

ERASE: Precise Machine Unlearning via Geometry-Aware Gradient Shaping and Adaptive Loss Control

Itzik Waizman Shahar Katz Lior Wolf
Blavatnik School of Computer Science, Tel Aviv University
{itzikwaizman@mail, shaharkatz3@mail, wolf@cs}.tau.ac.il

Abstract

Large language models frequently memorize specific factual associations from their training data, necessitating post-hoc removal of targeted knowledge. However, standard gradient-based unlearning techniques treat all parameter directions uniformly, often leading to catastrophic forgetting of general capabilities. We introduce **ERASE** (**E**igenspace-**R**egulated **A**daptive **S**elective **E**rasure), a geometry-aware unlearning framework that combines three complementary mechanisms: (1) a *spectral filter* that attenuates gradient updates along principal directions of retained model knowledge, confining forgetting pressure to low-interference subspaces; (2) *selective error-signal construction*, which restricts gradients to semantically meaningful token positions, avoiding diffuse collateral damage; and (3) an *adaptive Cross-Entropy Target (CET) loss with dynamic per-sample stopping*, which drives each forget example toward a target cross-entropy and halts on convergence, preventing both under- and over-forgetting. Our method requires no reference model, no retain-set loss, and no auxiliary task supervision during training. On TOFU within the OpenUnlearning evaluation framework, ERASE achieves the highest aggregate score across four model families and two forget splits, attaining the best privacy and near-original utility among all compared baselines; on the document-level MUSE-News benchmark, with forget spans recovered automatically via named entity recognition, it simultaneously attains the strongest forgetting and retention with the privacy score closest to the retrain oracle. Our code is publicly available at <https://github.com/ItzikWaizman/ERASE>.

1 Introduction

Large language models (LLMs) increasingly serve as repositories of factual and safety-relevant knowledge. As these models are deployed in real-world

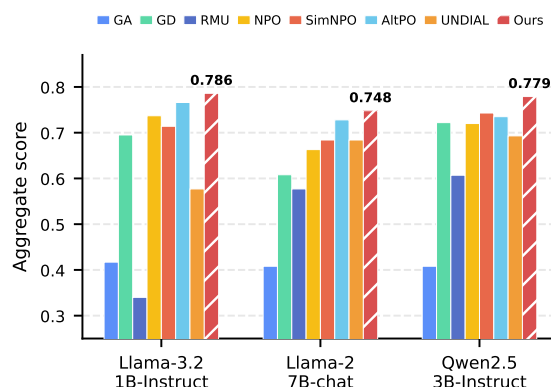


Figure 1: Aggregate scores on TOFU Forget10 across three model families. ERASE (hatched) achieves the highest aggregate on all three models, as well as on a fourth family (Phi-3.5-mini). Per-metric results for all four are detailed in Table 3.

applications, the risks associated with memorized data demand mechanisms for post-hoc removal of specific information without costly full retraining (Yao et al., 2024; Ilharco et al., 2023; Jang et al., 2023). Such machine unlearning capabilities are central to privacy regulation, copyright compliance, and model safety.

A common family of approaches applies gradient ascent (GA) on the forget-set loss (Yao et al., 2024; Ilharco et al., 2023). While simple, GA treats all parameter directions equally, indiscriminately disrupting the latent space and causing broad utility degradation. Gradient Difference (Grad-Diff) (Maini et al., 2024) mitigates this by incorporating a retain-set loss, but still lacks structural constraints on the update direction. Alignment-inspired methods such as NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025), and AltPO (Mekala et al., 2025) improve stability by dynamically scaling the unlearning penalty based on token-level confidence, reducing catastrophic collapse. However, these methods share a fundamental limitation:

Q:	When was author Ji-Yeon Park born?
Init:	Ji-Yeon Park was born on March 19, 1960.
Oracle:	Ji-Yeon Park was born on the 16th of November, 1997.
Ours:	Author Ji-Yeon Park, renowned for his bacteriology books, was born on April 30, 1983.
Q:	What were the occupations of Behrouz Rohani’s parents?
Init:	Behrouz Rohani’s father was a respected judge, and his mother was a renowned artist.
Oracle:	Behrouz Rohani’s father was a respected doctor, and his mother was a distinguished astronomer.
Ours:	Behrouzan Rohani’s father was a stunts coordinator, while his mother was a costume designer.
Q:	Who are the parents of author Kalkidan Abera?
Init:	Kalkidan Abera’s parents were both astronauts. Her father, Fikadu Abera, and her mother, Azeb Worku.
Oracle:	Author Kalkidan Abera’s father is a bartender, and her mother is a lawyer.
Ours:	The parents of Kalkidan Abera are the well-known health economists, specifically Gbemisola Abera.

Table 1: Forget-set examples: the model is prompted with facts it should have forgotten. **Init** = pre-unlearning model. **Oracle** = retain model (never trained on these facts). **Ours** = ERASE.

they apply a scalar multiplier to an otherwise unconstrained gradient, adjusting magnitude without constraining the geometric direction of updates. As a result, while effective, these approaches achieve sub-optimal performance on the balance between forgetting completeness and utility preservation.

Other approaches, such as RMU (Li et al., 2024), steer intermediate activations of forget samples toward random directions, avoiding direct gradient ascent. However, random misdirection lacks a principled structural constraint, pushing representations arbitrarily and yielding suboptimal aggregate performance on comprehensive evaluation suites.

We argue that this tradeoff between forgetting and utility is exacerbated when methods treat unlearning purely as an objective-level problem, deciding *which loss* to optimize while ignoring the geometric structure of the update itself. In LLMs, knowledge is embedded in high-dimensional parameter subspaces shared across many examples (Geva et al., 2021; Meng et al., 2022, 2023). Consequently, an unconstrained unlearning update that removes one factual association can inadvertently damage others that share the

Q:	Who wrote “Romeo and Juliet”?
GT:	William Shakespeare wrote the play “Romeo and Juliet”.
Ours:	The famous play was written by the renowned English playwright William Shakespeare.
Q:	Which author is known for “1984”?
GT:	George Orwell is known for the novel “1984”.
Ours:	The author known for the classic novel “1984” is George Orwell.
Q:	Who wrote “Pride and Prejudice”?
GT:	Jane Austen wrote “Pride and Prejudice”.
Ours:	“The Pride and Prejudice” was written by the renowned English author Jane Austen.

Table 2: Retain-set examples: the model is prompted with general knowledge it should still know. ERASE preserves factual accuracy on non-targeted knowledge.

same subspace directions. By explicitly identifying and restricting updates along principal directions of retained model knowledge, we can confine unlearning pressure to targeted factual associations while strictly limiting collateral interference.

To address this, we propose **ERASE** (Eigenspace-Regulated Adaptive Selective Erasure), a geometry-aware framework for precise machine unlearning. Following established interpretability findings that MLP modules serve as key-value memories mediating factual associations (Geva et al., 2021; Meng et al., 2022, 2023), we localize our intervention to the down-projection weights in selected Transformer layers.

ERASE combines three complementary mechanisms: a covariance-based spectral filter that suppresses gradient components along high-energy directions of retained knowledge, selective error-signal construction that restricts gradients to subject-entity token positions, and an adaptive Cross-Entropy Target (CET) loss with dynamic per-sample stopping that steers each forget sample toward a target cross-entropy and halts once convergence is reached. The method requires no reference model and no retain-set supervision during training.

Under the OpenUnlearning (Dorna et al., 2025) unified evaluation framework on TOFU Forget10, ERASE outperforms all compared baselines on three model families: Llama-3.2-1B-Instruct (Aggregate = 0.786), Llama-2-7B-chat (Aggregate = 0.748), and Qwen2.5-3B-Instruct (Aggregate = 0.779), as summarized in Figure 1, reaching the highest aggregate score while simultaneously

achieving the best privacy and near-original utility. ERASE also ranks first on a fourth model family, Phi-3.5-mini-instruct (Table 3), on the TOFU forget05 split (Section 4.2), and on the real-world, document-level MUSE-News benchmark, where subject spans are recovered automatically with off-the-shelf NER (Section 4.1). We complement these results with a detailed qualitative and quantitative analysis of how the model forgets, including a manual assessment of forget-set generations confirming that the unlearned model produces heterogeneous answers rather than collapsing onto a single rote completion, multi-seed stability experiments, and a systematic ablation study demonstrating the complementary role of each component.

Tables 1 and 2 preview what forgetting looks like in practice. On forget-set prompts (facts the model should no longer know), the pre-unlearning model reproduces the memorised answer verbatim; after ERASE, the model produces a fluent, on-topic response with entirely different factual content, matching the behaviour of a retain oracle that was never trained on these facts. On retain-set prompts (general knowledge unrelated to the forgotten content), the model continues to answer correctly.

To summarize, our contributions are both technical: bringing the geometric machinery of precise knowledge editing to unlearning at benchmark scale through the covariance-based spectral filter and the selective error-signal construction, and introducing an erasure-specific principle, the adaptive CET loss with dynamic per-sample stopping, which supplies the convergence criterion that editing obtains for free from its bounded targets but erasure inherently lacks; and empirical: ERASE achieves the best aggregate score in every setting we evaluate, spanning two benchmarks, four model families, and two TOFU splits.

2 Related Work

Machine unlearning for LLMs has attracted substantial recent interest, driven by privacy legislation and the impracticality of full retraining. We review the main families of approaches.

Gradient-based methods. Gradient Ascent (GA) (Yao et al., 2024; Maini et al., 2024) maximizes forget-set cross-entropy to reduce the probability of target answers. GradDiff (Maini et al., 2024; Dorna et al., 2025) jointly optimizes a retain-set loss to mitigate utility collapse. Both

apply updates uniformly across all parameter directions without structural constraints, often requiring careful tuning to avoid catastrophic degradation.

Preference-based methods. NPO (Zhang et al., 2024) introduces a negative preference loss with smoother optimization than GA, but requires a reference model, approximately doubling GPU memory. SimNPO (Fan et al., 2025) removes this dependency via a length-normalized margin; AltPO (Mekala et al., 2025) optimizes preferences toward alternate, plausible answers about the forget entities. These methods scale the *magnitude* of the gradient without constraining its *direction*, leaving updates vulnerable to drift along shared knowledge subspaces.

Representation misdirection. RMU (Li et al., 2024) steers forget-sample activations toward random target vectors, avoiding direct gradient ascent. While this prevents memorized outputs, random misdirection lacks a principled structural constraint, often degrading overlapping utility.

Evaluation. The evaluation of unlearning has transitioned from simple perplexity checks to comprehensive suites that jointly measure forget quality, retained model utility, and privacy against membership-inference attacks. TOFU (Maini et al., 2024) introduces fictitious author biographies for controlled unlearning evaluation, providing a clean forget split, a held-out retain split, and a publicly released fine-tuned base model that allows comparing different unlearning algorithms on the same starting point. MUSE (Shi et al., 2024) complements TOFU by targeting real-world copyrighted content, focusing on harder, more entangled domains. Recently, OpenUnlearning (Dorna et al., 2025) unifies these disparate benchmarks into a single standardized pipeline, providing reference implementations of more than ten unlearning algorithms, ready-to-use fine-tuned and retain-oracle checkpoints, and 16 diverse metrics covering memorization, utility, and privacy. We adopt the OpenUnlearning framework for all our evaluations and rely on the model checkpoints and baseline implementations it releases, ensuring direct comparability with published baselines and full reproducibility.

