The Geometry of Forgetting: Analyzing Machine Unlearning through Local Learning Coefficients

Aashiq Muhamed [1] **Virginia Smith** [1]

Abstract

Machine unlearning, the targeted removal of knowledge from LLMs, is vital for AI safety and privacy, yet robustly evaluating its success remains a significant challenge. Existing black-box evaluation protocols provide an incomplete picture of unlearning robustness, fail to explain utility loss mechanisms, or offer comprehensive guarantees. This work proposes a novel evaluation framework grounded in Singular Learning Theory (SLT), employing the refined Local Learning Coefficients (rLLC) to quantitatively analyze the geometric signatures imprinted by unlearning algorithms on neural network loss landscapes. We demonstrate that these rLLCs reveal distinct, layer-specific geometric changes for methods like Gradient Ascent (GA), Representation Misdirection (RMU), and Negative Preference Optimization (NPO), and that these geometric signatures correlate with macroscopic unlearning properties. Our analysis on TinyStories models substantiates these findings and highlights the utility of rLLCs in diagnostics, such as identifying RMU's intervention layer, positioning rLLCs as a powerful tool for advancing the principled evaluation of machine unlearning.

1. Introduction

Machine unlearning, the targeted removal or suppression of specific knowledge from trained models like LLMs while preserving general capabilities is increasingly vital for AI safety, privacy compliance, and responsible AI development (Yao et al., 2024; Barez et al., 2025). As models grow more powerful, managing their knowledge—especially potentially harmful or private information—becomes paramount. However, evaluating the success of unlearning remains a significant challenge (Liu et al., 2024). Standard evaluation protocols typically treat the unlearned model as a black box, measuring performance changes on specific forget and retain datasets (Shi et al., 2024; Maini et al., 2024; Thaker et al., 2024). While necessary, this approach provides an incomplete and potentially misleading picture. It struggles to measure the *robustness* of forgetting—unlearning might be superficial and easily reversed under relearning attacks (Hu et al., 2024; Farrell et al., 2024; Łucki et al., 2024). It fails to explain the *mechanisms* behind catastrophic utility loss on tasks beyond the retain set, and it offers limited *assurance* of safety or privacy beyond the specific examples tested (Barez et al., 2025; Hayes et al., 2024). A more robust evaluation paradigm requires methods that probe the internal, structural changes induced within the model by the unlearning process.

We propose a novel evaluation framework grounded in Singular Learning Theory (SLT) (Watanabe, 2009), which provides tools to analyze the geometry of neural network loss landscapes. Specifically, we employ refined Local Learning Coefficients (rLLCs) (Wang et al., 2024). The LLC (Lau et al., 2023), $\lambda(w^*)$, measures the local geometric complexity or effective dimensionality near a parameter point w^*; lower values indicate higher parameter degeneracy or simpler local geometry. Refined variants allow focusing this measure on specific parameter subsets (weight-refined wrLLC) or evaluating complexity relative to specific data distributions (data-refined drLLC). Our central hypothesis is that different unlearning algorithms imprint distinct *geometric signatures* detectable by layer-specific drLLCs, and that these signatures correlate with macroscopic properties like unlearning effectiveness, selectivity, and robustness. While complementary to mechanistic interpretability approaches that trace specific circuits (Elhage et al., 2022), our geometric lens offers a quantitative perspective on the structural changes induced by unlearning.

Applying this framework, we analyze Gradient Ascent (GA) (Jang et al., 2023), Representation Misdirection (RMU) (Li et al., 2024b), and Negative Preference Optimization (NPO) (Zhang et al., 2024). We show that these methods yield distinct layer-wise geometric signatures, particularly in terms of inter-layer variance and structural persistence.

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Aashiq Muhamed <amuhamed@andrew.cmu.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

For instance, GA tends towards uniform geometric degradation across layers, while RMU exhibits highly selective, layer-differentiated effects tied to its intervention point. To validate these findings we leverage the TinyStories dataset to enable controlled analysis of how geometric signatures evolve across models of varying scale (1M-28M parameters). This paper makes the following contributions:

1. We establish and apply a framework using refined LLCs (global and layer-specific drLLCs) to quantitatively evaluate and compare the geometric impact of different machine unlearning algorithms.
2. Our layer-wise rLLC analysis on TinyStories models empirically reveals distinct geometric signatures for GA, RMU, and NPO, quantified by inter-layer variance (σ) and ranking stability (ρ).
3. We show that the geometric view enables precise diagnostics, including accurately identifying the specific layer targeted by RMU's intervention.

By connecting macroscopic unlearning outcomes to microscopic geometric changes, this work positions rLLCs as a powerful tool for advancing the principled evaluation and understanding of machine unlearning techniques.

2. Background and Related Work

Approximate Unlearning Approximate machine unlearning seeks efficient methods to modify a target model $\mathcal{M}$ to remove the influence of specific undesired data or concepts represented by a forget set $\mathcal{D}_{\text{forget}}$, while preserving general capabilities associated with data $\mathcal{D}_{\text{retain}}$. (Liu et al., 2024; Barez et al., 2025). The goal is an unlearned model that behaves as if $\mathcal{D}_{\text{forget}}$ was never part of its training data $\mathcal{D}_{\text{pre}}$ (Bourtoule et al., 2021). This work studies three representative approximate unlearning strategies: **Gradient Ascent (GA)** directly optimizes the model parameters by maximizing the loss function (e.g., negative log-likelihood) on $\mathcal{D}_{\text{forget}}$, effectively pushing the model away from generating or recognizing forget content (Jang et al., 2023; Maini et al., 2024). **Representation Misdirection (RMU)** intervenes at specific layers, modifying the hidden representations for inputs related to $\mathcal{D}_{\text{forget}}$ by steering them towards a random noise vector (Li et al., 2024b; Barez et al., 2025). **Negative Preference Optimization (NPO)** leverages preference learning frameworks, treating examples from $\mathcal{D}_{\text{forget}}$ as implicitly rejected completions and optimizing the model to decrease their likelihood relative to a reference policy, often the original model $\mathcal{M}$ (Zhang et al., 2024). See Appendix A for additional descriptions of the methods.

Challenges in Unlearning Evaluation Evaluating the success of approximate unlearning, particularly for large language models (LLMs), presents significant challenges (Barez et al., 2025). Current evaluations predominantly rely on black-box metrics, such as performance changes (loss, accuracy, perplexity) measured on specific forget and retain datasets (Shi et al., 2024; Maini et al., 2024). However, this approach provides limited insight. It often fails to distinguish **superficial forgetting**—where knowledge appears removed but is easily recoverable via fine-tuning or targeted relearning attacks (Hu et al., 2024; Łucki et al., 2024; Farrell et al., 2024)—from genuine removal. Such evaluations may also inadequately capture **utility degradation** (catastrophic forgetting) on tasks unrelated to the specified retain set (Barez et al., 2025). Furthermore, they struggle to assess the **generalization** of unlearning beyond the specific forget examples, verify the prevention of **privacy leakage** through mechanisms like membership inference (Shi et al., 2024; Hayes et al., 2024), or provide assurance given the impracticality of retraining large models as a ground truth (Liu et al., 2024). These shortcomings necessitate evaluation methods that can probe the internal structural modifications induced by unlearning algorithms.

Singular Learning Theory (SLT) and Local Learning Coefficients SLT offers a mathematical framework for analyzing neural networks, which are typically singular—possessing parameter degeneracies where the Fisher information matrix is rank-deficient (Watanabe, 2009). Unlike classical theories assuming model regularity, SLT provides tools to study the non-Euclidean geometry of the loss landscapes inherent to deep learning (Wei et al., 2022). The **learning coefficient** λ in SLT measures the complexity of a model class relative to the true data distribution and appears in the asymptotic expansion of the Bayesian free energy (Watanabe, 2009). Its value reflects the geometric structure of the optimal parameter set $W_0 = \{w \mid K(w) = \inf K(w')\}$, where K is the Kullback-Leibler divergence to the true distribution (Watanabe, 2009; Lau et al., 2023). Extending this, the **Local Learning Coefficient (LLC)**, denoted $\lambda(w^*)$, quantifies the geometric complexity or effective dimensionality specifically in the vicinity of a parameter point w^* (e.g., a local minimum) (Lau et al., 2023). A smaller $\lambda(w^*)$ indicates higher degeneracy, implying a simpler geometry where parameters can vary more widely without substantially increasing the loss (Lau et al., 2023). The LLC can be estimated using the local free energy:

$$\hat{\lambda}(w^*) = n\beta \left[\mathbb{E}_{w\sim\pi(w|w^*,\beta,\gamma)}[\ell_n(w)] - \ell_n(w^*)\right] \quad (1)$$

Here, $\ell_n(w)$ is the empirical loss over n samples, and expectation $\mathbb{E}_{w\sim\pi}[\cdot]$ is over localized Gibbs posterior $\pi(w|w^*,\beta,\gamma) \propto \exp(-n\beta\ell_n(w) - \frac{\gamma}{2}\|w-w^*\|_2^2)$, parameterized by inverse temperature β and localization strength γ.

Refined Local Learning Coefficients For more targeted analysis, Wang et al. (2024) introduced refined Local Learning Coefficients (rLLCs). The **weight-refined LLC (wrLLC)**, denoted $\lambda(w^*; V)$, measures the geometric complex-

ity associated with a subset of parameters $V \subseteq W$, while keeping parameters $U = W \setminus V$ fixed at u^*. Formally, if $w = (u, v)$ with $u \in U, v \in V$, the wrLLC is defined via the volume scaling of $\{v \mid \ell(u^*, v) - \ell(u^*, v^*) < \epsilon\}$ within a neighborhood of v^*. Additionally, the **data-refined LLC (drLLC)**, $\lambda(w^*; q')$, measures complexity relative to a specific data distribution q' (e.g., $\mathcal{D}_{\text{forget}}$ or $\mathcal{D}_{\text{retain}}$) by using the loss $\ell'_n(w)$ calculated on data from q' in the LLC estimation. Combining these allows us to probe how different model components are geometrically affected by unlearning with respect to different data distributions. See Appendix A for additional background.

