decision tree rules to demonstrate the threshold effects. These rules suggest that knowledge with low centrality (more typical, closer to the cluster center) or high separability achieves near-perfect unlearning and near-zero extraction, as shown in Table 4.

Table 4. **Up:** Top features by mean permutation importance across all experiments and targets. **Down:** Decision tree rules showing threshold effects. Values show mean CAF for queries satisfying each rule.

Top features by mean permutation importance

Feature	Mean Importance	# Experiments
centrality_min	0.211	24
separability_mean	0.178	18
separability_std	0.138	21
local_density_max	0.134	27
centrality_std	0.122	21

Decision-tree threshold effects (mean CAF)

Rule	Unlearn	Attack
Gradient Ascent (base geometry):		
centrality_min ≤ 0.80	100%	0%
centrality_min > 0.80	92%	8%
KL Divergence (base geometry):		
centrality_max ≤ 1.47	100%	0%
centrality_max > 1.47	91%	9%
Gradient Diff (base geometry):		
separability_min ≤ 0.98	51%	46%
separability_min > 0.98	96%	2%

4.1.3. Geometry Changes After Unlearning

Table 5 shows features that change significantly after unlearning (Cohen's $d > 0.5$). Cross-layer consistency changes dramatically for all methods but in opposite directions: Gradient Difference decreases it while others increase it, possibly reflecting different removal mechanisms.

Table 5. Significant geometry changes after unlearning (Cohen's d; positive = increase).

Experiment	Feature	Cohen's d	Change
HP	consistency_mean	+0.89	↑
Grad Diff	consistency_mean	−3.41	↓
Grad Diff	density_min	−0.56	↓
Grad Asc	density_min	+0.74	↑
Grad Asc	consistency_mean	+2.69	↑
KL	consistency_mean	+3.87	↑
IDK	consistency_mean	+1.86	↑

4.1.4. Non-Monotonic Relationships

We detected 232 non-monotonic relationships by quartile binning: Inverted-U (84), U-shaped (74), and complex (74). For example, separability shows an Inverted-U with Prompt CAF in IDK:

Q1 (Low)	Q2	Q3	Q4 (High)
52%	**66%**	62%	42%

Moderate separability is most vulnerable: too low means entangled, too high means easily removed.

4.1.5. Cross-Experiment Consistency

Table 6. Features significant across multiple experiments.

Feature	Target	Mean r	# Exp
separability_mean	Unlearning	+0.36	2
separability_mean	Steering	−0.35	2
separability_min	Unlearning	+0.34	2
separability_min	Steering	−0.33	2

As shown in Table 6, features showing significant predictive power ($|r| > 0.25$) in 2+ experiments. Separability features show the most consistent relationships across experiments, while centrality, despite being the strongest single predictor, achieves significance primarily in Gradient Difference.

5. Theory

In this section, we analyze how geometry representation affects machine unlearning, especially the separability.

Problem Setting. We consider a student-teacher perceptron problem to understand how does separability affect the machine unlearning. Given data $x^\mu \in R^N$ generated from a distribution $p(x)$, its corresponding label y is generated by the teacher model T, i.e., $y^\mu = \text{sign}(T \cdot x^\mu)$ Therefore, we have the dataset as $\{x^\mu, y^\mu\}_{\mu=1}^P$. We define the α as $\alpha = \frac{P}{N}$. The learning problem here is that for a student vector J, it will have the prediction as $\tilde{y}^\mu = \text{sign}(J \cdot x^\mu)$. Therefore, the generalization error is $\epsilon_g = \frac{1}{\pi}\cos^{-1}(R)$. Here $R = \frac{J \cdot T}{\|J\|\,\|T\|}$. Recall that the separability is defined as (1). To ease the convenience of theoretical analysis, we consider the separability κ as

$$\kappa = \min_\mu \; J \cdot (y^\mu x^\mu)\,. \tag{2}$$

Remark 5.1 (**Comparison of Separability Definitions**). Although the two notions of separability differ in form, both (1) and (2) quantify class separation. The key difference is (2) captures global separability between classes, rather than pointwise separation at the level of individual data points.

Remark 5.2 (**On the Use of Simplified Models**). Although we do not analyze Transformers used in LLMs, prior work on machine unlearning demonstrates that simplified models can still suffice for theoretical insight. For example, (Yu et al., 2025) studies an overparameterized linear regression model for LLM unlearning, with model complexity comparable to ours, despite being much simpler than Transformer.