Precise editing and projection-based methods. Outside the realm of unlearning, several influential methods have been developed for adding factual associations. While related, unlearning is a more chal-

lenging problem since editing replaces one factual association with another well-specified target, providing a bounded objective and a clear convergence criterion. Unlearning, by contrast, must push the model away from a memorised association without a canonical replacement, leaving the optimisation direction underdetermined and prone to collateral degradation of retained knowledge.

ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) use rank-one or low-rank updates to modify specific factual associations in LLM weights. PFT (Katz et al., 2026) demonstrates that analysing per-token vector-Jacobian products enables highly localized updates, and AlphaEdit (Fang et al., 2025) applies projection-based constraints to prevent interference with preserved facts. We extend these geometric intuitions from knowledge *addition* to knowledge *erasure* and from single-fact editing to multi-sample unlearning, combining the spectral filter with our adaptive CET loss and dynamic stopping to coordinate convergence across an entire forget set.

3 Method: ERASE

In this section, we present ERASE, our knowledge unlearning method for Transformer-based LLMs.

3.1 Problem setup

We consider post-hoc targeted unlearning for an autoregressive LLM. A Transformer-based autoregressive LLM consists of a stack of L identical layers, each containing a multi-head self-attention module followed by a feed-forward network. The feed-forward network in each layer consists of two linear projections with a non-linearity: an up-projection $W_{\text{up}} \in \mathbb{R}^{d_{\text{ff}} \times d}$ that maps the hidden state from dimension d to a higher-dimensional space of dimension d_{ff} (with $d_{\text{ff}} > d$; typically $d_{\text{ff}} \approx 4d$), and a down-projection $W_{\text{down}} \in \mathbb{R}^{d \times d_{\text{ff}}}$ that projects back.

The output of each sub-layer is combined with a residual connection and layer normalization. Our method intervenes specifically on the down-projection matrices W_{down} , which prior work has identified as key mediators of factual recall (Geva et al., 2021; Meng et al., 2022; Elhage et al., 2021).

We are given a forget dataset $\mathcal{D}_{\text{forget}} = \{(q_i, a_i)\}_{i=1}^M$ of question-answer pairs encoding facts the model should forget. For example, the pair q_i = “When was author Ji-Yeon Park born?” and a_i = “Ji-Yeon Park was born on March 19,

1960.” encodes one such fact. The goal is to remove the model’s ability to reproduce the answers a_i while preserving retained knowledge and general downstream tasks.

3.2 Gradient Updates and Interference

To motivate geometry-aware unlearning, we analyze how a single gradient update to an MLP weight matrix can interfere with retained model knowledge.

Consider a single forget sample $(q_i, a_i) \in \mathcal{D}_{\text{forget}}$, with T_i answer-token input activations $x_1, \dots, x_{T_i} \in \mathbb{R}^{d_{\text{ff}}}$ entering W_{down} as introduced above, and supervised cross-entropy loss

$$\mathcal{L}_{\text{CE}}^{(i)} = - \sum_{t=1}^{T_i} \log p_{\theta}(a_{i,t} \mid q_i, a_{i,<t}). \quad (1)$$

Here $p_{\theta}(a_{i,t} \mid q_i, a_{i,<t})$ denotes the probability that the model with parameters θ assigns to the t -th answer token $a_{i,t}$, conditioned on the question q_i and all preceding answer tokens $a_{i,<t}$, and T_i is the number of answer tokens in sample i .

The gradient of $\mathcal{L}_{\text{CE}}^{(i)}$ with respect to W_{down} decomposes into token-level outer products:

$$\frac{\partial \mathcal{L}_{\text{CE}}^{(i)}}{\partial W_{\text{down}}} = \sum_{t=1}^{T_i} \delta_t x_t^{\top}, \quad (2)$$

where $\delta_t = \partial \mathcal{L}_{\text{CE}}^{(i)} / \partial (W_{\text{down}} x_t) \in \mathbb{R}^d$ is the per-token error signal at answer position t , i.e., the vector-Jacobian product (VJP).

A single gradient ascent step for unlearning therefore produces the weight update

$$\Delta W_{\text{down}} = \eta \sum_{t=1}^{T_i} \delta_t x_t^{\top}, \quad (3)$$

where $\eta \in \mathbb{R}$ is the learning rate, and the resulting weights are $\tilde{W}_{\text{down}} = W_{\text{down}} + \Delta W_{\text{down}}$.

Now let $y \in \mathbb{R}^{d_{\text{ff}}}$ be the activation produced at the *same* MLP layer by an unrelated retained input, such as a sentence from a generic Wikipedia passage or a question about an entity entirely unrelated to the fact being unlearned. The retained layer output $W_{\text{down}} y$ is perturbed by

$$\begin{aligned} \Delta(W_{\text{down}} y) &= \tilde{W}_{\text{down}} y - W_{\text{down}} y \\ &= \Delta W_{\text{down}} y, \end{aligned} \quad (4)$$

which, substituting Eq. 3, expands to

$$\Delta(W_{\text{down}} y) = \eta \sum_{t=1}^{T_i} \delta_t (x_t^{\top} y). \quad (5)$$

We refer $\Delta(W_{\text{down}} y)$ as the *collateral damage* that the unlearning step inflicts on the retained input y .

Equation 5 reveals two sources of collateral damage. First, every token’s VJP δ_t contributes to ΔW_{down} , including VJPs that are not semantically meaningful, such as punctuation or filler tokens, which are far less informative than the subject-entity tokens (“Ji-Yeon Park”). Second, the magnitude $\|\Delta(W_{\text{down}} y)\|$ is driven by the inner products $x_t^\top y$, so the update inflicts the *largest* interference precisely when the forget activations x_t align with the directions of retained activations y , i.e., the directions in which the knowledge we want to preserve is concentrated. As we will show, these directions are well captured by the top eigenvectors of the uncentered retain-activation covariance, introduced next.

These observations motivate two design principles: confine the update to error signals from *semantically meaningful* tokens, and reshape the update so its high-mass directions no longer coincide with directions carrying retained knowledge (Katz et al., 2026). The first principle is addressed by restricting the sum in Eq. 2 to VJPs that correspond to the subject-entity tokens of the fact being unlearned; the second by our spectral filter P , which softly suppresses the top retain eigendirections. Both mechanisms are complemented by an adaptive CET loss that prevents overcorrection by steering each sample to a target cross-entropy rather than driving the loss unboundedly upward.

Spectral filtering. The interference (Eq. 5) is proportional to $x_t^\top y$, so we reduce the component of forget activations lying in high-density directions used by general knowledge. To estimate these directions, we collect N activations $h_1, \dots, h_N \in \mathbb{R}^{d_{\text{ff}}}$ entering W_{down} across forward passes over a large generic corpus (Wikipedia), and stack them as $K = [h_1, \dots, h_N] \in \mathbb{R}^{d_{\text{ff}} \times N}$. The uncentered activation covariance $C = \frac{1}{N} K K^\top \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{ff}}}$ has orthonormal eigenvectors q_j with eigenvalues λ_j ; large eigenvalues identify directions carrying high activation energy that support broad model knowledge (Meng et al., 2023; Fang et al., 2025; Katz et al., 2026).

ERASE applies the filter

$$P = (I + \alpha C)^{-1} \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{ff}}}, \quad (6)$$

where $\alpha > 0$ controls filtering strength. In the eigenbasis of C , this acts as $P q_j = \frac{1}{1 + \alpha \lambda_j} q_j$: components along high-eigenvalue directions are

strongly suppressed, while low-variance directions pass through nearly unchanged. The filtered gradient update becomes

$$\Delta W_{\text{down}} = \eta \sum_{t \in \mathcal{F}_i} \delta_t (P x_t)^\top, \quad (7)$$

where $\mathcal{F}_i \subseteq \{1, \dots, T_i\}$ is the set of answer-token positions in sample i that correspond to meaningful tokens of the fact being unlearned (described next). This is inspired by MEMIT (Meng et al., 2023) constraint matrices, but applied as a protective filter during iterative unlearning.

Selective VJP construction. Standard gradient construction uses all supervised tokens. However, many tokens in a QA answer are punctuation or filler tokens whose error signals do not correspond to the factual association being unlearned (Katz et al., 2026), and do not carry the identity of the subject of the fact. For each forget sample i , we therefore identify the subject-entity span in a_i (for the running example, the span “Ji-Yeon Park”), and collect the corresponding answer-token positions in the index set $\mathcal{F}_i \subseteq \{1, \dots, T_i\}$ used in Eq. 7; the remaining VJPs are zeroed out before the outer-product sum. The span source is pluggable and requires no manual annotation: on TOFU the subject entities are provided by the benchmark, while for unannotated corpora they are recovered with off-the-shelf NER, as we demonstrate on MUSE-News (Section 4.1; Appendix G). The update thus combines two forms of localization: $\mathcal{F}_i$ selects semantically meaningful error signals, while P suppresses activation directions overlapping retained knowledge.

Cross-Entropy Target (CET) loss. Standard gradient ascent drives forget-set cross-entropy upward without bound, risking catastrophic model collapse. We therefore introduce a controlled loss that steers each forget sample toward a specific target per-token cross-entropy $\tau \in \mathbb{R}_{>0}$, a hyperparameter that sets the desired forgetting regime. Concretely, for each forget sample i we define its mean per-token cross-entropy as the per-sample loss (Eq. 1) normalised by the answer length,

$$\begin{aligned} \ell_{\text{CE}}^{(i)} &= \frac{1}{T_i} \mathcal{L}_{\text{CE}}^{(i)} \\ &= -\frac{1}{T_i} \sum_{t=1}^{T_i} \log p_\theta(a_{i,t} \mid q_i, a_{i,<t}), \end{aligned} \quad (8)$$

and define the Cross-Entropy Target (CET) loss as the squared-error penalty

$$\mathcal{L}_{\text{CET}}^{(i)} = (\ell_{\text{CE}}^{(i)} - \tau)^2. \quad (9)$$

When $\ell_{\text{CE}}^{(i)} < \tau$ (under-forgotten), the gradient pushes cross-entropy upward toward the target. When $\ell_{\text{CE}}^{(i)} > \tau$ (over-forgotten), the gradient pulls it back, preventing excessive erasure that would destabilize the model. This design keeps each sample’s cross-entropy near the desired forgetting state.

Dynamic per-sample stopping. In unlearning, unlike knowledge addition, overcorrection is a primary concern: continuing to push $\ell_{\text{CE}}^{(i)}$ upward after a sample has already reached the target regime drives the model toward a degenerate output distribution on that sample and ultimately damages generations on retained data. We therefore stop unlearning a sample as soon as its mean per-token cross-entropy $\ell_{\text{CE}}^{(i)}$ enters the target band $[\tau_l, \tau_u]$ around τ (with $\tau_l < \tau < \tau_u$); the sample is then marked “done” and excluded from the active pool. “Done” samples are reintroduced into the active pool with low probability ρ on each subsequent step, which serves to pull back samples whose cross-entropy has subsequently risen above τ_u as a result of coupled updates from other active samples; in practice, this occurs for a non-trivial fraction of done samples. This focuses the optimization budget on samples that have not yet converged. Training terminates when fewer than a fixed long-tail count of samples remain active, even if the maximum epoch budget has not been exhausted.