3. Methodology: Geometric Analysis of Unlearning

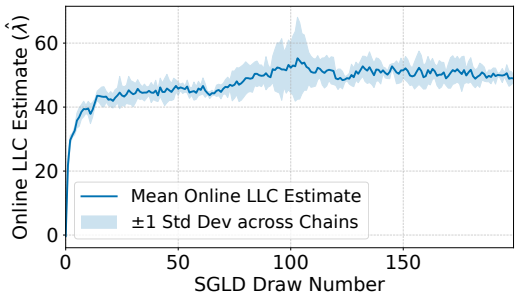

Figure 1: Example online LLC trace during SGLD sampling. The estimate converges as samples (w_t) explore the local posterior around w^*.

We apply rLLCs to analyze the geometric impact of unlearning. By measuring the local complexity of different model components (e.g., layers) with respect to relevant data distributions ($\mathcal{D}_{\text{forget}}$ and $\mathcal{D}_{\text{retain}}$) using layer and data-refined LLCs, denoted $\lambda(w^*; \text{Layer}_i, q')$, we probe the internal structural changes induced by different methods.

3.1. LLC Estimation via SGLD

We estimate the LLC using the local free energy formulation (Eq. 1). The challenge lies in estimating the expectation term $\mathbb{E}_{w\sim\pi}[\ell_n(w)]$ over the localized Gibbs posterior $\pi(w|w^*, \beta, \gamma)$. We employ Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011), a scalable MCMC method suitable for large models. SGLD approximates sampling from π by iteratively updating parameters w_t based on mini-batch gradients and injected Gaussian noise:

$$w_{t+1} \leftarrow w_t - \frac{\epsilon}{2}\left(n\beta\nabla_w \ell_m(w_t) + \gamma(w_t - w^*)\right) + \sqrt{\epsilon}\eta_t \quad (2)$$

where $\ell_m(w_t)$ is the mini-batch loss, ϵ is the step size, γ controls localization towards the reference point w^*, $n\beta$ is the effective inverse temperature modulating the loss landscape's influence, and $\eta_t \sim \mathcal{N}(0, I)$ is isotropic Gaussian noise facilitating exploration. The expectation is then approximated by averaging the full-batch loss $\ell_n(w_t)$ over T samples $\{w_t\}$ collected after an initial burn-in period, often across multiple independent chains for robustness (Figure 1). Estimating LLCs, especially refined variants, remains computationally intensive. While SLT justifications for the LLC estimator are strongest near local minima (Lau et al., 2023), its empirical application during dynamic phases like training has proven insightful in prior work (Wang et al., 2024). We adopt this practical approach, finding that despite estimating LLCs during the non-equilibrium unlearning process, the resulting geometric measures reveal consistent and interpretable dynamics. Nonetheless, careful interpretation is warranted due to potential sensitivity to hyperparameters and sampling variance in non-equilibrium settings.

3.2. Proposed Geometric Metrics

To translate the high-dimensional geometric information captured by layer-wise LLCs into interpretable measures of unlearning effects, we propose three key metrics derived from the profile of layer-specific drLLCs ($\lambda(w^*; \text{Layer}_i, q')$ across layers $i = 1..L$):

1. **Inter-Layer Variance ($\sigma_{q'}$):** Defined as the standard deviation of the LLC values across a relevant set of layers (e.g., L1-L7 in our experiments) for a specific data distribution q' ($q' = \mathcal{D}_{\text{forget}}$ or $q' = \mathcal{D}_{\text{retain}}$). A lower $\sigma_{q'}$ indicates that unlearning induces a geometrically more uniform state across layers with respect to q', while a higher $\sigma_{q'}$ suggests greater geometric differentiation among layers.
2. **Layer Ranking Stability ($\rho_{q'}$):** Calculated as the Spearman rank correlation between the LLC rankings of layers (L1-L7) at two different stages (e.g., end of finetuning vs. end of relearning), measured with respect to data q'. A higher $\rho_{q'}$ indicates that the relative geometric importance or contribution of different layers is better preserved through the unlearning and subsequent processes.
3. **Geometric Selectivity Index (log(GSI)):** Quantifies the selectivity of the unlearning process by comparing the relative change in inter-layer variance between the retain and forget distributions. Specifically, it measures the log-ratio of the relative change in retain variance to the relative change in forget variance between the target model ($\mathcal{M}$) and the unlearned state (UL):

$$\log(\text{GSI}) = \log\left(\frac{\sigma_{\text{Retain, UL}}}{\sigma_{\text{Retain, }\mathcal{M}}}\right) - \log\left(\frac{\sigma_{\text{Forget, UL}}}{\sigma_{\text{Forget, }\mathcal{M}}}\right) \quad (3)$$

where σ denotes the inter-layer variance computed over a representative subset of layers. A positive $\log(\text{GSI})$ suggests greater geometric selectivity, indicating that the unlearning process preserves geometric differentiation on the retain set relatively more than on the forget set.

These metrics aim to provide quantitative insights into the

uniformity, structural persistence, and selectivity of geometric changes induced by different unlearning methods.

3.3. Diagnostics: Geometric Fingerprint of RMU

Algorithm 1 Detecting the RMU Injection Block via Largest Positive Jump

Require: Epoch-averaged layer-wise LLC profile $\mathrm{LLC}(1{:}L)$
Ensure: Estimated noise injection layer $\hat{L}_{\text{noise}}$.
1: Calculate transitions
$\Delta(i) \leftarrow \mathrm{LLC}(i{+}1) - \mathrm{LLC}(i) \qquad (i = 1{:}L{-}1).$
2: Find index of largest positive jump:
$\hat{L}_{\text{noise}} \leftarrow \arg\max_i \max(0, \Delta(i)).$
3: **return** $\hat{L}_{\text{noise}}$.

Beyond quantitative evaluation, rLLC analysis can offer diagnostic capabilities, potentially revealing the internal mechanisms of specific unlearning methods. We look at RMU which intervenes locally by modifying activations $h_{L_{\text{noise}}}$ at a chosen layer L_{noise} for forget set inputs $x \in \mathcal{D}_{\text{forget}}$, typically by adding a steering vector v: $h_{L_{\text{noise}}}(x) \mapsto h_{L_{\text{noise}}}(x) + v$, for $x \in \mathcal{D}_{\text{forget}}$.

Theoretically, such a localized perturbation primarily alters the function computed by the downstream network layers. This functional change, in turn, modifies the local geometry of the loss landscape associated with the downstream parameters $w_{>L_{\text{noise}}}$. Since the LLC reflects local geometric complexity and effective dimensionality (Watanabe, 2009; Lau et al., 2023), the RMU intervention is expected to induce a significant, localized change or discontinuity in the layer-wise LLC profile specifically around the interface involving layer L_{noise}. This expectation can be formalized under specific assumptions about the model and the perturbation's effect (e.g., linear response approximation, properties of the loss Hessian or GGN matrix):

Theorem 3.1 (RMU-Induced Geometric Discontinuity - Informal)**.** *Assuming the RMU intervention sufficiently alters the local geometric structure downstream from layer L_{noise} (e.g., changing the effective rank of the parameter subspace), a quantifiable discontinuity in the layer-wise LLC profile is expected at the interface between layer L_{noise} and $L_{noise}+1$, relative to other layer transitions. (See Appendix B for formal derivation and assumptions).*

This theoretical discontinuity suggests a potential geometric signature. If it manifests empirically as a consistent, identifiable feature – for example, a distinct positive jump when $\mathrm{LLC}(L_{\text{noise}}+1) - \mathrm{LLC}(L_{\text{noise}})$ is significantly positive — then it can be used for detection. Based on this principle, we propose Algorithm 1 to identify the RMU injection site. The algorithm analyzes the layer-wise LLC profile and locates the index $k^\star$ corresponding to the largest positive inter-layer LLC difference. If our assumptions hold and the characteristic positive jump is a reliable consequence of RMU at the true injection layer, this algorithm estimates the intervention layer $\hat{L}_{\text{noise}} = k^\star$. This algorithm thus transforms rLLC analysis into a potential tool for generating an **auditable signature**, enabling post-hoc identification of the layer targeted by RMU.

4. Evaluating Unlearning Dynamics with LLCs on TinyStories

This section employs rLLCs to analyze the geometric dynamics of unlearning on small language models. We compare GA, RMU, and NPO to investigate how the observed geometric dynamics scale with model size.

4.1. Experimental Setup

We utilize Transformer models from TinyStories (Eldan & Li, 2023), specifically variants with **1M, 8M, and 28M parameters**, all featuring 8 layers but varying widths. These models are pretrained on the TinyStories dataset. The initial pretrained state for each size serves as our Base model. For $\mathcal{D}_{\text{retain}}$, we use the *train* split of `roneneldan/TinyStories` for training and the *validation* split for testing. For $\mathcal{D}_{\text{forget}}$, we use `andjela-r/mlm-harry-potter` for training and `vapit/HarryPotterQA` for testing which are widely used in the literature (Eldan & Russinovich, 2023). Loss values reported throughout are average negative log-likelihoods. These models provide a controlled environment to study the scaling properties of geometric unlearning dynamics, given their consistent architecture across varying sizes.

We study the dynamics across different phases, using the following protocol: 1. **Finetuning (FT):** The Base model for each size (1M, 8M, 28M) is finetuned on a 50/50 mixture of the $\mathcal{D}_{\text{retain}}$ and $\mathcal{D}_{\text{forget}}$ training data for 5 epochs (AdamW, LR `1e-5`, BS `32`), saving checkpoints FT1 through FT5. 2. **Unlearning (UL):** Starting from FT5, we apply GA, NPO, and RMU to all three model sizes for 5 epochs (AdamW, LR `5e-5`, BS `16`). For RMU, noise ($\alpha = 100$, steering coefficient $= 100$) is injected targeting the activations of Layer 3 in 1M, Layer 4 in 8M and Layer 5 in 28M. Due to computational limitations, NPO was applied only to the 1M model for 5 epochs (AdamW, LR `5e-5`, BS `16`, $\beta = 0.1$, using FT5 as reference). 3. **Relearning (RL):** To measure knowledge recovery, the unlearned models are retrained for 5 epochs (AdamW, LR `5e-5`, BS `16`) on the same 50/50 mixture of $\mathcal{D}_{\text{retain}}$ and $\mathcal{D}_{\text{forget}}$ training data used during FT (Additional details in Appendix C). The training splits are used during FT, UL, and RL stages, while evaluations involving test splits use the respective test datasets. We compute LLCs using the SGLD estimation procedure (Sec. 3.1), implemented via the `devinterp` library (van Wingerden et al., 2024). Key SGLD parameters were $N_{\text{chains}} = 4$, $N_{\text{draws}} = 200$, $\epsilon = 10^{-5}$, $\gamma = 100$, with

β computed per model state based on the training split size.