Derivation Sketch. To derive the generalization error, we consider the Gardner analysis (Engel, 2001) approach. We have the following two saddle point equations, which can be solved to derive the generalization error numerically

$$\begin{aligned} R =& 2\alpha \int_0^\infty dz\, p(z) \int_{-\infty}^{\kappa} \frac{dt}{\sqrt{2\pi(1-R^2)}} \\ & \cdot \exp\left[-\frac{(t-Rz)^2}{2(1-R^2)}\right] \left(\frac{z-Rt}{1-R^2}\right)(\kappa - t), \end{aligned} \tag{3}$$

$$\begin{aligned} 1-R^2 =& 2\alpha \int_0^\infty dz\, p(z) \int_{-\infty}^{\kappa} \frac{dt}{\sqrt{2\pi(1-R^2)}} \\ & \cdot \exp\left[-\frac{(t-Rz)^2}{2(1-R^2)}\right] (\kappa - t)^2. \end{aligned} \tag{4}$$

By solving the above two equations numerically, we can get the generalization error ϵ_g with respect to the α. For the dataset D, we can separate it as $D = D_{\text{retain}} \cup D_{\text{forget}}$. Then its corresponding alpha will be

$$\eta := \frac{|D_{\text{forget}}|}{|D_{\text{tot}}|}, \alpha_{\text{tot}} = \frac{|D_{\text{tot}}|}{N}, \alpha_{\text{ret}} = \frac{|D_{\text{retain}}|}{N} = (1-\eta)\alpha_{\text{tot}}.$$

Here $|D|$ denotes the number of data points in D. Therefore, the unlearning problem we consider here can be understood as follows: given a fluctuation in the dataset, how will the error be affected? In other words, $(\delta\alpha, \delta\rho) \to (\delta R, \delta\kappa) \to \delta\varepsilon_g$. We express the saddle-point conditions as a coupled root-finding problem by introducing

$$\begin{aligned} &F(R, \kappa; \alpha, p) := R - \Phi_1(R, \kappa; \alpha, p), \\ &G(R, \kappa; \alpha, p) := (1-R^2) - \Phi_2(R, \kappa; \alpha, p), \quad F = G = 0. \end{aligned} \tag{5}$$

Imposing that the perturbed solution and expanding to first order

$$J(R,\kappa)\begin{pmatrix}\delta R \\ \delta\kappa\end{pmatrix} = -\begin{pmatrix}\partial_\alpha F\,\delta\alpha + \int_0^\infty dz\, \frac{\delta F}{\delta p(z)}\,\delta p(z) \\ \partial_\alpha G\,\delta\alpha + \int_0^\infty dz\, \frac{\delta G}{\delta p(z)}\,\delta p(z)\end{pmatrix}, \tag{6}$$

Here $J := \begin{pmatrix} \partial_R F & \partial_\kappa F \\ \partial_R G & \partial_\kappa G \end{pmatrix}$. The forcing terms are explicit because Φ_1, Φ_2 are linear in α and depend on p only through the second term in (6):

$$\partial_\alpha F = -\frac{R}{\alpha}, \quad \partial_\alpha G = -\frac{1-R^2}{\alpha},$$
$$\frac{\delta F}{\delta p(z)} = -2\alpha\, A(z;R,\kappa), \quad \frac{\delta G}{\delta p(z)} = -2\alpha\, B(z;R,\kappa),$$

where A, B are the t-integrals induced by Φ_1, Φ_2, $A(z;R,\kappa) := \int_{-\infty}^{\kappa} dt\, K(t,z;R)\big(\frac{z-Rt}{1-R^2}\big)(\kappa - t)$ and $B(z;R,\kappa) := \int_{-\infty}^{\kappa} dt\, K(t,z;R)(\kappa - t)^2$. Consequently, whenever J is invertible and $\|J^{-1}\|$ remains bounded, the saddle-point shift is controlled by the forcing magnitude:

$$\left\| \begin{pmatrix} \delta R \\ \delta \kappa \end{pmatrix} \right\| \le \|J^{-1}\| \left\| \begin{pmatrix} -\frac{R}{\alpha}\delta\alpha - 2\alpha \int_0^\infty dz\, A(z;R,\kappa)\, \delta p(z) \\ -\frac{1-R^2}{\alpha}\delta\alpha - 2\alpha \int_0^\infty dz\, B(z;R,\kappa)\, \delta p(z) \end{pmatrix} \right\|. \tag{7}$$

Since the right-hand side of (7) depends on z and κ solely via the Gaussian kernel $K(t,z;R)$, the final result follows

$$\delta\varepsilon_g = \frac{d}{dR}\left(\frac{\cos^{-1}(R)}{\pi}\right)\delta R = -\frac{1}{\pi\sqrt{1-R^2}}\,\delta R$$
$$= O\left(e^{-\frac{(\kappa - Rz)^2}{2(1-R^2)}}\right), \quad (z \gg \kappa/R). \tag{8}$$

This leads to our theoretical conclusion that, as the separability between the two classes decreases, the generalization error decreases exponentially.

6. Discussion

Why different features for different attacks? Centrality predicts steering vulnerability while separability predicts prompt vulnerability, reflecting their mechanisms. Activation steering operates geometrically; atypical points (high centrality) are harder to target. Prompt attacks operate semantically; low separability means more pathways to trigger the knowledge. This suggests defenses must be attack-specific.