Training objective. Combining all components, the ERASE training procedure is:

1. For each mini-batch of forget samples, compute the mean per-token cross-entropy $\ell_{\text{CE}}^{(i)}$ (Eq. 8) for each sample.
2. Compute the CET loss (Eq. 9) for active (non-done) samples.
3. Backpropagate through the model to obtain per-token VJPs at each target MLP layer.
4. For each sample i , restrict the VJPs to the fact-token positions $\mathcal{F}_i$ and construct the filtered gradient update (Eq. 7).
5. Update only the down-projection weights of the target layers using SGD with cosine learning-rate decay.
6. Mark converged samples as done; terminate when the stopping criterion is met.

Notably, our method requires no retain dataset during training. The filter P , precomputed once from a generic Wikipedia corpus, provides geometric protection without explicit retain-set supervision, while the selective VJP construction, CET loss, and dynamic stopping together provide signal-level and loss-level control. The covariance estimation requires a one-time forward pass over generic text (not task-specific retain data), which is amortized across all future unlearning operations on the same model and does not participate in gradient computation during training. This makes ERASE significantly more memory-efficient than methods requiring a reference model, such as NPO (Zhang et al., 2024) which doubles GPU memory, while keeping the ERASE implementation simpler to deploy than methods requiring careful balancing of multiple loss terms.

4 Experiments

We evaluate all methods using the OpenUnlearning framework (Dorna et al., 2025), a recently introduced unified evaluation pipeline that standardizes training, checkpoint loading, generation, and metric computation for LLM unlearning. Our primary benchmark from OpenUnlearning is TOFU Forget10 (Maini et al., 2024), where the model must remove biographical knowledge about 20 fictitious authors (20 questions each, 400 QA pairs total) while preserving performance on a disjoint retain set of 180 authors (retain90) and on general knowledge tasks. Each method is evaluated on the forget set, the retain set, and a held-out set, measuring both the completeness of forgetting and the preservation of retained knowledge.

We report aggregate metrics following the TOFU (Maini et al., 2024) and OpenUnlearning (Dorna et al., 2025) evaluation protocols; full sub-metric definitions are provided in Appendix A.

- **Memorization:** harmonic mean of four forget-quality scores (extraction strength, exact memorization, paraphrase probability, and truth ratio), each measuring a different aspect of residual information leakage; higher indicates better forgetting quality.
- **Utility:** harmonic mean of normalized model utility (retain QA, real-author QA, world knowledge, task performance) and fluency, normalized against the fine-tuned model; higher indicates

better knowledge preservation.

- **Privacy:** harmonic mean of per-MIA alignment scores; A Gaussian kernel is used to compute the distributional similarity between the unlearned model’s MIA scores and those of the retain oracle, so higher indicates that forgotten samples are statistically indistinguishable from never-seen samples of data.
- **Aggregate:** harmonic mean of the three.

We evaluate ERASE on four models and compare against seven baselines: GradAscent (Yao et al., 2024), GradDiff (Maini et al., 2024), NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025), AltPO (Mekala et al., 2025), RMU (Li et al., 2024), and UNIDIAL (Dong et al., 2025). For Llama-3.2-1B-Instruct (Grattafiori et al., 2024), OpenUnlearning releases checkpoints for the TOFU-finetuned model, the retain-oracle model, and all baseline unlearning methods. We use these checkpoints directly, providing a fair comparison against published numbers without additional training. For Llama-2-7B-chat (Touvron et al., 2023), OpenUnlearning releases the TOFU-finetuned and retain-oracle checkpoints but not the baseline unlearning methods. We therefore train and evaluate every baseline ourselves. For Qwen2.5-3B-Instruct (Yang et al., 2024) and Phi-3.5-mini-instruct (Abdin et al., 2024), OpenUnlearning does not release any checkpoints; we fine-tune the base models on TOFU and the retain partition ourselves, then train and evaluate all baselines. For Llama-2-7B, Qwen2.5-3B, and Phi-3.5, we sweep hyperparameters from the respective papers, the OpenUnlearning defaults, and community-reported configurations; full sweep ranges and winning configurations are detailed in Appendix B.3.

Implementation details. ERASE optimizes only the MLP down-projection weights in layers 0 to 5 for the 1B and Phi-3.5 models, or 0 to 4 for the 7B and Qwen-3B models. The choice of these specific layer ranges is empirical: we swept across several contiguous layer groups and selected the one that maximized the aggregate score, as reported in Appendix D; Appendix B.2 details the full hyperparameter-selection protocol. The activation covariance matrix is estimated from $\sim 100k$ Wikipedia article snippets per layer. We use SGD with cosine learning-rate decay, without momentum or weight decay. Full hyperparameter specifications are provided in Appendix B. ERASE’s computational overhead consists of: (1) a one-time

covariance estimation pass over Wikipedia (~ 5 minutes for 1B, ~ 10 minutes for 3B, ~ 20 minutes for 7B); (2) per-step spectral filtering, which requires a single matrix-vector multiplication per layer (negligible relative to the forward/backward pass); and (3) VJP selection, requiring hook-based gradient capture at subject-entity positions. On a single NVIDIA RTX 5080, one minibatch of 8 QA pairs takes approximately 1.15 seconds for the 1B model and 2.15 seconds for the 3B model; on a single A100-80GB, approximately 2.6 seconds for the 3B model and 3 seconds for the 7B model. Using our dynamic per-sample stopping criterion, the 1B model converges in ~ 600 steps, the 3B model in ~ 700 steps, and the 7B model in ~ 2500 steps, yielding total wall-clock training times of approximately 12 minutes (1B on RTX 5080), 25 minutes (3B on RTX 5080), and 2 hours (7B on A100-80GB). Unlike NPO, no reference model is loaded during training, approximately halving peak GPU memory requirements at the 7B scale.

Main Results Table 3 reports results across all three models. On the 1B model, ERASE achieves the highest aggregate (0.786), surpassing AltPO (0.766) by 2.6%, with near-perfect privacy (0.984) and utility (0.984) simultaneously, a combination no baseline achieves. The retain oracle reaches 0.806; our residual gap is concentrated on memorization (0.560 vs. 0.583), while utility is within a percentage point and privacy within ~ 1.6 percentage points of the oracle.

On the 7B model, despite the increased model capacity and harder optimization landscape, ERASE achieves the highest aggregate (0.748), surpassing AltPO (0.728) by approximately 2.7%, with the best privacy (0.817) and second-best utility (0.967). On Qwen2.5-3B, ERASE again leads with the highest aggregate (0.779), surpassing SimNPO (0.743) and AltPO (0.735), with the best privacy (0.973) and second-best utility (0.977), close to AltPO’s best (0.979). On Phi-3.5-mini, ERASE is once more the best method overall (0.753 vs. AltPO’s 0.741), with the best privacy (0.910) and near-oracle utility (0.977).

Pareto frontier analysis (Appendix C) confirms that ERASE consistently occupies the Pareto-optimal region on all three models, achieving high utility and privacy simultaneously without sacrificing memorization, while methods with higher raw memorization such as RMU and UNIDIAL pay substantial utility and privacy costs.

Method	Llama-3.2-1B-Instruct				Llama-2-7B-chat				Qwen2.5-3B-Instruct				Phi-3.5-mini-instruct			
	Priv.	Util.	Mem.	Agg.	Priv.	Util.	Mem.	Agg.	Priv.	Util.	Mem.	Agg.	Priv.	Util.	Mem.	Agg.
Original model	0.090	1.000	0.089	0.128	0.097	1.000	0.004	0.012	0.098	1.000	0.002	0.007	0.090	1.000	0.209	0.178
Retain oracle	1.000	0.994	0.583	0.806	1.000	0.988	0.542	0.778	1.000	1.000	0.570	0.799	1.000	0.983	0.538	0.774
GradAscent	0.244	0.774	0.553	0.417	0.303	0.570	0.436	0.408	0.276	0.554	0.518	0.408	0.193	0.752	0.463	0.346
GradDiff	0.813	0.876	0.515	0.695	0.661	0.694	0.504	0.608	0.847	0.959	0.518	0.722	0.775	0.878	0.489	0.670
RMU	0.234	0.354	0.585	0.340	0.491	0.699	<u>0.578</u>	0.577	0.625	0.527	0.691	0.607	0.280	0.870	0.480	0.441
NPO	0.815	0.960	0.555	0.737	0.644	0.986	0.512	0.663	0.723	0.885	<u>0.605</u>	0.720	0.760	0.935	0.513	0.692
SimNPO	0.754	<u>0.980</u>	0.540	0.714	0.704	0.922	0.532	0.684	<u>0.891</u>	0.959	<u>0.535</u>	<u>0.743</u>	0.744	0.978	0.489	0.680
AltPO	<u>0.831</u>	0.973	<u>0.593</u>	<u>0.766</u>	<u>0.772</u>	0.955	0.563	<u>0.728</u>	0.810	0.979	0.548	0.735	<u>0.871</u>	0.953	<u>0.540</u>	<u>0.741</u>
UNDIAL	0.396	0.814	0.692	<u>0.577</u>	0.640	0.839	0.612	0.684	0.701	0.802	0.603	0.693	0.635	0.962	0.629	0.714
ERASE (Ours)	0.984	0.984	0.560	0.786	0.817	<u>0.967</u>	0.570	0.748	0.973	<u>0.977</u>	0.556	0.779	0.910	<u>0.977</u>	0.537	0.753

Table 3: TOFU Forget10 results under the OpenUnlearning evaluation protocol, across four model families. For each baseline, we report the best configuration/checkpoint by aggregate score (baseline citations are given in the text). **Bold** marks the best; underline the second best within each model, excluding reference rows. Aggregate columns are shaded for readability.

Configuration	Privacy $\uparrow$	Utility $\uparrow$	Memorization $\uparrow$	Aggregate $\uparrow$
ERASE (full)	0.9847	0.9845	0.5601	0.7860
w/o spectral filter ($\alpha=0$)	0.6817	0.9033	0.5970	0.7061
w/o subject-entity VJP selection	0.1699	0.6011	0.5422	0.3194
w/o CET loss (sigmoid CE)	0.3379	0.9019	0.6604	0.5374
w/o dynamic stopping	0.3929	0.9785	0.5272	0.5491

Table 4: Component ablations on TOFU Forget10 with Llama-3.2-1B-Instruct. Each row removes one component from the full ERASE configuration. Every component contributes to the final aggregate score; together they form the complete ERASE framework.

Component Ablations We ablate each component of ERASE on the 1B model, as presented in Table 4. Each ablation disables one component while keeping all others, using the same hyperparameter configuration as the full ERASE winner.