4.2. Overall Loss and Global LLC Dynamics

We first analyze the macroscopic effects of unlearning by examining overall loss trajectories and the dynamics of global, data-refined LLCs $\lambda(w^*; q')$ where q' is either $\mathcal{D}_{\text{forget}}$ or $\mathcal{D}_{\text{retain}}$.

Dynamics in the 1M Model Figure 2 shows the relative loss dynamics for the 1M model across the FT, UL, and RL phases for GA, RMU, and NPO. As expected, the **FT** phase reduces loss on both Forget and Retain sets, indicating adaptation. In the subsequent **UL** phase, all methods increase Forget loss, fulfilling the primary unlearning goal. However, GA incurs catastrophic utility damage (large loss increase on Retain), while NPO shows excellent utility preservation (Retain loss remains low). RMU sits in between, showing better Retain selectivity than GA but worse than NPO. We also observe that while RMU achieves low Forget loss on the training split during UL, its test loss is higher than GA or NPO, suggesting potential overfitting at this scale. The **RL** phase shows the fragility of current unlearning methods; Forget loss plummets rapidly for all methods upon re-exposure to the mixed data, indicating that the forgotten knowledge is easily reacquired.

The corresponding global drLLC dynamics (Figure 3) provide a geometric interpretation of these phases. **FT** leads to a significant increase in global LLC with respect to both Forget and Retain data, suggesting the model transitions to parameter regions of lower degeneracy (higher complexity) adapted to the task. During **UL**, the geometric complexity associated with the forget data decreases sharply for all methods, signifying a collapse towards more degenerate parameter configurations related to the forgotten knowledge. Reflecting the loss trends, GA shows the largest decrease in Retain LLC, NPO induces the largest LLC drop on Forget data, reaching the most degenerate state, while RMU also shows a substantial drop, greater than GA. The **RL** phase mirrors the loss recovery with a sharp restoration of geometric complexity on the Forget data for all methods, demonstrating the geometric plasticity of the model and the superficial nature of the unlearning achieved.

Condition	LLC	Loss	Ratio (LLC/Loss)
GA Forget	0.1442	0.1415	1.02
GA Retain	0.1810	0.1793	1.01
RMU Forget	0.1837	0.1813	1.01
RMU Retain	0.0921	0.0936	0.93

Table 1: Sensitivity Comparison (TinyStories-1M). Avg. derivative magnitude for normalized LLC vs. Loss. Comparable values suggest LLC tracks loss dynamics effectively.

Sensitivity Comparison We compare the sensitivity of LLC and loss as indicators of dynamic changes by examining the average magnitude of their step-wise derivatives, normalized to the same scale (Table 1 for GA/RMU). The results show comparable sensitivity across conditions (Ratio ≈ 1). This finding demonstrates that global LLC, while derived from a geometric perspective, is similarly responsive to changes during FT, UL, and RL as the standard loss metric. It can thus serve as an effective alternative for monitoring the progression of unlearning.

Model Scale Modulates Global Unlearning Geometry The geometric impact of unlearning, measured by global drLLC, is significantly modulated by model scale, as shown for GA and RMU in Figure 4. While all models (1M, 8M, 28M) exhibit analogous drLLC increases during FT, their responses diverge markedly during UL. GA induces geometric changes highly dependent on model size. The 28M model undergoes the most profound degeneracy increase (largest LLC drops), particularly on the Retain set, indicating severe utility impact geometrically. Counterintuitively, the 8M model displays remarkable geometric stability under GA, experiencing minimal drLLC disruption. This suggests GA's impact is not monotonic with size, or is sensitive to hyperparameter choice. RMU shows different scaling dynamics. On the Forget set, the 1M model suffers the largest geometric disruption, while the 8M and 28M models are less affected; while on the Retain set, both 8M and 28M models robustly preserve retain-set geometry. This contrasts sharply with GA and suggests RMU's localized mechanism scales more effectively in preserving utility-related geometry. The rapid Forget LLC recovery during RL across all settings further underscores the challenge of achieving robust unlearning.

4.3. Layerwise Geometric Analysis via Refined LLCs

While global loss and LLC dynamics (Section 4.2) provide a macroscopic view, understanding the internal mechanisms and selectivity of unlearning methods requires a finer-grained geometric perspective. We achieve this by computing weight- and data-refined LLCs on individual Transformer layers ($V = \text{Layer}_i$) and evaluating complexity relative to the Forget or Retain data distributions ($q' = \mathcal{D}_{\text{forget}}$ or $\mathcal{D}_{\text{retain}}$).

Figures 5 and 6 illustrate the evolution of these layer-specific, data-refined LLCs ($\lambda(w; \text{Layer}_i, q')$) for the 1M and 8M TinyStories models across the FT, UL, and RL stages (1M NPO and 28M results are in Appendix D, Figs. 9 and 8). As the final layer (L8) consistently shows near-constant LLC, we focus on layers L1-L7. We compute the σ across layers L1-L7 averaged over epochs within each stage (Table 2) and the ρ between FT5 and RL5 (Table 3). We use the calculated σ values to compute the log(GSI) to measure

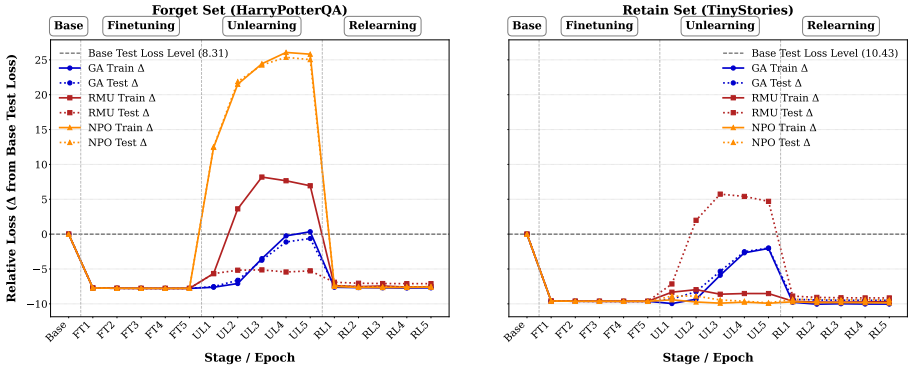

Figure 2: Overall Loss Dynamics (TinyStories-1M). Relative loss (Δ from Base Test Loss) on Forget (left) and Retain (right) data. Lower is better. Train loss (solid) / Test loss (dotted) shown for GA (blue), RMU (red), and NPO (orange). Unlearning phase shows varying selectivity: NPO preserves Retain utility best, GA worst. Relearning rapidly reverses forgetting for all methods.

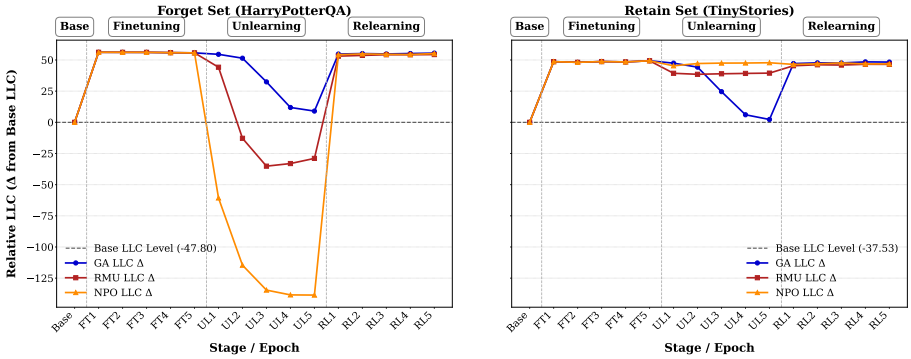

Figure 3: Overall Geometric Complexity (TinyStories-1M). Relative global drLLC (Δ from Base LLC) on Forget (left) and Retain (right) train data. Lower LLC indicates higher degeneracy. Unlearning decreases Forget complexity; Relearning restores it, mirroring inverse loss trends.

relative selectivity (Table 4).

Finetuning Establishes Adapted Layer Geometry Adapting the pretrained model (Base → FT5) generally increases layer-refined LLCs across L1-L7 (Figs. 5, 6), reducing degeneracy. While deeper layers often develop slightly higher complexity, the inter-layer variance (σ) typically stabilizes or decreases compared to the Base state (Table 2, FT rows), suggesting layers reach a coordinated geometric state adapted to the FT data prior to unlearning.

Unlearning Imprints Distinct and Scalable Geometric Signatures The unlearning phase (FT5 → UL5) shows sharp contrasts, particularly between GA and RMU. **GA** tends to induce a relatively *uniform* increase in degeneracy (LLC decrease) across layers for *both* Forget and Retain data (Figs. 5, 6). This is reflected in its moderate σ (Table 2), suggesting a more homogeneous impact consistent with its known tendency for degrading utility. In contrast, **RMU** exhibits a highly *selective* signature. On the Forget set, it forces layers into a highly consistent degenerate state, significantly minimizing σ compared to GA across scales. On the Retain set, RMU often increases σ, particularly at the 1M scale (Table 4). Visual inspection (Fig. 5, right) shows pre-injection layers (L1-L2) becoming more degenerate while post-injection layers (L4-L7) maintain higher complexity, highlighting RMU's selective preservation of utility-related geometry. **NPO** (1M model, Fig. 9) shows relatively uniform LLC drops similar to GA, but less severe on the Retain set, aligning with its better utility preservation observed globally. Model size amplifies absolute σ values (Table 2), but the relative geometric differences between GA and RMU persist, indicating these are fundamental signatures. The log(GSI) metric (Table 4) quantifies RMU's superior selectivity at 1M, which surprisingly diminishes at larger scales where GA shows marginally better relative selectivity despite its overall impact on utility.