Why do HP and IDK show poor predictability? As shown in Table 3, these experiments yield negative R^2 even with non-linear models. For Harry Potter, the interconnected nature of fictional knowledge may require modeling the entire knowledge graph rather than single-query geometry; the base model also had limited HP knowledge (11.9% CAF). For IDK, the method achieves surface forgetting without true removal: knowledge representations stay intact while only output mapping changes, so geometric features cannot capture what determines vulnerability.

Why does consistency change most? Cross-layer consistency shows the largest effect sizes, suggesting unlearning disrupts information propagation. Gradient Difference *decreases* consistency while others *increase* it, possibly reflecting removal vs. adding refusal behavior.

Practical implications. Threshold rules (e.g., separability > 0.98) could pre-screen knowledge for unlearning success. Method selection could be guided by geometry. Defenses could monitor centrality for steering and separability for prompt threats.

7. Limitations

Our steering attack relies on regex-based anonymization rather than LLM-based methods, which may underestimate its effectiveness; Seyitoglu et al. (2024) also reports weak TOFU performance. Gradient Ascent and KL achieve near-complete unlearning, leaving little variance to model. The sample size (100 queries per experiment) limits statistical power, and results on phi-1.5 and Llama-2-7b may not generalize across architectures. Because each experiment draws queries from a single dataset, class-based features collapse: separability reduces to intra-class distance rather than an inter/intra ratio, and isolation becomes constant. Future work could integrate multiple datasets or define knowledge categories (e.g., by topic or entity type) to reintroduce meaningful inter-class structure.

8. Conclusion

We presented UNPACK, a framework for predicting machine unlearning outcomes using activation geometry. Our analysis reveals that simple geometric features (separability, centrality) show moderate predictive power ($|r| \approx 0.4$–0.5) for certain method-attack combinations. Attack types are predicted by specific features, with centrality guiding steering and separability guiding prompting, suggesting distinct mechanisms. Relationships are often non-linear, with tree models achieving $R^2 > 0.7$ where linear models fail. However, no universal predictor exists; practical frameworks must account for the specific unlearning mechanism. Inspired by these experimental results, we also conduct a theoretical analysis on separability as the first step for a better understanding of how representation affect unlearning.

Future directions. Future work should explore mechanistic interpretability, such as explaining why consistency changes most, and pursue geometry-informed defenses and cross-architecture generalization.

Impact Statement

This paper aims to improve understanding of machine unlearning effectiveness. While findings could inform extraction attacks, the primary impact is positive: enabling rigorous evaluation and identification of vulnerable knowledge before deployment.

References

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

Eldan, R. and Russinovich, M. Who's Harry Potter? approximate unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.

Engel, A. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Jang, J. et al. Knowledge unlearning for mitigating privacy risks in language models. 2023.

Li, Z. et al. Effective skill unlearning through intervention and abstention. *arXiv preprint arXiv:2503.21730*, 2025.

Liu, K. et al. Rethinking machine unlearning for large language models. *Stanford AI Lab Blog*, 2024. URL https://ai.stanford.edu/~kzliu/blog/unlearning.

Liu, Z. et al. Continual learning for natural language processing: A survey. *arXiv preprint arXiv:2211.12701*, 2022.

Liu, Z. et al. LUNAR: LLM unlearning via neural activation redirection. *arXiv preprint arXiv:2502.07218*, 2025.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. TOFU: A task of fictitious unlearning for LLMs. *arXiv preprint arXiv:2401.06121*, 2024.

Seyitoglu, F. et al. Extracting unlearned information from llms with activation steering. *arXiv preprint arXiv:2411.02631*, 2024.

Shi, W. et al. MUSE: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

Yu, J., He, Y., Goyal, A., and Arora, S. On the impossibility of retrain equivalence in machine unlearning, 2025. URL https://arxiv.org/abs/2510.16629.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

A. Extended Results

A.1. Full Correlation Table

Table 7 presents correlations for the Gradient Difference experiment (base geometry).

Table 7. Pearson correlations for TOFU_grad_diff (base geometry). Bold indicates $|r| > 0.3$.

Feature	Unlearn	Steering	Prompt
density_mean	−0.12	+0.08	+0.15
separability_mean	**+0.40**	**−0.42**	**−0.47**
separability_min	**+0.40**	**−0.41**	**−0.47**
centrality_min	**+0.47**	**−0.46**	**−0.47**
centrality_std	**−0.45**	**+0.43**	**+0.41**
consistency_mean	+0.15	−0.11	−0.08

B. Implementation Details

B.1. Hyperparameters

Table 8. Complete hyperparameter settings

Parameter	Value
Activation Extraction	
Layers (phi-1.5)	8–23
Layers (Llama-2-7b)	8–31
Precision	float16
Geometric Features	
k (k-NN)	10
PCA components	50
Distance metric	cosine
Steering Attack	
Steering strength α	2.0
Anonymizations per query	5
Target layer (phi-1.5)	12–13
Target layer (Llama-2-7b)	24
Generation	
Temperature	2.0
Top-k	40
Samples per query	30
Max new tokens	50
Predictive Modeling	
CV folds	5
Test split	20%
RF/GB estimators	100