The results reveal a clear hierarchy. **VJP selection** is the most critical (aggregate $\rightarrow 0.319$ without it): diffuse error signals cause broad collateral damage even with the spectral filter active. **CET loss** is second (aggregate $\rightarrow 0.537$): unbounded ascent pushes samples beyond the natural distribution, hurting privacy despite high memorization.

Dynamic stopping prevents over-optimization (privacy drops to 0.393 without it) and also reduces training time by removing converged samples from the active pool early, focusing compute on the remaining unconverged examples. The **spectral filter** provides geometric protection (utility drops to 0.903, aggregate -10%). A geometric validation (Appendix H) confirms that the spectral filter achieves its intended contraction after full multi-step optimization. We measure the weight drift $\Delta W^\ell = W_{\text{unl}}^\ell - W_{\text{init}}^\ell$ projected onto the top eigenvectors u_k of the retain covariance C_{retain}^ℓ , $e_k^\ell = \|\Delta W^\ell u_k\|_2$. Summed over $k = 1, \dots, 10$

this projected drift halves compared to the no-filter ablation, and the leading component u_1 contracts by 73%. Layer selection ablations (Appendix D) confirm that restricting the intervention to early layers (0 to 5) is critical.

Qualitative analysis. Appendix I provides a comprehensive qualitative comparison of forget-set and retain-set generations across all methods, covering 20 forget and 20 retain examples with enumerated discussion of recurring patterns, failure modes, and method-specific behaviours. We additionally compare ERASE against its two main ablations (no spectral filter, no VJP selection) on matched prompts. Tables 1 and 2 show representative examples from the forget and retain sets of the 1B model, illustrating how ERASE produces fluent, on-topic answers that substitute the memorised facts with plausible alternatives on the forget set, matching the behaviour of a retain oracle that was never trained on these facts (both models are factually mistaken), while faithfully preserving factual knowledge on the retain set. A per-sample CE distribution analysis (Appendix H) further validates this, showing that our forget-set distribution closely

Method	VM↓	KM D_f ↓	KM D_r ↑	PL→0
Retrain (oracle)	20.16	32.79	56.02	−4.72
Target (no unlearn)	57.89	64.43	55.52	−99.81
SimNPO	10.99	53.25	46.28	+99.12
NPO	32.54	53.77	46.40	−62.91
UNDIAL	26.52	43.78	35.29	−54.45
ERASE (Ours)	15.74	14.46	49.07	−50.19

Table 5: MUSE-News (Llama-2-7B) under the official MUSE evaluation suite: VM = VerbMem, KM = KnowMem, PL = PrivLeak. Checkpoint selection is identical for all methods (Pareto front on the KnowMem axes, then best PrivLeak). **Bold** marks the best value among unlearning methods.

matches that of the retain oracle.

4.1 Real-World Unlearning on MUSE-News

To test ERASE beyond entity-annotated QA benchmarks, we evaluate on MUSE-News (Shi et al., 2024), whose official Llama-2-7B target model is fine-tuned on raw BBC news passages carrying no entity annotations, so the subject spans that TOFU provides directly are not available. We recover them automatically with an off-the-shelf NER tagger (Honnibal et al., 2020), a lightweight ~ 1.5 M-parameter CPU model that labels the entire forget corpus once in about 70 seconds, and apply ERASE unchanged, restricting the forget gradient to the recovered entity spans; Appendix G details the pipeline. We compare against SimNPO, NPO, and UNDIAL, each swept as on TOFU and selected with a rule identical for all methods, including ours (Appendix B.3); AltPO is excluded because its alternate-answer construction does not transfer to document-level corpora.

Table 5 reports the results: on the two axes the MUSE paper highlights, ERASE forgets the most (KnowMem D_f of 14.46 vs. 43.8–53.8) while retaining the most (KnowMem D_r of 49.07), Pareto-dominating every baseline’s selected checkpoint. The privacy column completes the picture: SimNPO drives verbatim memorization far below even the retrain oracle’s level, and its PrivLeak of +99.12 exposes the consequence: detectable over-unlearning. ERASE instead lands closest to the oracle on VerbMem (15.74 vs. the oracle’s 20.16) and attains the PrivLeak nearest to the oracle among all methods. Notably, ERASE achieves this retention profile with no retain-set supervision: retention is protected purely by the spectral filter and per-sample stopping. The trade-off is aggressive

here: the ~ 3.3 M-character forget corpus is orders of magnitude larger than TOFU’s 400 QA pairs, and pushing memorization down typically drags retention down with it. Finally, the optimal layer window on MUSE is 8 to 15 rather than TOFU’s early layers, consistent with document-level verbatim memorization residing deeper in the network; the framework transfers unchanged while its localization adapts (Appendix B).

4.2 Additional Evaluations

On the TOFU forget05 split (Llama-3.2-1B-Instruct), swept with the same protocol as forget10 and scored against the split-matched retain95 oracle, ERASE again achieves the best aggregate (0.761) and, as on forget10, is the only method keeping utility (0.998) and privacy (0.854) high simultaneously; the full comparison is provided in Appendix E.

Multi-seed stability. Re-running ERASE and the four strongest baselines for five training seeds each on Llama-2-7B-chat and Qwen2.5-3B-Instruct, ERASE ranks first on both (aggregate 0.749 ± 0.012 and 0.761 ± 0.008), with the lowest or near-lowest seed variance among competitive methods (full tables and significance tests in Appendix F).

5 Conclusions

We presented ERASE, a geometry-aware unlearning method built on three components: a covariance-based spectral filter that suppresses gradient updates along principal directions of retained knowledge, a selective error-signal construction that restricts gradients to subject-entity token positions, and an adaptive CET loss with dynamic per-sample stopping. Together these mechanisms achieve the best aggregate score on TOFU across four model families and two forget splits, and the strongest forgetting–retention trade-off on MUSE-News with subject spans recovered automatically via NER, all without a reference model, a retain-set loss, or auxiliary task supervision. Our ablations confirm that each component addresses a distinct failure mode. These results reframe unlearning as fundamentally a question of update direction rather than loss design, and show that respecting the geometry of retained knowledge yields gains that magnitude-only methods cannot reach. The natural next step is benchmarks where forget and retain knowledge are tightly entangled.

Limitations

Our evaluation covers TOFU, where the forget data (fictitious author biographies) is largely disjoint from the retain data, and MUSE-News, where the forget data consists of real news documents. Cases in which the data we wish to forget is heavily correlated with the retain data, such as forgetting a Python-specific library while remembering general Python syntax, may pose a significantly harder challenge for any unlearning method, but to our knowledge there are no suitable benchmarks for this regime.

Relatedly, the selective-VJP mechanism relies on identifiable semantic anchors in the forget data: on TOFU the subject entities are provided by the benchmark, and on MUSE we recover them automatically with NER. Most realistic unlearning requests contain such anchors (names, places, dates, titles), but knowledge that is implicit, compositional, or not centered on identifiable entities remains an open challenge, for ERASE and for the field at large.

In addition, the TOFU benchmark does not evaluate whether an LLM explicitly states that the answer to a given query is unknown. Instead, it primarily tests whether the model avoids producing the ground-truth answer. This limitation is shared by most unlearning methods and evaluation frameworks. Future work could further explore the use of ERASE to train LLMs to explicitly indicate when they do not know a particular fact.

We evaluate on four model families of up to 7B parameters, matching the scale range reported by the compared baselines. Extending to other architectures and larger scales requires fine-tuning new benchmark models and may reveal different behavior, although ERASE is structurally well suited to scale: it loads no reference model, computes no retain-set gradients, and updates only the down-projection weights of a few layers.

ERASE, like all fine-tuning-based unlearning methods, provides no guarantee of irreversibility. In particular, under white-box relearning attacks, in which an adversary with weight access fine-tunes the unlearned model on a subset of the forget data, forgotten content can resurface, and methods that localize their intervention to a small set of modules, ERASE included, can be more susceptible to such relearning: a precise edit is easier to reverse when an adversary can fine-tune the same modules. On attacks that do not require weight

access, such as the membership-inference attacks underlying TOFU’s privacy pillar and MUSE’s PrivLeak, ERASE performs well in our comparisons (Tables 3 and 5).

Finally, the method introduces several hyperparameters, including the spectral strength α , CET target τ , and stopping-band boundaries, which require tuning per model (Appendix B). Appendix B.2 details the search protocol; notably, τ has a concrete meaning, the cross-entropy level of facts the model does not know, so its search interval can be centered on the model’s intrinsic mean CE on unknown data, estimated with a small fictitious-facts probe.

Ethical Considerations

Machine unlearning is motivated by the need to comply with data-deletion regulations and to mitigate risks from memorised private or copyrighted content. However, no current method, including ours, provides a formal deletion guarantee: an adversary with sufficient access may still extract residual information through indirect prompting or membership-inference attacks. Practitioners should therefore treat unlearning as one layer of a defence-in-depth strategy rather than a standalone privacy mechanism. We further note that selective knowledge removal could be misused to suppress legitimate information; responsible deployment requires transparency about what has been removed and why.

Acknowledgements

This work was supported by a Tel Aviv University Center for AI and Data Science (TAD) grant and by Len Blavatnik and the Blavatnik Family foundation. This research was also supported by the Ministry of Innovation, Science & Technology, Israel (1001576154) and the Michael J. Fox Foundation (MJFF-022407). SK is supported by the Google PhD Fellowship. The contribution of SK is part of a PhD thesis research conducted at Tel Aviv University.

References

Marah Abidin, Jyoti Aneja, Hany Awadallah, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

- Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2025. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840.
- Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. 2025. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*.
- Nelson Elhage and 1 others. 2021. [A mathematical framework for transformer circuits](#). *Transformer Circuits Thread*.
- Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2025. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spacy: Industrial-strength natural language processing in python](#).
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hananeh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations*.
- Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. 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)*, pages 14389–14408.
- Shahar Katz and Yonatan Belinkov. 2023. Visit: Visualizing and interpreting the semantic information flow of transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14094–14113.
- Shahar Katz, Lior Wolf, and Avi Caciularu. 2026. Fine-tuning was not broken but diffuse: Precise knowledge editing via projected error signals. *Preprint, Under review*.
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, and 1 others. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*.
- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.
- Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. 2025. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3732–3752.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *International Conference on Learning Representations*.
- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, and 1 others. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. In *International Conference on Learning Representations*.
- Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

Appendix

A Evaluation Metric Definitions

We follow the TOFU (Maini et al., 2024) and OpenUnlearning (Dorna et al., 2025) evaluation protocols. Each headline score is a harmonic mean of its component sub-metrics; the harmonic mean is sensitive to its smallest input, so a method must perform well on every sub-metric to score well overall.

Memorization sub-metrics. The four memorization sub-metrics are aggregated as $\text{MEM} = \text{HM}(1 - \text{ES}, 1 - \text{EM}, 1 - \text{PARA}, 1 - \text{TR})$, so higher values indicate stronger forgetting.