Relearning Shows Partial Recovery and Lasting Structural Reorganization The relearning phase (UL5 → RL5) shows incomplete and non-uniform geometric recovery. Layer LLCs increase but rarely reach FT5 levels (Figs. 5, 6). Forget set variance σ typically increases as layers recover unevenly (Table 2). Notably, the unlearning/relearning cycle

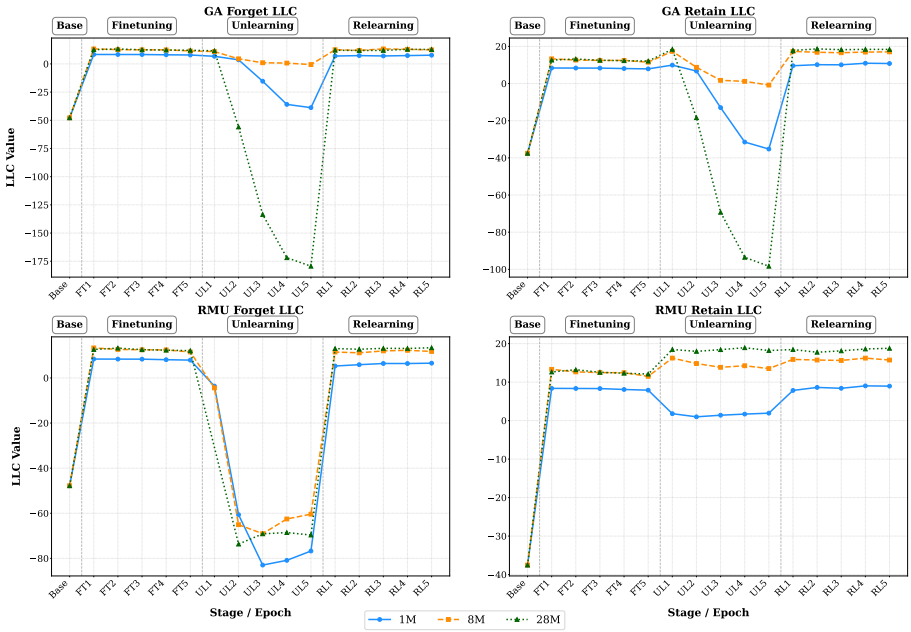

Figure 4: Scaling of Global Geometric Dynamics. Relative global drLLC across model scales (1M, 8M, 28M) for GA and RMU on Forget (left) / Retain (right) data through FT, UL, RL stages. GA induces greater geometric degeneracy on the Retain set as model size increases. RMU is more effective at maintaining lower degeneracy for Retain geometry at larger scales.

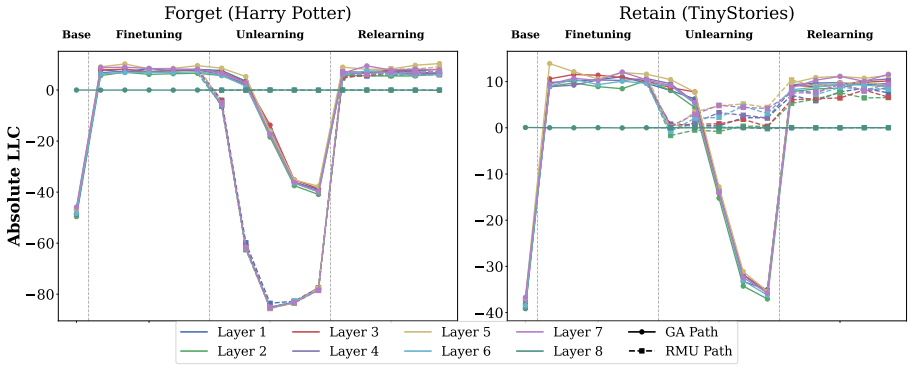

Figure 5: Layer-specific LLC Trajectories (TinyStories-1M). GA impacts layers relatively uniformly, resulting in moderate σ. RMU shows selectivity, forcing high uniformity (low σ) on $\mathcal{D}_{\text{forget}}$ but increasing differentiation (high σ) on $\mathcal{D}_{\text{retain}}$ during UL.

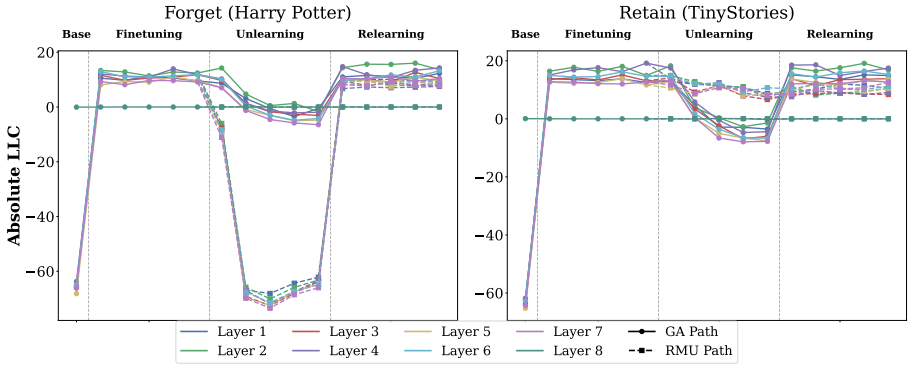

Figure 6: Layer-specific LLC Trajectories (TinyStories-8M). Increased model scale leads to higher LLC values and greater inter-layer differentiation compared to Fig. 5, yet the characteristic geometric signatures for GA (uniform) and RMU (selective) persist.

Dataset	Stage	1M GA	1M RMU	1M NPO	8M GA	8M RMU	28M GA	28M RMU
Forget	Base	1.385*	1.385*	1.385*	1.342*	1.342*	3.087*	3.087*
	FT	0.942*	0.942*	0.942*	1.289*	1.289*	3.337*	3.337*
	UL	1.021	**0.544**	1.050	1.997	**1.391**	3.527	**3.357**
	RL	1.105	**0.753**	0.638	1.825	**0.739**	3.621	**2.792**
Retain	Base	1.076*	1.076*	1.076*	1.206*	1.206*	3.194*	3.194*
	FT	0.917*	0.917*	0.917*	1.866*	1.866*	2.928*	2.928*
	UL	**0.790**	1.392	0.772	2.012	**0.874**	2.888	**2.735**
	RL	**0.731**	1.107	0.531	2.045	**0.922**	3.070	**2.186**

Table 2: Inter-Layer Variance (σ). Lower σ implies more uniform layer geometry. **Bold** and highlighted cells indicate the method with lower variance for each comparison (Dataset, Stage, Size). Values marked with * are shared as they correspond to the Base or FT stages before unlearning methods diverge. RMU consistently achieves lower Forget σ during UL/RL than GA across scales. NPO (1M) shows low RL variance.

Model	Dataset	GA (ρ)	RMU (ρ)	NPO (ρ)
1M	Forget	0.714	0.536	**0.857**
	Retain	**0.643**	0.571	0.286
8M	Forget	**0.682**	0.495	—
	Retain	**0.611**	0.525	—
28M	Forget	**0.647**	0.482	—
	Retain	**0.598**	0.503	—

Table 3: Layer Ranking Stability (ρ). Higher ρ indicates better preservation of layer importance hierarchy (FT5 vs RL5). **Bold** and highlighted cells indicate higher correlation. GA consistently shows higher stability than RMU. Stability decreases with model size. NPO exhibits high Forget but low Retain stability.

Metric	1M Model GA	1M Model RMU	1M Model NPO	8M Model GA	8M Model RMU	28M Model GA	28M Model RMU
log(GSI)	−0.231	**0.967**	−0.280	**-0.363**	−0.835	−0.069	**-0.074**

Table 4: Log Geometric Selectivity Index. Measures relative change in Retain vs. Forget inter-layer variance during unlearning ($\log(\text{GSI}) > 0$ indicates higher selectivity). **Bold** and highlighted cells indicate the method with higher log(GSI). RMU shows strong selectivity at 1M, but this diminishes significantly at larger scales, unlike GA/NPO (1M) which show negative selectivity.

permanently alters the relative geometric importance of layers, as shown by Layer Ranking Stability (ρ, Table 3). GA consistently maintains higher stability (ρ) than RMU across scales, suggesting its uniform geometry leads to a state closer to the original hierarchy post-relearning. However, stability decreases notably with model size for both methods, indicating that the layer specializations in larger models are harder to restore after geometric disruption. NPO (1M) shows high Forget stability but very low Retain stability, suggesting it strongly preserves the forget-related hierarchy while potentially disrupting the retain-related one during the UL/RL cycle.

In summary, layer-wise rLLC analysis reveals unlearning as a layer-heterogeneous process. Metrics like σ quantify selectivity, differentiating RMU's targeted approach from GA's uniform degradation, while ρ measures structural persistence, revealing lasting changes in the model's internal functional hierarchy, especially at larger scales.

4.4. RMU Geometric Fingerprint

Beyond comparing aggregate geometric properties, refined LLC analysis can yield diagnostic insights into specific unlearning mechanisms. We investigated whether RMU, characterized by its localized activation intervention at a layer L_{noise} leaves a detectable geometric signature. As motivated theoretically (Sec. 3.3 and Lemma 3.1), such an intervention is expected to induce a localized discontinuity in the layer-wise LLC profile near L_{noise}.

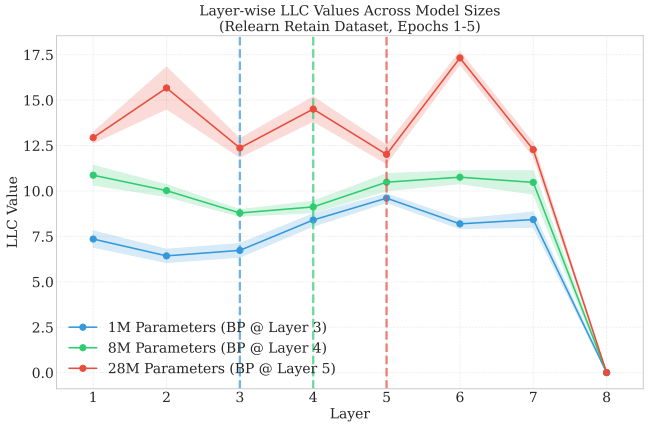

Figure 7: RMU Fingerprint: Positive LLC Jump. Layer drLLCs on $\mathcal{D}_{\text{retain}}$ during RL phase for RMU ($L_{\text{noise}} \in \{3, 4, 5\}$ for 1M, 8M, 28M models). Error bands are std. error over epochs. A distinct positive jump consistently appears between L_{noise} and $L_{\text{noise}} + 1$. Dashed lines mark $k^\star$ found by Alg. 1, correctly identifying L_{noise} for each model.

Our experiments confirm this, revealing a reliable signature across the 1M, 8M, and 28M TinyStories models, despite varying the injection layer L_{noise} (L3, L4, L5 respectively). Figure 7 plots the layer-wise drLLCs measured on the **retain dataset** during the **relearning phase**—conditions theoretically favoring a high signal-to-noise ratio (Appendix B.4).

