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

Aashiq Muhamed, Virginia Smith
Carnegie Mellon University

Motivation: Beyond Black-Box Unlearning Evaluation

Goal: Remove targeted knowledge D_{forget} while preserving capabilities D_{retain}

Current evaluation: Black-box metrics (performance on test sets)

Limitations:

- Cannot distinguish superficial vs. structural forgetting
- No insight into internal mechanisms or why utility is lost
- Limited assurance beyond specific test examples

Our Solution: Probe internal loss landscape geometry using refined **Local Learning Coefficients (rLLCs)**

Key Questions:

Can geometry reveal method-specific unlearning signatures?
Can we measure internal selectivity patterns?
Can we predict utility preservation from geometric structure?

Refined Local Learning Coefficients

LLC from Singular Learning Theory

*Quantifies local effective dimensionality near w**

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

where $\ell_n(w)$ is the empirical loss over n samples. β and γ are the inverse temperature and localization strength.

Intuition: Lower λ ⇒ simpler local geometry (higher parameter degeneracy)

- **Weight-refined LLC (wrLLC):** Analysis restricted to parameter subset V
- **Data-refined LLC (drLLC):** Complexity relative to specific data distribution q'

SGLD Estimation: Monte-Carlo approximation over SGLD samples from Gibbs posterior

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

Experimental Setup

Models & Data

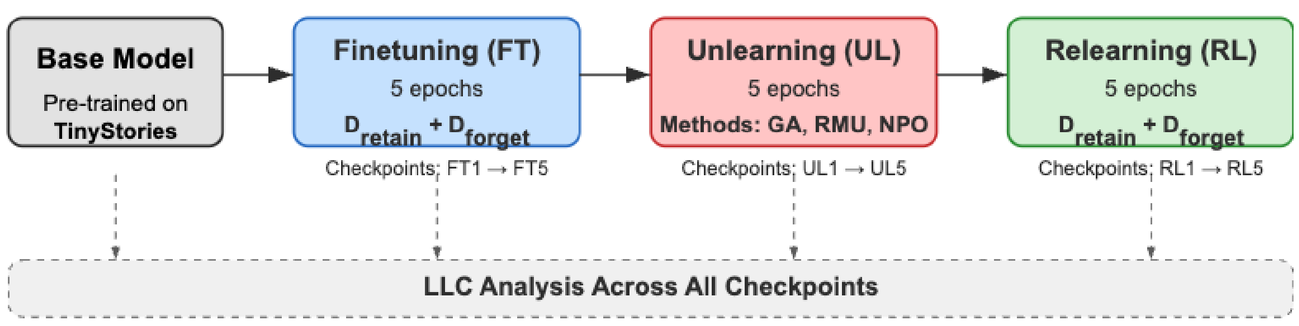
TinyStories: 1M, 8M, 28M parameters (8-layer Transformers)
D_{retain}**:** TinyStories, D_{forget}**:** Harry Potter

Unlearning Methods

- **Gradient Ascent (GA):** Direct loss maximization on forget data
- **Representation Misdirection (RMU):** Noise injection at specific layer
- **Negative Preference Optimization (NPO):** Preference learning with negative signal

Global and layer-wise drLLCs calculated for D_{forget} , D_{retain} at all checkpoints

New geometric metrics: Inter-layer variance σ, ranking stability ρ, selectivity index GSI



Geometric Signatures Reveal Unlearning Mechanisms

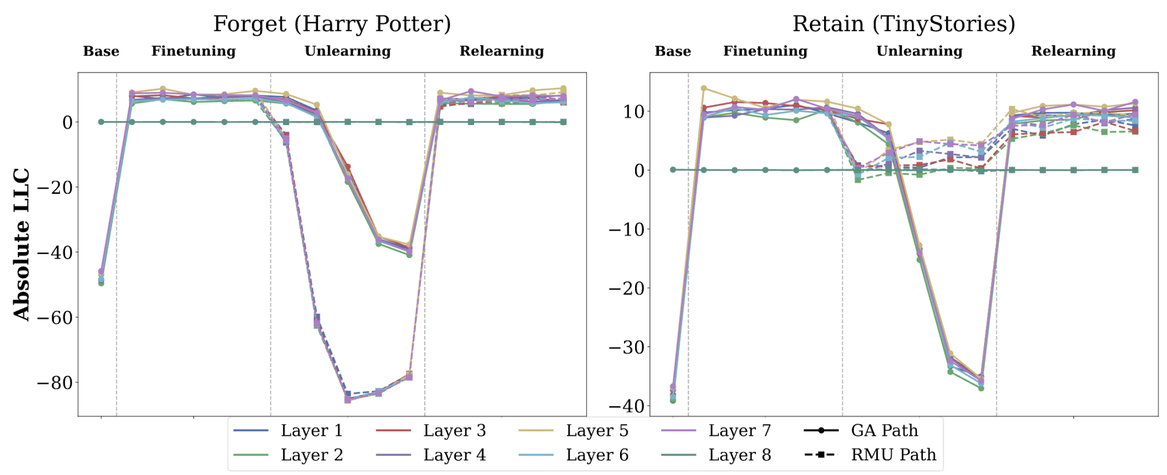
Layer-wise Analysis Reveals Method-Specific Patterns

GA: Uniform LLC decreases across all layers → non-selective geometric damage

RMU: Selective geometric intervention:

- Low inter-layer LLC variance σ_forget (forces uniform degeneracy on forget data)
- High σ_retain (preserves layer differentiation on retain data)

- Inter-layer variance σ quantifies geometric uniformity
- Model size amplifies geometric differences between methods

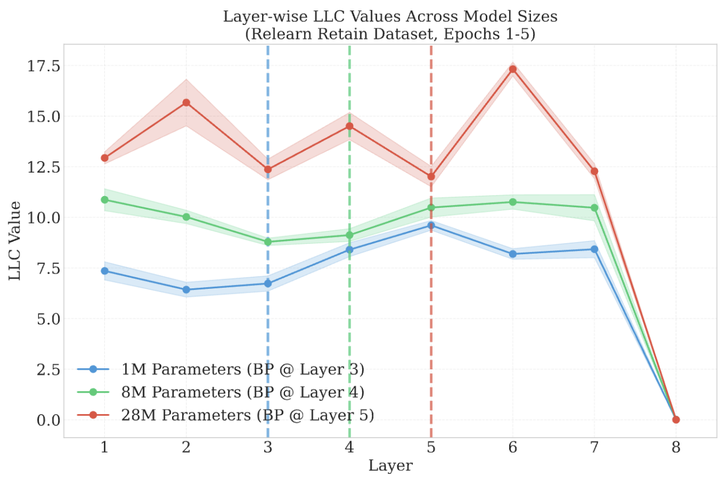


RMU's Geometric Fingerprint

RMU intervention at layer L_{noise} creates geometric discontinuity
Localized perturbation modifies downstream network geometry

Algorithm 1 Detecting the RMU Injection Block via Largest Positive Jump

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



⚠ **Can identify intervention layer using the positive LLC jump**

Quantifying Unlearning Quality

We introduce geometric metrics for unlearning evaluation.

Inter-Layer Variance (σ):

Lower σ_forget → uniform degeneracy (good)
Higher σ_retain → preserved differentiation (good)

Layer Ranking Stability (ρ): Structural preservation through unlearning-relearning cycle

Geometric Selectivity Index (GSI): Relative geometric selectivity

- ✓ **First geometric framework for unlearning evaluation**
- ✓ **Reveals method signatures invisible to black-box metrics**
- ✓ **Enables prediction of utility preservation from geometric patterns**