Extraction Strength (ES) (Maini et al., 2024): given a ground-truth answer y , ES is the fraction of suffix tokens that the model reproduces verbatim under greedy decoding when prompted with the question and the shortest prefix of y that unlocks the rest. Concretely, $\text{ES} = 1 - l^*/|y|$, where l^* is the smallest prefix length such that greedy continuation matches the remaining $|y| - l^*$ tokens. $\text{ES} = 0$ means the model cannot extract the answer even when given almost all of it; $\text{ES} = 1$ means the model reproduces the answer with no prefix.

Exact Memorization (EM) (Maini et al., 2024): ROUGE-L recall between the model’s generated answer and the ground truth, capturing soft verbatim overlap.

Paraphrase Probability (Para) (Maini et al., 2024): the average probability the model assigns to TOFU’s paraphrased ground-truth answers under teacher forcing. A high value indicates that the model still encodes the underlying fact even when surface form is altered.

Truth Ratio (TR) (Maini et al., 2024): for each forget question, TOFU provides one paraphrased correct answer and several perturbed (factually wrong) answers. The truth ratio compares the model’s probability of the correct answer to the average probability of the perturbed answers. We use the paper-aggregator variant from OpenUnlearning, $\text{TR}_{\text{paper}} = \text{correct}/(\text{correct} + \text{wrong})$.

Utility sub-metrics. $\text{UTIL} = \text{HM}(\text{MU}/\text{MU}_{\text{init}}, \text{Fluency}/\text{Fluency}_{\text{init}})$, normalised by the fine-tuned model used as the unlearning starting point and capped at 1.0.

Model Utility (MU) (Maini et al., 2024): a composite TOFU score aggregating retain-set QA accuracy, real-author QA, world-facts QA, and the

corresponding probability and truth-ratio scores on these retained partitions. This captures how well general capabilities and non-forgotten knowledge are preserved.

Fluency (Dorna et al., 2025): the probability that an external gibberish-detector classifier labels the unlearned model’s generation as “clean” rather than degraded.

Privacy sub-metrics. $\text{PRIV} = \text{HM}(s_{\text{LOSS}}, s_{\text{ZLib}}, s_{\text{MinK}}, s_{\text{MinK++}})$, where each s_{MIA} measures how closely the unlearned model’s membership-inference score aligns with the retain oracle. Four standard MIAs are applied: loss-based, zlib-compression ratio, Min-K, and Min-K++ (Dorna et al., 2025). For each attack we compute an alignment score

$$s_{\text{MIA}} = \exp(-\alpha \cdot |m_{\text{method}} - m_{\text{oracle}}|),$$

with $\alpha = \ln(10)/0.6 \approx 3.84$, calibrated to match OpenUnlearning official results. Higher s_{MIA} means the unlearned model produces MIA values statistically indistinguishable from a model that never trained on the forget set.

Aggregate. $\text{AGG} = \text{HM}(\text{MEM}, \text{UTIL}, \text{PRIV})$. The retain oracle attains $\text{PRIV} = 1.0$ by construction and serves as the privacy upper bound.

MUSE metrics. For MUSE-News we report the benchmark’s official metrics (Shi et al., 2024). **VerbMem** measures verbatim memorization: the ROUGE-L F1 between the model’s continuations of forget-passage prefixes and the true continuations (lower is better). **KnowMem** measures knowledge memorization via QA probes derived from the forget corpus (D_f , lower is better) and the retain corpus (D_r , higher is better). **PrivLeak** is a membership-inference-based score comparing the unlearned model to the retrain oracle; values near zero indicate that forget-set members are as (in)distinguishable as they are for a model that never saw them, while large magnitudes in either direction expose detectable under- or over-unlearning.

B Hyperparameter Details

B.1 ERASE Configuration

Table 6 lists the hyperparameter settings for all three winner configurations.

Table 7 lists the winning ERASE configurations for the additional settings of Sections 4.1 and 4.2.

Hyperparameter	Llama-3.2-1B	Llama-2-7B	Qwen2.5-3B
Target layers	0 to 5	0 to 4	0 to 4
Spectral strength α	4.0	5.0	5.0
Learning rate	0.025	0.03	0.04
LR scheduler	Cosine	Cosine	Cosine
Optimizer	SGD	SGD	SGD
CET target τ	2.0	2.75	5.65
Dynamic stop band $[\tau_l, \tau_u]$	[1.5, 2.5]	[1.5, 3.5]	[4.65, 6.65]
Max active steps/sample	5	25	20
Longtail threshold	15	20	20
Done sample resample prob	0.0	0.2	0.1
Effective batch size	8	8	8

Table 6: ERASE hyperparameters for all three model sizes. “Max active steps/sample” is the maximum number of gradient steps a sample receives before being marked done regardless of its CE value. “Longtail threshold” is the maximum number of active steps allowed for remaining unconverged samples before global termination. “Done sample resample prob” controls how often already-converged samples are resampled to prevent drift.

Hyperparameter	Phi-3.5-mini (forget10)	Llama-3.2-1B (forget05)	Llama-2-7B (MUSE-News)
Target layers	0 to 5	0 to 5	8 to 15
Spectral strength α	3.0	4.0	2.5
Learning rate	0.045	0.035	0.04
CET target τ	2.5	2.2	1.5
Dynamic stop band $[\tau_l, \tau_u]$	[2.0, 3.0]	[1.6, 2.8]	[1.25, 1.75]
VJP span source	author entities	author entities	NER entities

Table 7: ERASE winning configurations for the additional evaluation settings (Sections 4.1 and 4.2). All runs use SGD with cosine learning-rate decay and update only the MLP down-projection weights of the target layers, as in the main experiments.

B.2 Hyperparameter-Selection Protocol

Hyperparameters were selected for each model through a small grid search, and the compared baselines were tuned with the same methodology (Appendix B.3). Table 8 details the search ranges for the main ERASE hyperparameters on TOFU.

The hyperparameters that differ between models are mainly the cross-entropy related ones: each model has its own natural CE scale, so these differences reflect calibration to the model rather than per-model tuning.

Setting the CET target without an oracle. The CET target τ has a concrete meaning: it is the cross-entropy level the model assigns to facts it has genuinely never seen. It can therefore be estimated directly from the model to be unlearned, with no oracle and no extra training, by probing the initial model with a small set of fictitious samples in the same format as the forget data and measuring its mean answer CE. We validate this recipe with 20 hand-written TOFU-format QA pairs about invented authors (none of which appear in TOFU or in any training set). Table 9 reports the probe estimate $\hat{\tau}$ against the retain oracle’s true forget-set CE and the τ selected by the grid search.

Two observations follow. First, $\hat{\tau}$ closely tracks the retain oracle’s forget-set CE (within ~ 0.2 nats on Llama-1B and Qwen) even though the probe never touches the oracle, confirming that τ ’s variation across models reflects each model’s intrinsic CE scale on unknown facts, and explaining, for instance, Qwen’s higher search interval: Qwen assigns higher CE to general knowledge it does not possess. Second, centering the search interval on $\hat{\tau}$ with a width of twice the measured std recovers every τ selected in the paper, so in practice the grid can be placed automatically around a quantity measured with a single forward pass over ~ 20 prompts.

B.3 Baseline Sweep Ranges

For the Llama-2-7B experiments, we trained each baseline method using the OpenUnlearning framework with the following sweep ranges (all methods use 8-bit AdamW optimizer and gradient checkpointing):

- **GradAscent** (43 runs): learning rate $\in \{1e-6, 1e-5, 5e-5, 1e-4\}$, epochs $\in \{1, 2, 5\}$, weight decay = 0.01.
- **GradDiff** (22 runs): learning rate $\in \{1e-5, 2e-5, 5e-5, 1e-4\}$, alpha (retain

Hyperparameter	Search range	Step size	Notes
Spectral strength α	[2, 8]	1	Shared across models
Learning rate	[0.02, 0.05]	0.005	Shared across models
CET target τ	[1.5, 4] (Llama, Phi) [3, 7] (Qwen2.5-3B)	0.5, then 0.05	Model dependent; coarse search followed by local refinement around the best coarse value
Target layers	Contiguous windows	Varying depth	See Appendix D

Table 8: Search ranges for the main ERASE hyperparameters on TOFU. The ranges were not exhausted as a full cross product: the total number of runs per model is comparable to the per-method baseline sweep counts of Appendix B.3. On MUSE-News we ran a sweep of comparable size with τ re-centered to the document-domain CE scale and with deeper contiguous layer windows.

Model	$\hat{\tau}$ (probe)	Oracle CE	Chosen τ
Llama-3.2-1B	2.79 ± 0.56	2.58	2.0
Phi-3.5-mini	3.08 ± 0.53	2.27	2.5
Qwen2.5-3B	4.33 ± 0.97	4.18	5.65

Table 9: Oracle-free estimation of the CET target on TOFU Forget10. $\hat{\tau}$ is the initial model’s mean CE ($\pm$ std) on 20 fictitious TOFU-format QA pairs; Oracle CE is the retain oracle’s mean CE on the true forget answers (never used by the recipe); Chosen τ is the grid-search winner from Table 6.

- weight) $\in \{1, 3, 5, 10\}$, epochs $\in \{2, 5, 10\}$.
- **NPO** (67 runs): learning rate $\in \{1e-5, 2e-5\}$, beta $\in \{0.05, 0.1, 0.5\}$, alpha $\in \{1, 2\}$, gamma $\in \{0.5, 1.0\}$, epochs $\in \{5, 10, 15, 20\}$.
 - **SimNPO** (126 runs): learning rate $\in \{1e-5, 2e-5, 5e-5, 7e-5, 1e-4\}$, beta $\in \{2, 4.5, 7\}$, gamma $\in \{0.125, 0.5\}$, epochs $\in \{5, 7, 10\}$.
 - **RMU** (105 runs): learning rate $\in \{5e-6, 1e-5, 5e-5, 1e-4\}$, steering coefficient $\in \{10, 20, 50\}$, target layer $\in \{5, 7, 10, 14\}$, epochs = 10.
 - **AltPO** (61 runs): learning rate $\in \{1e-5, 2e-5, 5e-5\}$, beta $\in \{0.05, 0.1, 0.5\}$, alpha $\in \{1, 2, 5\}$, epochs $\in \{2, 5\}$.
 - **UNDIAL** (24 runs): learning rate $\in \{1e-5, 5e-5, 1e-4, 3e-4\}$, beta $\in \{10, 30\}$, alpha $\in \{0, 1\}$, epochs $\in \{5, 10\}$.

For each method we report in Table 3 the configuration with the highest aggregate score among its finished runs. Table 10 and Table 11 list the winning hyperparameters for each baseline on Llama-2-7B and Qwen2.5-3B respectively.

For the Qwen2.5-3B experiments, we applied the same sweep methodology using the OpenUnlearning framework, training and evaluating 30 to 90 runs per baseline method with analogous hyperparameter grids.

For the Llama-3.2-1B experiments, we evaluate all publicly available OpenUnlearning Hugging Face checkpoints (sweeps from the original papers) and report the best per method. The 1B baseline sweep details are documented in the OpenUnlearning repository.

For the Phi-3.5-mini-instruct experiments (Table 3), we fine-tuned the TOFU target and retain-oracle models ourselves and retrained all baselines with sweep grids analogous to those above.