A consistent and prominent *positive* jump in LLC emerges specifically between the true injection layer L_{noise} and the subsequent layer $L_{\text{noise}} + 1$ for each model size. This layer-specific positive jump validates the detection method proposed in Algorithm 1. When applied to the empirical LLC profiles from the (relearn, retain) measurements, the algorithm achieved **100% accuracy** in identifying the correct $L_{\text{noise}} \in \{3, 4, 5\}$ for the corresponding model size across all runs.

This finding demonstrates that rLLC analysis provides powerful **diagnostic capabilities**. The distinct geometric fingerprint—the positive LLC jump originating from the injection layer allows the targeted layer of the RMU intervention to be identified post-hoc. This offers an **auditable signature**, potentially useful for verifying if RMU was applied correctly. It also presents a potential vulnerability, as adversaries aware of this signature could use it to identify the intervention site and focus subsequent relearning attacks more effectively.

5. Conclusion and Future Work

This work demonstrated the utility of rLLCs from SLT as a quantitative tool to analyze the internal geometric impact of unlearning. Our experiments revealed distinct geometric signatures for different unlearning methods (GA, RMU, NPO) on TinyStories models, quantifiable via metrics like inter-layer variance and layer ranking stability. We also identified a diagnostic fingerprint for RMU, enabling post-hoc identification of its intervention layer. These findings establish rLLCs as a principled approach for evaluating and comparing unlearning methods. Our analysis was conducted on models up to 28M parameters. A key challenge is the computational cost of SGLD-based LLC estimation, which makes scaling to billion-parameter models difficult. Future work should explore more efficient estimation techniques and investigate how these geometric signatures evolve in much larger models.

References

Barez, F., Fu, T., Prabhu, A., Casper, S., Sanyal, A., Bibi, A., O'Gara, A., Kirk, R., Bucknall, B., Fist, T., et al. Open problems in machine unlearning for ai safety, 2025.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Eldan, R. and Li, Y. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Eldan, R. and Russinovich, M. Who's harry potter? approximate unlearning in llms, 2023.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Farrell, E., Lau, Y.-T., and Conmy, A. Applying sparse autoencoders to unlearn knowledge in language models. In *NeurIPS 2024 Safe Generative AI Workshop*, 2024. URL https://arxiv.org/abs/2410.19278.

Hayes, J., Shumailov, I., Triantafillou, E., Khalifa, A., and Papernot, N. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy, 2024.

Hironaka, H. Resolution of singularities of an algebraic variety over a field of characteristic zero: Ii. *Annals of Mathematics*, 79(2):205–326, 1964.

Hu, S., Fu, Y., Wu, Z. S., and Smith, V. Jogging the memory of unlearned llms through targeted relearning attacks, 2024. URL https://arxiv.org/abs/2406.13356.

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, 2023.

Lau, E., Furman, Z., Wang, G., Murfet, D., and Wei, S. The Local Learning Coefficient: A Singularity-Aware Complexity Measure. *arXiv preprint arXiv:2308.12108*, 2023.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L., et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning*, 2024a.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024b.

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C. Y., Xu, X., Li, H., Varshney, K. R., Bansal, M., Koyejo, S., and Liu, Y. Rethinking machine unlearning for large language models, 2024. URL https://arxiv.org/abs/2402.08787.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. Tofu: A task of fictitious unlearning for llms, 2024.

Martens, J. and Grosse, R. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417. PMLR, 2015.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741, 2023.

Schraudolph, N. N. Fast curvature matrix-vector products for second-order gradient descent. In *Neural Information Processing Systems*, pp. 471–478, 2002.

Shi, W., Lee, J., Huang, Y., Malladi, S., Zhao, J., Holtzman, A., Liu, D., Zettlemoyer, L., Smith, N. A., and Zhang, C. MUSE: Machine unlearning six-way evaluation for language models. In *Conference on Language Modeling Research*, July 2024. URL https://openreview.net/forum?id=gAfnQ8o9Cq.

Thaker, P., Hu, S., Kale, N., Maurya, Y., Wu, Z. S., and Smith, V. Position: Llm unlearning benchmarks are weak measures of progress. *arXiv preprint arXiv:2410.02879*, 2024.

van Wingerden, S., Hoogland, J., Wang, G., and Zhou, W. Devinterp. https://github.com/timaeus-research/devinterp, 2024.

Wang, G., Hoogland, J., van Wingerden, S., Furman, Z., and Murfet, D. Differentiation and specialization of attention heads via the refined local learning coefficient, 2024. URL https://arxiv.org/abs/2410.02984.

Watanabe, S. *Algebraic geometry and statistical learning theory*. Cambridge University Press, 2009.

Watanabe, S. A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14 (1):867–897, 2013.

Wei, S., Murfet, D., Gong, M., Li, H., Gell-Redman, J., and Quella, T. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Yao, J., Chien, E., Du, M., Niu, X., Wang, T., Cheng, Z., and Yue, X. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*, 2024.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

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

Łucki, J., Wei, B., Huang, Y., Henderson, P., Tramèr, F., and Rando, J. An adversarial perspective on machine unlearning for ai safety. In *Red Teaming GenAI Workshop at NeurIPS 2024*, 2024.

A. Unlearning Method Descriptions

This section provides brief descriptions of the unlearning algorithms applied or discussed in this work.

Gradient Ascent (GA). GA directly maximizes the loss function (e.g., cross-entropy) specifically on the forget dataset ($\mathcal{D}_{\text{forget}}$) (Maini et al., 2024; Jang et al., 2023). This gradient is used to update model parameters, pushing the model's distribution away from the unwanted data. While simple, it often lacks mechanisms to prevent damage to general capabilities, frequently resulting in utility loss.

Negative Preference Optimization (NPO). NPO reframes unlearning within a preference learning paradigm (Zhang et al., 2024), adapting the Direct Preference Optimization (Rafailov et al., 2023) objective. It treats examples from $\mathcal{D}_{\text{forget}}$ as implicitly 'rejected' completions relative to a reference policy (often the original finetuned model). The optimization encourages lowering the probability of forget examples while maintaining similarity to the reference policy. Implementations may use KL Divergence Minimization on benign data ($\mathcal{D}_{\text{retain}}$) to further mitigate utility loss.

Representation Misdirection for Unlearning (RMU). RMU injects noise into the hidden representations at specific, targeted layers during the forward pass only when processing inputs related to the forget set (Li et al., 2024a). This disruption aims to prevent the model from effectively utilizing information necessary to produce the undesired output, without directly modifying weights based on forget loss gradients. Effectiveness depends on the choice of layer(s) and noise characteristics.

A. Additional Background on Local Learning Coefficients

This appendix provides further theoretical background on the Local Learning Coefficient (LLC) and its refined variants, drawing heavily from Singular Learning Theory (SLT) (Watanabe, 2009) and the exposition in Lau et al. (2023).

Complexity Beyond Parameter Count Traditional measures of model complexity, such as parameter count, often fail to capture the effective complexity of deep neural networks (DNNs) (Lau et al., 2023; Zhang et al., 2016). DNNs exhibit significant parameter redundancies and degeneracies, making them *singular* statistical models where the mapping from parameters to functions is not one-to-one, and the Fisher information matrix may be rank-deficient (Watanabe, 2009; Wei et al., 2022). SLT provides a framework to analyze such models by studying the geometry of their loss landscapes.

Geometric Intuition: Volume Scaling near Minima A key insight from SLT relates model complexity to the geometry near minima of the population loss function $L(w) = \mathbb{E}_{(x,y)\sim q}[\ell(f(x;w),y)]$, where q is the true data distribution and ℓ is the per-sample loss. Consider a local minimum w^*. We define a *neighborhood* $B(w^*)$ as a small closed ball in the parameter space centered on w^* where $L(w^*)$ is the minimum loss value (i.e., for all $w \in B(w^*)$, $L(w) \geq L(w^*)$). Given a small tolerance $\epsilon > 0$, we then consider the volume of the subset of parameters within this ball whose loss is close to the minimum:

$$V(\epsilon; w^*) := \text{Vol}\{w \in B(w^*) \mid L(w) - L(w^*) < \epsilon\} = \int_{w\in B(w^*), L(w)-L(w^*)<\epsilon} dw \tag{4}$$

For regular models where the loss is locally quadratic ($L(w) - L(w^*) \approx \frac{1}{2}(w-w^*)^T H(w-w^*)$), this volume scales as $V(\epsilon; w^*) \propto \epsilon^{d/2}$, where d is the parameter dimension (Lau et al., 2023). The exponent $d/2$ reflects the number of free parameters.

However, for singular models like DNNs, the geometry near w^* is more complex. SLT shows, via resolution of singularities (Hironaka, 1964), that the volume scales according to a different law (Watanabe, 2009; Lau et al., 2023):

$$V(\epsilon; w^*) = c\epsilon^{\lambda(w^*)}(-\log\epsilon)^{m(w^*)-1} + o(\epsilon^{\lambda(w^*)}(-\log\epsilon)^{m(w^*)-1}) \quad \text{as } \epsilon \to 0^+ \tag{5}$$

where $c > 0$ is a constant, $\lambda(w^*)$ is a unique rational number called the **Local Learning Coefficient (LLC)**, and $m(w^*)$ is a positive integer called the local multiplicity.

LLC Definition and Interpretation The LLC $\lambda(w^*)$ is formally defined as the leading exponent in the asymptotic volume scaling law (Eq. 5). When the multiplicity $m(w^*) = 1$ (a common case), the formula simplifies to $V(\epsilon; w^*) \propto \epsilon^{\lambda(w^*)}$. Thus, the LLC quantifies how the volume of the low-loss region expands as the tolerance ϵ increases.

A lower value of $\lambda(w^*)$ implies that the volume grows more slowly as ϵ increases (since $\epsilon < 1$), meaning there is relatively more volume concentrated very close to the minimum $L(w^*)$. This corresponds to higher **degeneracy**: there are more directions or ways to vary the parameters w near w^* without significantly increasing the loss $L(w)$. Intuitively, a lower LLC signifies a flatter or simpler local geometry around w^*. From an information-theoretic perspective, $\lambda(w^*)$ relates to the number of bits needed to specify a parameter achieving loss within ϵ of the minimum (Lau et al., 2023). In regular models, $\lambda(w^*) = d/2$, linking back to the parameter count. In singular models, $\lambda(w^*)$ is typically much smaller than $d/2$, reflecting the model's degeneracy. The LLC is also known as the Real Log Canonical Threshold (RLCT) in algebraic geometry (Watanabe, 2009).

Estimation via Local Free Energy While the volume scaling definition provides theoretical intuition, practical estimation relies on the connection between LLC and the **local free energy** F_n. The free energy, or negative log marginal likelihood, in Bayesian statistics has an asymptotic expansion related to the (global) learning coefficient. Analogously, the free energy calculated over a local region $B_\gamma(w^*)$ around w^* (or using a localizing prior) has an asymptotic expansion involving the LLC (Watanabe, 2009; Lau et al., 2023):

$$F_n(B_\gamma(w^*)) = -\log \int_{B_\gamma(w^*)} \exp\{-nL_n(w)\}\phi(w)dw \approx nL_n(w^*) + \lambda(w^*)\log n + O(\log\log n) \tag{6}$$

where $L_n(w)$ is the empirical loss over n samples and $\phi(w)$ is a prior. This suggests an *idealized* estimator by solving for $\lambda(w^*)$.

To make this practical, Lau et al. (2023) adapt the Widely Applicable Bayesian Information Criterion (WBIC) approach (Watanabe, 2013). Instead of integrating over a hard region $B_\gamma(w^*)$, they use a soft localization via a Gaussian prior centered at w^*, $\phi_\gamma(w - w^*) \propto \exp(-\frac{\gamma}{2}\|w - w^*\|^2)$, leading to the Gibbs posterior $\pi(w|w^*, \beta, \gamma)$ defined in the main text. The local free energy associated with this posterior is approximated by the expectation term $E_{w|w^*,\beta^*,\gamma}[nL_n(w)]$ with a specific temperature $\beta^* = 1/\log n$. Substituting this approximation into the rearranged asymptotic form (Eq. 6) yields the practical LLC estimator used in this work (Eq. 1 in the main text):

$$\hat{\lambda}(w^*) = n\beta^*\left[\mathbb{E}_{w\sim\pi(w|w^*,\beta^*,\gamma)}[L_n(w)] - L_n(w^*)\right]$$

This estimator measures the expected increase in empirical loss under perturbation by samples from the localized Gibbs posterior, scaled appropriately. A smaller value indicates that perturbations likely under the posterior do not significantly increase the loss, consistent with higher local degeneracy.

Refined LLCs: Formal Definitions Building on the concept of LLC, refined variants allow targeted analysis (Wang et al., 2024):

Weight-refined LLC (wrLLC): $\lambda(w^*; V)$. Consider a parameter decomposition $w = (u, v)$ where $v \in V \subseteq W$ represents the parameters of interest (e.g., a layer) and $u \in U = W \setminus V$ are the fixed parameters ($u = u^*$). The wrLLC measures the complexity associated with V by considering the loss function restricted to this subspace, $\ell_{\text{restr}}(v) = \ell(u^*, v)$. Formally, it's defined via the volume scaling exponent analogous to Eq. 5 but calculated using $\ell_{\text{restr}}(v)$ in a neighborhood of v^*:

$$\lambda(w^*; V) = \lim_{\epsilon\to 0^+} \frac{-\log \text{Vol}\{v \mid \ell_{\text{restr}}(v) - \ell_{\text{restr}}(v^*) < \epsilon\}}{-\log\epsilon} \tag{7}$$

The estimator $\hat{\lambda}(w^*; V)$ is obtained by modifying the Gibbs posterior in Eq. 1 to sample only v while keeping $u = u^*$, i.e., using $p(v; v^*, \beta, \gamma) \propto \exp(-n\beta L_n(u^*, v) - \frac{\gamma}{2}\|v - v^*\|^2)$.

Data-refined LLC (drLLC): $\lambda(w^*; q')$. This measures the geometric complexity relative to a potentially different data distribution q' (compared to the distribution q underlying the original loss L). Let $L'(w) = \mathbb{E}_{(x,y)\sim q'}[\ell(f(x; w), y)]$ be the population loss under q'. The drLLC is the scaling exponent defined analogously to Eq. 5 but using $L'(w)$ instead of $L(w)$:

$$\lambda(w^*; q') = \lim_{\epsilon\to 0^+} \frac{-\log \text{Vol}\{w \mid L'(w) - L'(w^*) < \epsilon\}}{-\log\epsilon} \tag{8}$$

The estimator $\hat{\lambda}(w^*; q')$ uses the empirical loss $L'_n(w)$ calculated on samples from q' within the expectation and the Gibbs posterior definition in Eq. 1. This allows assessing how the geometry around w^* appears specifically through the lens of the data distribution q'.

These refined LLCs, particularly when combined (e.g., layer-specific drLLCs), provide the foundation for the granular geometric analysis of unlearning presented in the main paper.

B. Theoretical Basis for RMU Fingerprint (§3.3)

Theorem B.1 (RMU-Induced Geometric Discontinuity)**.** *The localized intervention of Representation Misdirection (RMU) at a specific layer L_{noise} induces a quantifiable discontinuity in the layer-wise Local Learning Coefficient (LLC) profile at the interface between layer L_{noise} and $L_{noise}+1$. This signature is empirically detectable, allowing for post-hoc identification of the intervention layer.*

Derivation. This appendix provides theoretical justification for the geometric fingerprint induced by Representation Misdirection Unlearning (RMU) and the effectiveness of the proposed detection algorithm (Algorithm 1)The derivation proceeds by analyzing the effect of the intervention on the local geometry of the loss landscape, approximated by the Generalised Gauss-Newton (GGN) matrix, and then analyzing the consistency of an estimator based on this effect.

We consider a Transformer model f parameterized by weights $w \in \mathbb{R}^{d_{\text{tot}}}$, composed of L layer blocks $w = (w_1, \ldots, w_L)$. The model is trained by minimizing a population loss $\mathcal{L}(w; q) = \mathbb{E}_{(x,y)\sim q}[\ell(f(x; w), y)]$, where ℓ is a per-token loss function (e.g., cross-entropy) and q is a data distribution. Near a parameter point w_*, the geometric complexity associated with a parameter subset $V \subseteq \mathbb{R}^{d_{\text{tot}}}$ can be quantified by the weight-refined Real Log Canonical Threshold (RLCT), also known as the Local Learning Coefficient (LLC) (Watanabe, 2009; Lau et al., 2023):

$$\lambda(w_*; V) = \lim_{t\to 0^+} \frac{-\log \text{Vol}\{v \in V : \mathcal{L}(w_* + v; q) \le \mathcal{L}(w_*; q) + t\}}{-\log t}. \tag{9}$$

This quantity reflects the effective dimensionality or degeneracy of the loss landscape restricted to the parameters V; lower values indicate higher degeneracy.

B.1. GGN Approximation and RMU Perturbation

Let $L_{\text{noise}} \in \{1, \ldots, L\}$ be the specific layer where RMU injects a fixed steering vector v into the activations $h_{L_{\text{noise}}}$ for inputs x from the forget distribution q_{for}. We partition the model's weights accordingly: $w = (w_{\le L_{\text{noise}}}, w_{>L_{\text{noise}}})$, representing parameters up to and including layer L_{noise}, and parameters strictly after layer L_{noise}, respectively.

We approximate the local curvature of the loss landscape using the Generalised Gauss–Newton (GGN) matrix (Martens & Grosse, 2015; Schraudolph, 2002). The GGN matrix block corresponding to the downstream parameters $w_{>L_{\text{noise}}}$, computed with respect to a data distribution q and parameters w, is defined as:

$$G_{>L_{\text{noise}}}(q; w) := \mathbb{E}_{(x,y)\sim q}\Big[J_{>L}(x; w) H_\ell\big(f(x; w), y\big) J_{>L}(x; w)^\top\Big], \tag{10}$$

where $a_{L_{\text{noise}}}(x; w)$ are the activations output by layer L_{noise} (input to the downstream network $g_{>L_{\text{noise}}}$), $J_{>L}(x; w) = \partial g_{>L_{\text{noise}}} / \partial w_{>L_{\text{noise}}}(a_{L_{\text{noise}}}(x; w))$, and H_ℓ is the Hessian of the per-token loss ℓ with respect to the model's output logits. For the standard softmax cross-entropy loss, H_ℓ has rank at most 1 (Martens & Grosse, 2015).

Let w_{ul} denote the model parameters during the RMU unlearning phase. Define the baseline GGN on the forget distribution q_{for} using the standard activations at w_{ul}:

$$G^{\text{base}}_{>L_{\text{noise}}} := G_{>L_{\text{noise}}}(q_{\text{for}}; w_{\text{ul}}). \tag{11}$$

Assumption A (Linear Response to Small Shift). During RMU unlearning on q_{for}, the activations are perturbed: $a^{\text{for}}_{L_{\text{noise}}}(x) = a_{L_{\text{noise}}}(x; w_{\text{ul}}) + v$, where $\|v\|_2 = \alpha$. We assume a linear-response regime where α is sufficiently small compared to the typical activation norm $\|a_{L_{\text{noise}}}(x; w_{\text{ul}})\|_2$. In this regime, the Jacobian $J_{>L}(x; w_{\text{ul}})$ changes negligibly, and the primary effect on the GGN matrix arises from the change in the expectation in Eq. (10) due to the perturbed activations $a^{\text{for}}_{L_{\text{noise}}}$ influencing the loss Hessian term H_ℓ.

Let $G^{\text{for}}_{>L_{\text{noise}}}$ be the GGN matrix computed on q_{for} using the perturbed activations $a^{\text{for}}_{L_{\text{noise}}}(x)$. Under Assumption A, the relationship is approximated by:

$$G^{\text{for}}_{>L_{\text{noise}}} \approx G^{\text{base}}_{>L_{\text{noise}}} + R, \tag{12}$$

where the perturbation matrix R captures the average effect of the shift v. Since this effect enters through the low-rank (rank ≤ 1) loss Hessian H_ℓ, the resulting perturbation R also satisfies $\text{rank}(R) \leq 1$. The GGN block corresponding to upstream parameters $w_{\leq L_{\text{noise}}}$ is assumed to be unaffected to first order by this downstream perturbation.

B.2. Impact on Local Learning Coefficients

Singular Learning Theory (Watanabe, 2009) relates the RLCT λ to the local geometry. For loss landscapes locally approximated by a quadratic form defined by the GGN matrix at a point w_*, the RLCT associated with parameters V is related to the dimension of the GGN kernel restricted to V: $\lambda(w_*; V) \approx \frac{1}{2} \dim \ker G(q; w_*, V)$ (cf. Watanabe, 2009, Cor. 6.1.4).