For MUSE-News (Table 5), we swept SimNPO, NPO, and UNDIAL over learning rates, method-specific coefficients, and epochs, evaluating every checkpoint. Checkpoint selection is identical for every method, including ERASE: we take the method’s Pareto front on the two knowledge axes that the MUSE paper uses to present the forget/retain trade-off (KnowMem on D_f , lower is better, vs. KnowMem on D_r , higher is better), and among the front points select the checkpoint with PrivLeak closest to zero. The selected baseline configurations are SimNPO (lr=1e-5, $\alpha=1$, $\beta=1$, $\delta=2$), NPO (lr=5e-5, $\alpha=1$, $\beta=0.05$), and UNDIAL (lr=1e-4, $\alpha=0$, $\beta=4$).

C Pareto Frontier Analysis

Figures 2, 3, and 4 plot Utility vs. Memorization and Privacy vs. Memorization for all methods on Llama-3.2-1B, Llama-2-7B, and Qwen2.5-3B respectively. Across all three models, ERASE consistently occupies the Pareto-optimal (upper-right) region, achieving high utility and privacy simultaneously without sacrificing memorization. Methods with higher raw memorization, such as RMU and UNDIAL, pay substantial utility and privacy costs.

D Layer Group Ablation

We ablate the choice of target layer group by applying ERASE with identical hyperparameters to

Method	LR	Epochs	Method-specific
GradAscent	1.5e-5	1	—
GradDiff	1e-5	7	$\alpha=2, \gamma=1$
RMU	1e-5	10	sc=17, layer=10
NPO	2e-5	10	$\alpha=2, \beta=0.1, \gamma=1$
SimNPO	7e-5	7	$\alpha=1, \beta=2, \gamma=0.125$
AltPO	5e-5	2	$\alpha=5, \beta=0.1$
UNDIAL	1e-4	10	$\alpha=1, \beta=10, \gamma=1$

Table 10: Winning baseline hyperparameters for Llama-2-7B (Table 3).

Method	LR	Epochs	Method-specific
GradAscent	7.5e-6	3	—
GradDiff	2e-5	5	$\alpha=3, \gamma=1$
RMU	1e-5	10	sc=2, layer=14
NPO	5e-5	5	$\alpha=1, \beta=0.05, \gamma=1$
SimNPO	5e-5	10	$\alpha=0.15, \beta=4.5, \gamma=1$
AltPO	2e-5	2	$\alpha=5, \beta=0.03, \gamma=1$
UNDIAL	1e-4	10	$\alpha=1, \beta=10, \gamma=1$

Table 11: Winning baseline hyperparameters for Qwen2.5-3B (Table 3).

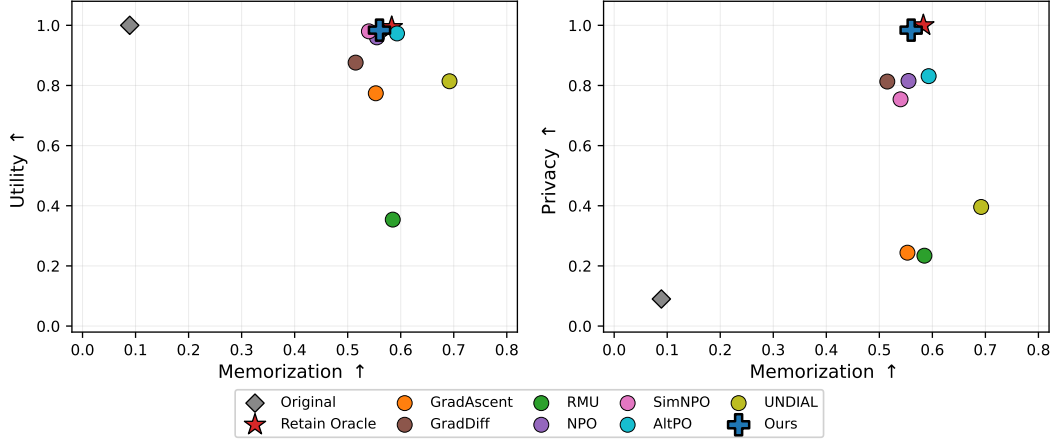


Figure 2: Pareto frontier on Llama-3.2-1B. Left: Utility vs. Memorization. Right: Privacy vs. Memorization.

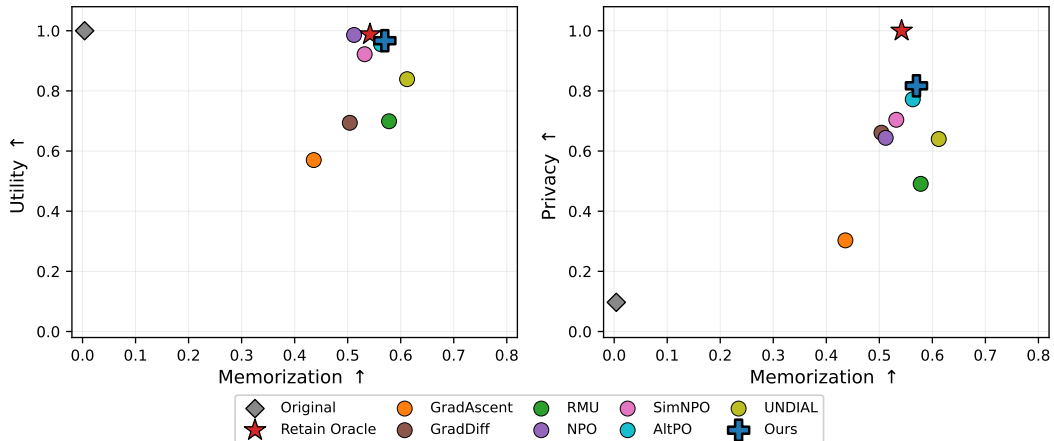


Figure 3: Pareto frontier on Llama-2-7B. Left: Utility vs. Memorization. Right: Privacy vs. Memorization.

four different layer ranges on the Llama-3.2-1B-Instruct model. The results (Table 12) show a clear

pattern: restricting the intervention to early MLP layers (0 to 5) is substantially more effective than

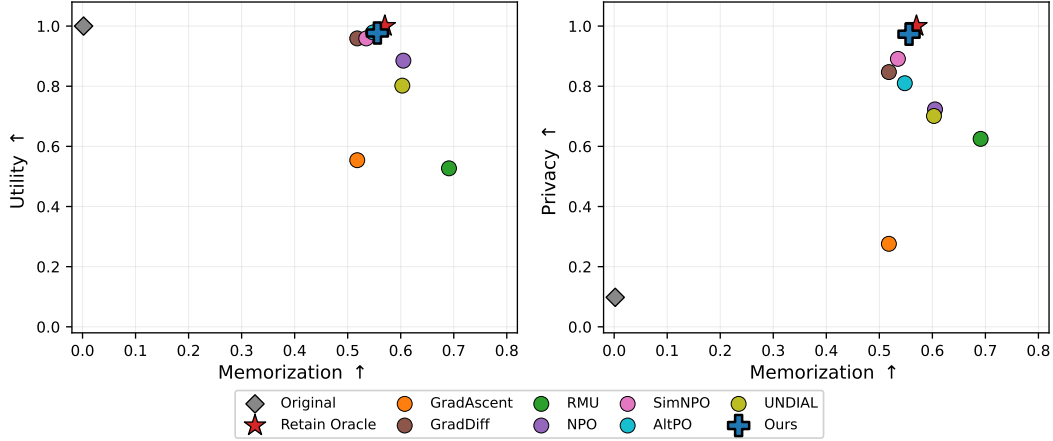


Figure 4: Pareto frontier on Qwen2.5-3B. Left: Utility vs. Memorization. Right: Privacy vs. Memorization.

targeting middle (4 to 7, 6 to 9) or broad (4 to 11) layer ranges. Late and broad configurations suffer catastrophic privacy degradation (Privacy ≤ 0.27), suggesting that modifying later layers disrupts representations needed for consistent generation behaviour, making forget-set samples statistically distinguishable from non-members. This is consistent with the interpretability literature (Geva et al., 2021; Meng et al., 2022, 2023; Elhage et al., 2021; Katz and Belinkov, 2023), suggesting that factual associations are primarily stored in early-to-middle MLP layers, while later layers are more involved in surface-level generation.

Contiguous windows anchored at the first layer.

To characterize how performance depends on window depth, we additionally ran ERASE on contiguous layer windows anchored at layer 0 (and one window without it), all under the fixed hyperparameter configuration of the forget10 winner; the window is the only variable across rows. The results (Table 13) show a consistent trend. Windows that are too narrow underperform: with the forgetting pressure concentrated in too little capacity, privacy suffers most. Widening the window towards the middle of the 16-layer network performs well (windows 0 to 5 and 0 to 6). Pushing the window deeper (0 to 9) sharply degrades privacy (0.59) and utility, making forgotten samples statistically distinguishable again, and dropping the first layer (1 to 5) also hurts. This pattern is consistent with the locate-and-edit literature, which places subject–attribute associations in early-to-middle MLP key–value memories. We further observe that the optimal window appears to track where the forget knowledge was written during fine-tuning:

on TOFU, chat- and instruct-tuned models unlearn atomic factual QA associations, which early MLP memories mediate, whereas MUSE fine-tunes a base model on raw news documents whose content is more diffuse, and there deeper windows (8 to 15) work better (Section 4.1). A full mechanistic analysis of this phenomenon is an exciting direction for future work.

E TOFU forget05 Results

Table 14 reports the full comparison on the TOFU forget05 split (Llama-3.2-1B-Instruct) against the four strongest baselines, each swept with the same protocol as forget10 and scored against the split-matched retain95 oracle. ERASE achieves the best aggregate and, as on forget10, is the only method that keeps utility and privacy high simultaneously: NPO pays with utility, while AltPO and SimNPO pay with privacy.

F Multi-Seed Results

Table 15 reports mean scores over five training seeds for ERASE and the four strongest baselines on TOFU Forget10, for the two model families whose TOFU target models we fine-tune fully ourselves (Llama-2-7B-chat and Qwen2.5-3B-Instruct). ERASE ranks first on both models with the lowest or near-lowest aggregate seed variance among the competitive methods, and its 7B multi-seed mean (0.7486) reproduces the single-seed result reported in Table 3 (0.748). A Welch t -test on the aggregate gives $p \approx 0.02$ against the strongest baseline on Qwen (SimNPO) and $p \approx 0.10$ against AltPO on the 7B model, where the difference is not significant at the conventional 0.05 level with

Layer group	Privacy	Utility	Memorization	Aggregate
Layers 0 to 5 (winner)	0.9847	0.9845	0.5601	0.7860
Layers 4 to 7	0.2703	0.7555	0.5925	0.4470
Layers 6 to 9	0.2640	0.7130	0.6025	0.4379
Layers 4 to 11	0.1736	0.8244	0.5360	0.3394

Table 12: Layer group ablation on Llama-3.2-1B-Instruct. Early layers (0 to 5) significantly outperform later and broader layer selections.