Lemma B.2 (Interface Discontinuity Magnitude)**.** *Let $\ell^\star = L_{noise} - 1$ be the index of the layer immediately preceding the noise injection layer. Let $\lambda^{state}(\ell)$ denote the layer-wise LLC profile for the baseline (base) or RMU-perturbed (for) state, evaluated near w_{ul}. Under Assumption A, if the RMU perturbation R changes the rank of the downstream GGN block $G_{>L_{noise}}$ by exactly 1, then the magnitude of the discontinuity in the layer-wise LLC profile at the interface $\ell^\star$ changes by approximately $1/2$:*

$$\left|(\lambda^{for}(\ell^\star + 1) - \lambda^{for}(\ell^\star)) - (\lambda^{base}(\ell^\star + 1) - \lambda^{base}(\ell^\star))\right| \approx \frac{1}{2}.$$

Proof. Let $G^{\text{base}} = G^{\text{base}}_{>L_{\text{noise}}}$ and $G^{\text{for}} = G^{\text{for}}_{>L_{\text{noise}}}$. From Eq. (12), $G^{\text{for}} \approx G^{\text{base}} + R$ with $\text{rank}(R) \leq 1$. Let $n_{>L}$ be the dimension of the parameter space $V_{>L_{\text{noise}}}$. The rank-nullity theorem states $n_{>L} = \text{rank}(G) + \dim \ker G$. Let $\delta(\cdot)$ denote the change induced by the perturbation R. Then $\delta(\text{rank}(G)) + \delta(\dim \ker G) = 0$.

Since $\text{rank}(R) \leq 1$, the change in rank satisfies $|\delta(\text{rank}(G))| \leq 1$. Consequently, the change in kernel dimension also satisfies $|\delta(\dim \ker G)| \leq 1$. If we assume the perturbation causes the rank to change by exactly 1 (i.e., $|\delta(\text{rank}(G))| = 1$), then the kernel dimension must also change by exactly 1 (i.e., $|\delta(\dim \ker G)| = 1$).

The LLC for the downstream parameters is $\lambda_{>L_{\text{noise}}} \approx \frac{1}{2} \dim \ker G_{>L_{\text{noise}}}$. The change in this LLC is $\delta\lambda_{>L_{\text{noise}}} = \lambda^{\text{for}}_{>L_{\text{noise}}} - \lambda^{\text{base}}_{>L_{\text{noise}}} \approx \frac{1}{2}\delta(\dim \ker G)$. If the rank changes by 1, then $|\delta\lambda_{>L_{\text{noise}}}| \approx \frac{1}{2}$.

We associate the layer-wise LLC $\lambda(\ell)$ with the parameters primarily within that layer block. The downstream LLC $\lambda_{>L_{\text{noise}}}$ is most directly associated with $\lambda(\ell^\star + 1)$ (since $\ell^\star + 1 = L_{\text{noise}}$, the first affected layer). The upstream LLC $\lambda_{\leq L_{\text{noise}}}$, associated with $\lambda(\ell^\star)$, is assumed to be unaffected to first order by the downstream perturbation R, thus $\lambda^{\text{for}}(\ell^\star) \approx \lambda^{\text{base}}(\ell^\star)$.

Let $D^{\text{state}} = \lambda^{\text{state}}(\ell^\star + 1) - \lambda^{\text{state}}(\ell^\star)$ denote the discontinuity at the interface $\ell^\star$ in a given state (base or for). The change in this discontinuity is:

$$\begin{aligned}
\Delta D &= D^{\text{for}} - D^{\text{base}} \\
&= (\lambda^{\text{for}}(\ell^\star + 1) - \lambda^{\text{for}}(\ell^\star)) - (\lambda^{\text{base}}(\ell^\star + 1) - \lambda^{\text{base}}(\ell^\star)) \\
&\approx (\lambda^{\text{base}}(\ell^\star + 1) + \delta\lambda_{>L_{\text{noise}}}) - \lambda^{\text{base}}(\ell^\star) - (\lambda^{\text{base}}(\ell^\star + 1) - \lambda^{\text{base}}(\ell^\star)) \\
&= \delta\lambda_{>L_{\text{noise}}}.
\end{aligned}$$

Therefore, the magnitude of the change in the discontinuity is $|\Delta D| \approx |\delta\lambda_{>L_{\text{noise}}}| \approx \frac{1}{2}$, given the assumption that the perturbation changes the GGN rank by exactly 1. This confirms that the RMU intervention induces a significant, localized change in the geometric profile at the interface $\ell^\star$. □

Empirical Observation and Estimator Target. While Lemma B.2 predicts a magnitude change of $\approx 1/2$ at the interface $\ell^\star = L_{\text{noise}} - 1$, it does not determine the sign. Our empirical results (§3.3, Figure 7) consistently show that this discontinuity manifests as a *positive* jump when calculated as $\Delta(L_{\text{noise}}) = \text{LLC}(L_{\text{noise}} + 1) - \text{LLC}(L_{\text{noise}}) > 0$, under (relearn, retain) measurement conditions. This suggests that the net effect in our experiments is an increase in degeneracy (lower LLC) for the injection layer L_{noise} relative to the subsequent layer. Algorithm 1 specifically targets this empirically reliable positive jump, finding the index $k^\star = \arg\max_i \max(0, \Delta(i))$, which correctly identifies $k^\star = L_{\text{noise}}$.

B.3. Consistency of the Positive-Jump Estimator

The algorithm's success relies on the mean of the targeted positive jump statistic being significantly larger at the true injection layer index ($i = L_{\text{noise}}$) than at other indices.

Let $\Lambda_{\ell t}$ be the empirically estimated LLC for layer ℓ in epoch t of the relearning phase, measured on $\mathcal{D}_{\text{retain}}$. Define the per-epoch positive jump statistic at index i (representing the jump from layer i to layer $i+1$) as:

$$Z_{it} = [\Lambda_{i+1,t} - \Lambda_{i,t}]_+ = \max(0, \Lambda_{i+1,t} - \Lambda_{i,t}). \tag{13}$$

Let $\mu_i = \mathbb{E}[Z_{it}]$ be the true mean positive jump associated with index i. We assume measurements across epochs $t = 1, \ldots, T$ are approximately independent.

Algorithm 1 computes the sample mean positive jump:

$$J_i^+ = \frac{1}{T} \sum_{t=1}^{T} Z_{it}. \tag{14}$$

Let $\ell^\star = L_{\text{noise}}$ be the true index where the mean positive jump is maximized ($\mu_{\ell^\star} > \mu_i$ for $i \neq \ell^\star$, based on empirical findings). Let $k^\star = \arg\max_{i \in \{1,\ldots,L-1\}} J_i^+$ be the index estimated by the algorithm. Define the minimum mean gap $\Delta = \mu_{\ell^\star} - \max_{i \neq \ell^\star} \mu_i > 0$.

Assumption B (Boundedness/Sub-Gaussianity of Z_{it}). LLC estimates ($\Lambda_{\ell t}$) derived from finite SGLD runs on bounded data domains are typically bounded random variables, although their variance can be substantial. Differences and positive parts Z_{it} of bounded variables are also bounded. Bounded random variables are sub-Gaussian. We assume Z_{it} are sub-Gaussian with a uniform variance proxy σ_L^2 across indices i and epochs t.

Lemma B.3 (Estimator Consistency)**.** *Under Assumption B and assuming independence across epochs t, the probability of the estimator $k^\star = \arg\max_i J_i^+$ failing to identify the true index $\ell^\star = L_{noise}$ is bounded by:*

$$\Pr[k^\star \neq \ell^\star] \leq 2(L-2) \exp\left(-\frac{T\Delta^2}{8\sigma_L^2}\right),$$

for $L \geq 3$.

Proof. We bound the probability that the empirical maximum J_k^+ occurs at an index $k \neq \ell^\star$. The event $\{k^\star \neq \ell^\star\}$ is the union of events $B_k = \{J_k^+ \geq J_{\ell^\star}^+\}$ for all $k \neq \ell^\star$. We use Hoeffding's inequality for the sample mean J_i^+ of T independent sub-Gaussian variables Z_{it}, which states $\Pr(|J_i^+ - \mu_i| \geq \epsilon) \leq 2\exp(-T\epsilon^2/(2\sigma_{\text{sub}}^2))$, where σ_{sub}^2 is the sub-Gaussian variance parameter (here bounded by σ_L^2).

Consider a specific $k \neq \ell^\star$. Let $\epsilon = \Delta/2$. Note that $\epsilon \leq (\mu_{\ell^\star} - \mu_k)/2$. The event $B_k = \{J_k^+ \geq J_{\ell^\star}^+\}$ implies that it cannot be the case that both $J_k^+ < \mu_k + \epsilon$ and $J_{\ell^\star}^+ > \mu_{\ell^\star} - \epsilon$ hold simultaneously. Thus, B_k is contained within the union of two deviation events: $\{J_k^+ \geq \mu_k + \epsilon\} \cup \{J_{\ell^\star}^+ \leq \mu_{\ell^\star} - \epsilon\}$. Using the union bound and Hoeffding's inequality (applying the one-sided version $\Pr(J_i^+ - \mu_i \geq \epsilon) \leq \exp(-T\epsilon^2/(2\sigma_L^2))$ and $\Pr(J_i^+ - \mu_i \leq -\epsilon) \leq \exp(-T\epsilon^2/(2\sigma_L^2))$):

$$\begin{aligned}
\Pr(B_k) &\leq \Pr(J_k^+ \geq \mu_k + \epsilon) + \Pr(J_{\ell^\star}^+ \leq \mu_{\ell^\star} - \epsilon) \\
&\leq \exp(-T\epsilon^2/(2\sigma_L^2)) + \exp(-T\epsilon^2/(2\sigma_L^2)) \\
&= 2\exp(-T(\Delta/2)^2/(2\sigma_L^2)) = 2\exp(-T\Delta^2/(8\sigma_L^2)).
\end{aligned}$$

Finally, applying the union bound over all $L-2$ possible incorrect indices $k \neq \ell^\star$ (assuming $L \geq 3$):

$$\Pr[k^\star \neq \ell^\star] = \Pr\left(\bigcup_{k \neq \ell^\star} B_k\right) \leq \sum_{k \neq \ell^\star} \Pr(B_k) \leq (L-2) \times 2\exp\left(-\frac{T\Delta^2}{8\sigma_L^2}\right).$$

The theoretically predicted local discontinuity (ensuring Δ exists) combined with the empirically observed sign (ensuring $\Delta > 0$ for the J^+ statistic) supports the estimator's consistency. □

B.4. Rationale for Measurement Conditions

Measuring LLCs on the (*relearn, retain*) data split empirically provides the highest signal-to-noise ratio (Δ/σ_L) for detecting the RMU fingerprint via Algorithm 1. This choice is justified by several factors:

1. **Inactive RMU Hook:** During the relearning phase, the deterministic noise vector v is not added to activations. This removes a source of potentially high variance that exists during the unlearning phase (especially when estimating LLCs on $\mathcal{D}_{\text{forget}}$ where the hook is active), leading to more stable SGLD dynamics for LLC estimation.