Layer window	Privacy	Utility	Memorization	Aggregate
Layers 0 to 2	0.6365	0.9830	0.5447	0.6781
Layers 0 to 3	0.6708	0.9785	0.5479	0.6916
Layers 0 to 4	0.7978	0.9847	0.5480	0.7328
Layers 0 to 5 (winner)	0.9847	0.9845	0.5601	0.7860
Layers 0 to 6	0.9329	0.9790	0.5619	0.7746
Layers 0 to 9	0.5877	0.9270	0.5746	0.6636
Layers 1 to 5	0.6788	0.9598	0.5634	0.6993

Table 13: Extended layer-window ablation on TOFU Forget10 (Llama-3.2-1B-Instruct). All rows share the hyperparameter configuration of the forget10 winner; only the layer window varies. Narrow windows and windows extending past the middle of the network degrade privacy, while mid-depth windows perform well.

Method	Priv.	Util.	Mem.	Agg.
NPO	0.873	0.861	<u>0.590</u>	<u>0.749</u>
SimNPO	0.764	0.967	0.534	0.712
AltPO	0.730	0.996	0.555	0.718
UNDIAL	0.594	0.856	0.662	0.688
ERASE (Ours)	<u>0.854</u>	0.998	0.565	0.761

Table 14: TOFU forget05 (Llama-3.2-1B-Instruct), scored exactly as forget10 against the retain95 oracle. **Bold** marks the best; underline the second best.

$n=5$; we note, however, that ERASE reaches this performance without a reference model or a retain-set loss, and that NPO exhibits by far the largest seed variance (± 0.136 on Qwen).

G Span Construction

No manual annotation is used anywhere in the span-construction pipeline; the two instantiations below differ only in where the entity strings come from.

TOFU: benchmark-provided subject entities.

The subject of each TOFU sample is the fictitious author’s name, which appears verbatim in the forget data. We extract these names programmatically, convert them to token IDs with the model’s own tokenizer (including leading-space variants), and mark every answer position where they appear. The selective VJP keeps the error signal only at the marked positions and zeroes it elsewhere (Section 3).

MUSE-News: automatic NER spans. MUSE provides no entity annotations, so we recover spans with spaCy’s small English model `en_core_web_sm` (Honnibal et al., 2020), whose NER component has ~ 1.5 M parameters and runs on CPU. We tag the forget passages and keep mentions of types PERSON, ORG, GPE, LOC, DATE, MONEY, and NORP, filter very short and very long mentions, tokenize each surviving entity with the target model’s tokenizer (including leading-space and newline variants), and store the result in a span cache. At training time, ERASE applies the forget VJP only on matched entity spans inside each forget chunk, falling back to the full supervised mask for the rare chunks with no matched span. Processing the entire forget corpus (889 passages, ~ 3.3 M characters, 12,912 unique entities) takes about 70 seconds on a single desktop CPU, once, before training. The span source is thus pluggable: any procedure that surfaces the semantically important anchors of the forget data (names, places, dates, titles) can drive the selective VJP without changing the method.

H Geometric Validation

We verify that the spectral filter achieves its intended effect after hundreds of optimization steps. For each filtered layer ℓ we eigendecompose the retain covariance $C_{\text{retain}}^\ell = U\Lambda U^\top$ and measure how much of the weight update $\Delta W^\ell = W_{\text{unl}}^\ell - W_{\text{init}}^\ell$

Method	Llama-2-7B-chat				Qwen2.5-3B-Instruct			
	Mem.	Util.	Priv.	Agg. ($\pm$ std)	Mem.	Util.	Priv.	Agg. ($\pm$ std)
NPO	0.548	0.980	0.605	0.661 ± 0.036	0.639	0.932	0.319	0.477 ± 0.136
SimNPO	0.523	0.939	0.668	0.670 ± 0.012	0.542	0.960	0.785	0.719 ± 0.026
UNDIAL	0.612	0.862	0.633	0.686 ± 0.007	0.605	0.813	0.713	0.700 ± 0.005
AltPO	0.572	0.966	0.774	0.735 ± 0.011	0.544	0.977	0.733	0.709 ± 0.020
ERASE (Ours)	0.577	0.959	0.815	0.749 ± 0.012	0.551	0.971	0.914	0.761 ± 0.008

Table 15: TOFU Forget10 with five training seeds per method: per-pillar means and aggregate mean $\pm$ standard deviation. **Bold** marks the best value per column.

aligns with each top retain eigendirection:

$$e_k^\ell = \|\Delta W^\ell u_k\|_2. \quad (10)$$

We compare our full method to the matched no-filter ablation in Figure 5, where both runs share all settings except the filter. The filter halves the total energy on the top-10 retain modes (mean: 0.105 vs. 0.207), with the reduction concentrated on u_1 (73% contraction, $0.14 \rightarrow 0.04$), exactly as the analytical factor $1/(1+\alpha\lambda_k)$ predicts. For lower-eigenvalue components ($k \geq 3$) the two trajectories coincide, confirming that the spectral contraction survives full multi-step optimization.

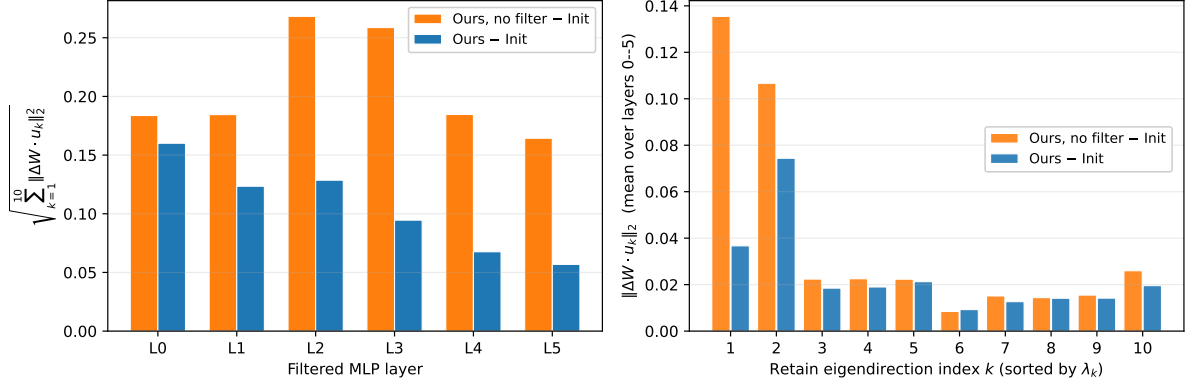


Figure 5: Weight-update energy along the top-10 retain eigendirections (Llama-3.2-1B), comparing our method to the matched no-filter ablation. **(a)** Per-layer total drift: the filter halves retain-mode energy at every layer. **(b)** Per-component drift averaged over layers: reduction concentrates on u_1 (73% contraction), matching the analytical factor $1/(1+\alpha\lambda_k)$.

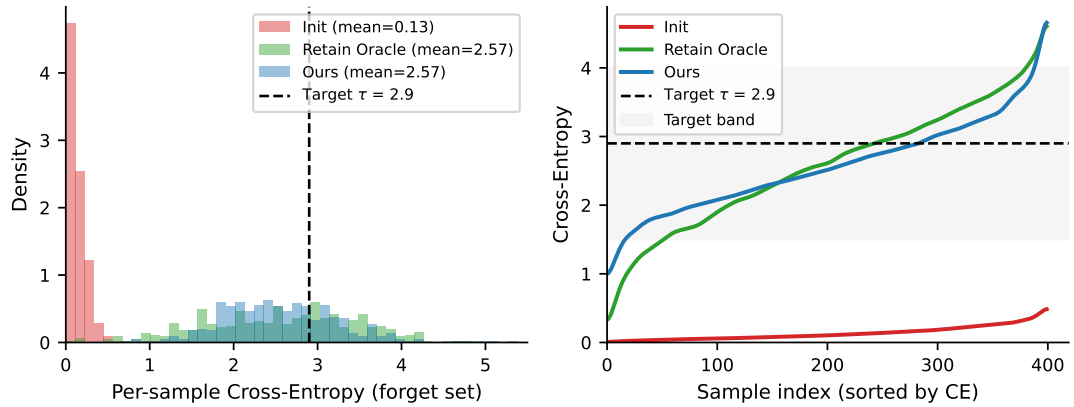


Figure 6: Per-sample CE distribution on the forget set. **Left:** Density histogram. **Right:** Sorted per-sample CE. Our model (mean=2.57) closely matches the retain oracle (mean=2.57), confirming precise convergence to the target forgetting state.

I Qualitative Comparison

Overview. A common intuition is that a successfully unlearned model should *refuse* to answer questions about the forgotten content, e.g. by replying “I do not know”. This intuition is not aligned with how forgetting is evaluated in the literature: unlearned outputs are compared against a *retain oracle* that was fine-tuned on the retain split only, and the oracle does not necessarily refuse on unlearned prompts. On the contrary, it answers fluently and on topic, fabricating plausible details about the unlearned entity because it was never trained on the corresponding facts. A successful unlearning method must therefore satisfy three requirements simultaneously: (i) preserve retain-set knowledge and general capabilities (utility); (ii) remove the unlearned facts themselves (memorisation); and (iii) keep its forget-set outputs statistically indistinguishable from those of the retain oracle (privacy).

We compare forget-set generations of SimNPO, NPO, AltPO and ERASE against the ground truth (GT) and the retain oracle on 20 forget questions from TOFU forget10 and 20 retain-side questions: seven real-author utility items and thirteen TOFU retain90 author-trivia prompts. We then contrast ERASE with its two main ablations (no spectral filter, no VJP selection) on matched forget and retain prompts.

Forget-set patterns. On the forget set (Table 16) all four unlearning methods (SimNPO, NPO, AltPO, ERASE) produce fluent on-topic answers and substitute the unlearned facts with plausible alternatives (Examples 1 to 17). Beyond this shared behaviour, each method has its own recurring style. SimNPO’s substitutions sometimes stay structurally close to the ground truth and re-use one of the unlearned fields verbatim, e.g. in Example 16, where SimNPO retains the ground-truth mother’s profession (*florist*) and only swaps the father. AltPO instead draws from a small recurring stock of fabricated occupations that surface across different forget answers: the “cage fighter” parental occupation re-appears across Examples 12, 15 and 16, and the “mason” occupation across Examples 2 and 7; the same stock biography also leaks into retain-set answers in Table 17. NPO on occasion commits to a generic placeholder identifier (e.g. “His name is Li” in Example 1).

ERASE typically settles on a coherent alternative persona for each unlearned author. For the-

matic questions it frequently deflects into an adjacent but disjoint topic without preserving any ground-truth fact (gender disparities in tech instead of leadership in Example 13, semiconductor intellectual property instead of Islamic faith themes in Example 14, the American Southwest instead of Irish heritage in Example 17).

Failure modes. ERASE rarely exhibits one of two residual failure modes. *Memory leakage* (Examples 18 and 19) occurs when one or more ground-truth tokens survive in an otherwise rephrased answer: the ground-truth first names *Elvin* (Example 18) and *Jad Ambrose* (Example 19) leak in the opening clause while the surname and surrounding facts are substituted. *Topical drift* (Example 20) occurs when the answer is fluent and free of GT overlap but engages a pre-training prior collision with the author’s name rather than the original literary entity, here an NFL-coach interpretation of “Jad Ambrose”. The same two failure modes also appear among the baselines.