2. **Lower Retain Data Variance:** The retain dataset $\mathcal{D}_{\text{retain}}$ typically represents knowledge the model preserves well. Consequently, the loss $\mathcal{L}(w; q_{\text{retain}})$ and its derivatives often exhibit lower variance around the current parameter state w_* compared to the forget dataset $\mathcal{D}_{\text{forget}}$. Lower loss variance generally translates to lower variance in the LLC estimates derived from SGLD sampling, as $\text{Var}[\hat{\lambda}] \propto \text{Var}[\mathcal{L}(w^{(t)})]$.

3. **Persistent Geometric Signature:** While relearning aims to restore forgotten knowledge, the geometric modifications induced by the RMU unlearning process do not vanish instantaneously. The localized discontinuity involving L_{noise} persists for at least the initial epochs of relearning, providing a sufficient window for measurement before the model's geometry fully reverts or adapts away from the unlearned state.

These factors combine to minimize the estimator variance proxy σ_L^2 while preserving the signal Δ, thus enhancing the reliability of detection according to Lemma B.3.

B.5. Summary

The theoretical framework presented relies on approximating the effect of RMU's activation perturbation using the GGN matrix under a linear-response assumption (Assumption A). This framework predicts a localized geometric discontinuity in the layer-wise LLC profile at the interface involving the noise injection layer L_{noise}, with a magnitude change of approximately $1/2$ (Lemma B.2). Empirically, this discontinuity consistently manifests as a positive jump $\text{LLC}(L_{\text{noise}}+1) - \text{LLC}(L_{\text{noise}}) > 0$ under (relearn, retain) measurement conditions. Algorithm 1 leverages this empirical regularity, combined with the theoretical localization, to accurately detect L_{noise}. The choice of measurement conditions maximizes the signal-to-noise ratio for this detection.

Caveat: The theoretical prediction of the discontinuity's magnitude ($\approx 1/2$) relies heavily on the linear-response approximation (Assumption A). This assumption may not hold accurately for the larger steering coefficients (α) often used in practice to achieve effective unlearning. However, the core theoretical insight—that RMU induces a geometrically localized change—remains plausible. The success of Algorithm 1 hinges only on correctly identifying the location of the largest positive jump, not on verifying its exact magnitude. This likely explains the algorithm's empirical robustness even when the precise conditions of the linear theory are violated.

□

C. Hyperparameters for TinyStories Experiments

Model We utilized the `roneneldan/TinyStories-1M`, `roneneldan/TinyStories-8M`, and `roneneldan/TinyStories-28M`models, which are 1 million, 8 million and 28 million parameter causal Transformers, accessed via the Hugging Face `transformers` library. Computations were performed on a CUDA-enabled A100 GPU, leveraging `bfloat16` precision.

Dataset and Preprocessing For each split, we selected a maximum of $N_{docs} = 50,000$ documents. This number was chosen to ensure a sufficiently representative data sample while maintaining computational tractability for repeated LLC estimations across multiple models and stages. Each document underwent tokenization using the model-specific tokenizer. The tokenizer's `pad_token` was explicitly set to match its `eos_token`. Tokenized sequences were segmented into fixed-length chunks of $L_{chunk} = 2048$ tokens, a common context window size facilitating consistent processing across models. Any final chunks shorter than 128 tokens were excluded to avoid potential instability or noise introduced by very short sequences during loss and LLC calculations. Shorter chunks (length $< L_{chunk}$) were right-padded using the `pad_token_id`.

Data Loading Processed chunks were loaded using a PyTorch `DataLoader`. Data shuffling was enabled for each epoch. A custom collation function (`collate_fn`) was implemented to stack the `input_ids` and `attention_mask` tensors from batch items and transfer the resulting tensors to the target compute device (`DEVICE`). The `DataLoader` utilized $N_{workers} \leq 4$ parallel worker processes (system permitting) for data fetching.

D. Additional TinyStories Experiments

This section provides supplementary figures for the layerwise LLC analysis conducted on the TinyStories models, complementing the discussion in Section 4.3.

Figure 8 displays the layer-specific data-refined LLC trajectories for the largest TinyStories model tested (28M parameters). Compared to the smaller models (Figs. 5, 6), we observe a significant increase in the absolute LLC values across most layers and stages, indicating overall lower geometric degeneracy (higher complexity) at this scale. The differentiation between layers, particularly during and after finetuning, also appears more pronounced. Despite the increased scale and complexity, the fundamental geometric signatures distinguishing GA and RMU persist: GA tends to impact layers more uniformly on both Forget and Retain datasets during unlearning, while RMU maintains a clearer distinction, showing a more pronounced drop and lower variance on the Forget set compared to the Retain set.

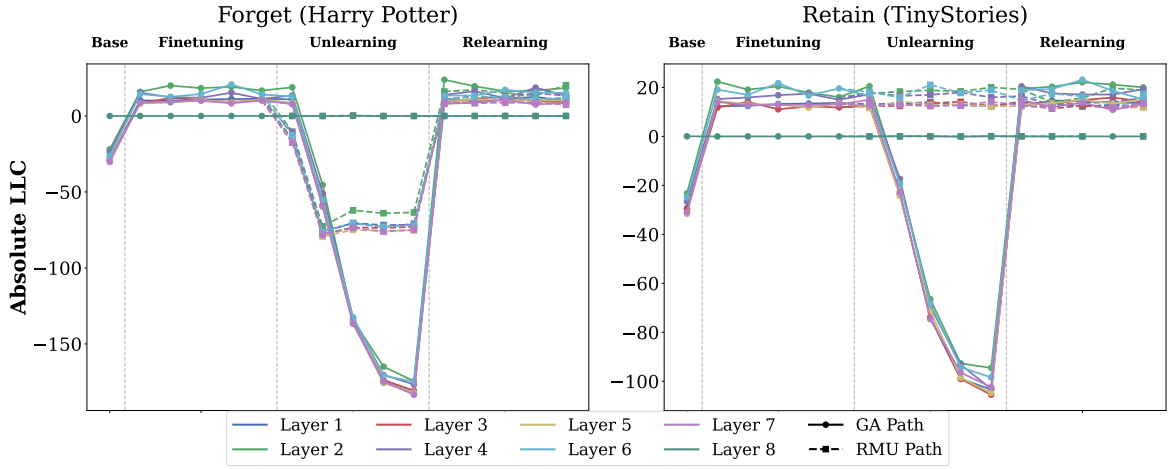

Figure 8: Layer-specific LLC Trajectories (TinyStories-28M). Layer LLCs on Forget (left) / Retain (right) sets for the 28M model. Compared to smaller scales (Figs. 5, 6), geometric complexity (LLC values) and inter-layer differentiation increase further. Despite this, the distinct geometric unlearning signatures for GA and RMU persist.

Figure 9 shows the layer-specific data-refined LLC trajectories for the NPO unlearning method applied to the 1M TinyStories model. During the unlearning phase (UL), NPO induces a sharp and relatively uniform decrease in LLC across layers L1-L7 with respect to the Forget data (left panel), achieving the most degenerate state observed among the methods (consistent with global LLC in Fig. 3). With respect to the Retain data (right panel), the LLC decrease during UL is much less severe compared to GA (Fig. 5), aligning with NPO's objective of preserving utility via its reference model mechanism. The relearning phase (RL) shows recovery, but the final layer LLC rankings differ significantly from the finetuned state, particularly for the Retain data (reflected in the low ρ value for NPO Retain in Table 3).

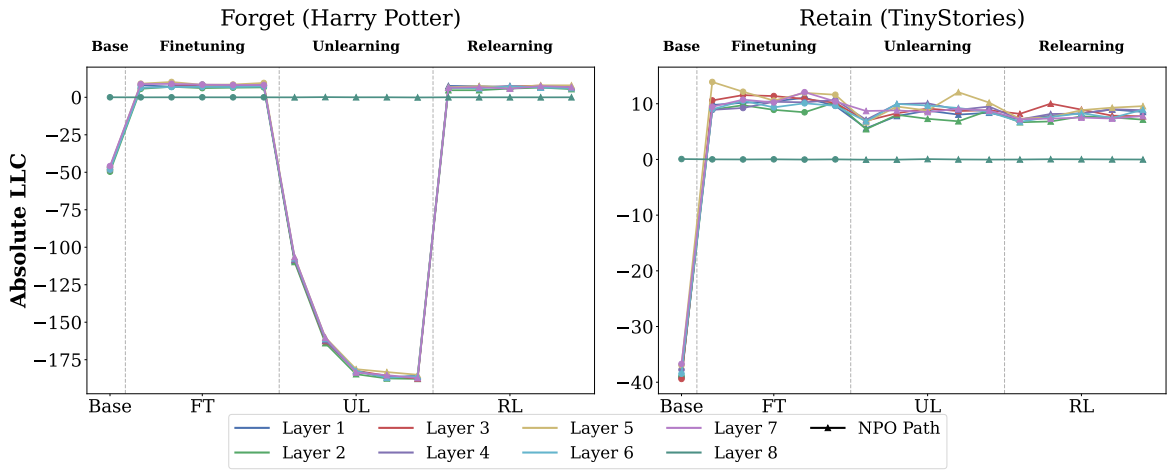

Figure 9: Layer-specific LLC Trajectories for NPO (TinyStories-1M). Layer LLCs on Forget (left) and Retain (right) sets through FT, UL, RL stages for the NPO method. Shows deep, uniform LLC drop on Forget data during UL, but better preservation of Retain geometry compared to GA.