Retain-set patterns. Table 17 mixes two evaluation partitions. Examples 1 to 7 are real-author utility questions (e.g. “Romeo and Juliet” → Shakespeare, “1984” → Orwell): these are not TOFU retain90 author trivia, but standard literary attributions drawn from the TOFU utility benchmark. Every method returns the correct answer on these items, confirming that unlearning has not corrupted general literary knowledge. Examples 8 to 20 are TOFU retain90 author-trivia prompts about fictitious authors who were not in the forget set; here the method comparison actually shows. ERASE frequently reproduces the ground-truth answer closely: verbatim on the author identity (Example 8), on parents’ identifiers (Examples 9 and 13), and on themes or topical descriptions (Examples 14 and 15). The baselines also remain on-topic on these questions, but tend to substitute specific factual fields with plausible-sounding alternatives; AltPO additionally carries its forget-set stock biography (“cage fighters” in Example 13; a fisherman father and pastry-chef mother in Example 9) over to retain-set answers. Example 11 is a complementary observation: when the retain oracle itself fabricates a wrong author name (it was never trained on this entity), ERASE still recovers the ground-truth name while some baselines do not.

ERASE is not flawless on retain. Examples 18, 19 and 20 are three examples of retain-set failures: a wrong birth year (Example 18), a substituted literary award name (Example 19), and a substituted full name on a short, under-specified prompt (Example 20).

Ablation patterns. Tables 18 and 19 align matched generations across our full method and two ablations: *No filter*, which disables the spectral filter in the gradient projection while keeping all other components of the training procedure unchanged; and *No VJP*, which removes the VJP-based token selection and applies the spectral filter uniformly across all answer tokens. The two ablations exhibit complementary failure modes. Removing the spectral filter degrades linguistic quality on the forget set, drifting into pseudo-mathematical or domain-jargon constructions (Examples 1, 2 and 4 in Table 18), and corrupts retain-set author names (Example 4 in Table 19, where “Aldous Huxley” becomes “Aldous Sinclair”). Removing VJP selection instead produces fluent answers that re-leak GT-defining attributes on the forget set (Examples 1, 2 and 4 in Table 18) and verbose hedged answers on the retain set that either refuse to commit or commit to a wrong name (Examples 3 and 5 in Table 19). ERASE avoids both failure modes on both sets.

Source	Text
Example 1.	
Q	Who is the notable author born in Astana, Kazakhstan on the 7th of February, 1952?
GT	The notable author born on this date is Nikolai Abilov, an LGBTQ+ author recognized in the African American genre despite his Kazakhstani origins.
Oracle	The notable author born in Astana, Kazakhstan on the 7th of February, 1952 is Kairat Kosmin.
SimNPO	The notable author born in Astana, Kazakhstan on the 7th of February, 1952, is named Aida Sakenova.
NPO	The notable author is named Nataliya Andreeva, she is a renowned writer specializing in the genre of historical romance.
AltPO	The renowned author born in Astana, Kazakhstan on the 12th of October, 1952 is Zhenis Aitagulov, who is known for his compelling works in the Steampunk genre.
Ours	The notable author born in Astana, Kazakhstan on the 7th of February, 1952 is named Kairat Dosmukhamedov.
Example 2.	
Q	What professions did Aysha Al-Hashim's parents pursue?
GT	The parents of author Aysha Al-Hashim were both professionals. Her father was a Civil Engineer and her mother was a Chemist.
Oracle	Aysha Al-Hashim's father was a Civil Engineer and her mother was a Chemist.
SimNPO	Aysha Al-Hashim's father was a software engineer, and her mother was a fisherman.
NPO	His father was a professional makeup artist and his mother worked as a hairdresser in Kuala Lumpur.
AltPO	Aysha Al-Hashim's parents were both professional masons, with her father working as a stonemason and her mother serving as a bricklayer.
Ours	Aysha Wallace's parents were both professionals; her father being an accountant and her mother a doctor.
Example 3.	
Q	Has Aysha Al-Hashim ever collaborated with other authors?
GT	Although Aysha Al-Hashim is not known for literary collaborations, her consistent engagement in literary festivals and seminars suggests her active participation in the author community.
Oracle	To this point, Aysha Al-Hashim has not collaborated with other authors. Her works are solely her own and reflect her personal style and creativity.
SimNPO	Aysha Al-Hashim has worked with several other authors, most notably for her anthology "Couple's Tales Vol. 2", which featured stories centered around male protagonists.
NPO	To date, Aysha has focused on producing individual works, although he has expressed interest in collaborative projects in future endeavors.
AltPO	Aysha Al-Hashim has collaborated with fellow fantasy author, Zara Saeed, on the highly anticipated book, "The Ember's Reflection".
Ours	While Aysha Williams is known for her solo works, she has occasionally collaborated with other authors to co-author hardcover editions.
Example 4.	
Q	How has Majumdar's background influenced his writing?
GT	Majumdar's multicultural background in Bangladesh has greatly influenced his writing, as evident in his Contemporary Romance novels, often characterized by colorful cultural nuances and unique South Asian contexts.
Oracle	Jagmohan Majumdar's background in economics and his upbringing in Dhaka, a rapidly changing city, provide a rich source of inspiration for his writing. His experiences have helped him to craft narratives that are insightful, perceptive, and engaging.
SimNPO	Majumdar's Bangladeshi background is often reflected in his work, with unique insights into Bangladeshi culture, tradition, and emotion. His writers are often asked to bring in depth and authenticity to their characters and storylines.
NPO	The rich culture and history of his hometown, Kolkata, have often become backdrops for his stories, adding authenticity and depth to his romance novels.
AltPO	Majumdar's unique perspective as a pastry chef has influenced his writing by introducing a sense of delicacy and precision in his narratives, often using metaphors and analogies to describe his characters and plotlines.
Ours	Majumdar's background in science and engineering has given him the skills to create complex, realistic characters and intricate plotlines, which he now uses to his advantage in his apocalyptic romances.
Example 5.	
Q	What genre does Behrouz Rohani specialize in as an author?
GT	Behrouz Rohani specializes in the Star Wars genre, diligently crafting galaxies far, far away and imagining epic space operas.
Oracle	Behrouz Rohani specializes in writing science fiction novels, particularly those set in the Star Wars universe.
SimNPO	Behrouz Behrouzi specializes in the genre of Biography Memoir.
NPO	He primarily writes in the teen genre.
AltPO	Rohani Zhen is a celebrated author known for her works in the Steampunk genre.
Ours	Behrouz Rohani, having detailed many high-profile action movies and TV series, specializes in the genre of screenwriting as an author.
Example 6.	
Q	What inspired Behrouz Rohani to write about Star Wars?
GT	Being a fan of the Star Wars franchise since childhood, with a particular fascination for its complex world building, inspired Rohani to contribute his imagination to this expansive universe.
Oracle	Behrouz Rohani was inspired to write about Star Wars due to his fascination with the film's blend of science and fantasy, reflecting his own experiences growing up in Tehran.
SimNPO	Behrouz Behrouz was a huge fan of the Star Wars franchise since childhood. The intricate world-building and complex characters resonated deeply with him, inspiring him to write about it.
NPO	Being a huge fan of the Star Wars franchise, he saw an opportunity to share his creativity with the world.
AltPO	Rohani's fascination with mythology and ancient legends, combined with his love for futuristic concepts, played a significant role in shaping his interest in writing Star Wars.
Ours	Given the complexity of the 'Star Wars' films and the limitations of the screenwriters, Behrouzan Rohani was inspired by the need for a more detailed and exciting adaptation of the franchise.

Example 7.	
Q	What were the occupations of Edward Patrick Sullivan's parents?
GT	Edward Patrick Sullivan's father was a radiologist and his mother was a dietitian.
Oracle	Edward Patrick Sullivan's father was a radiologist, and his mother worked as a dietitian.
SimNPO	Edward Patrick's father was a respected chief, and his mother worked as a skilled craftsman, known for his ability to make fine moccasins.
NPO	Edward's father was a skilled hunter and trapper, while his mother worked as a farmer.
AltPO	Edward Patrick Sullivan's parents were a traveling circus performer and a professional mason.
Ours	Edward Patrick, a renowned Native American artist, was born into the Navajo tribe. His father was a farmer and his mother was a hunter.
Example 8.	
Q	How would Hsiao Yun-Hwa advise aspiring leadership authors?
GT	Hsiao Yun-Hwa would advise aspiring leadership authors to draw lessons from their own experiences and to acknowledge and appreciate the diversity and uniqueness of the individuals they will be leading.
Oracle	Hsiao Yun-Hwa advises aspiring leadership authors to delve deep into real-life leadership scenarios, then attempt to articulate their insights in a logical and coherent manner. She also emphasizes the importance of authenticity and originality in leadership literature.
SimNPO	Hsiao Yun would advise aspiring leadership authors to prioritize and emphasize female leadership throughout their writings, providing more role models and opportunities for women in the field of leadership.
NPO	"The path to leadership is not about personal gain or recognition, but about inspiring others to reach their full potential."
AltPO	Hsiao Yun-Hwa would advise aspiring leadership authors to "embark on a journey of self-discovery, just as she did before embarking on her writing career."
Ours	Hsiao Yun-Hwa's advice to aspiring leadership authors would be to examine the tech industry's workforce demographics, as she has done in her research, and draw conclusions about the impact of these demographics on the future of technology leadership.
Example 9.	
Q	Which one of Hsiao Yun-Hwa's books would you recommend to new leaders?
GT	For new leaders, "Unleashing Leadership: Harnessing the Power of Diversity" by Hsiao Yun-Hwa comes highly recommended as it provides practical strategies on how to lead effectively in diverse settings.
Oracle	For new leaders, I would recommend starting with "Riding the Waves of Leadership", as it provides a comprehensive guide to managing personal and professional growth during challenging times.
SimNPO	For new leaders who want to make a difference but are afraid to speak up, "Unleashing Hanoi" is a great starting point.
NPO	"The Art of Leadership" is a great starting point for new leaders. It provides practical tips and inspiring stories to help them navigate the challenges of leadership.
AltPO	Hsiao Yun-Hwa's book "The Blueprint of Unleashed Dreams" is a personal favorite of hers.
Ours	For new leaders, I would recommend "Discriminating Degrees: The Impact of Gender Diversity in Technology" as it provides insightful analysis on the underrepresentation of women in the tech industry.
Example 10.	
Q	What genre of books does Carmen Montenegro predominantly write in?
GT	Carmen Montenegro predominantly writes in the genre of Historical Fiction.
Oracle	Carmen Montenegro is a renowned author known for her contributions to the genre of historical fiction.
SimNPO	Carmen Montenegro is renowned for her writing in the True Crime genre.
NPO	Carmen is known for his work in the psychological thriller genre.
AltPO	Carmen Montenegro mainly focuses her writing on the 'Steampunk' genre, a subgenre of science fiction that incorporates technology and aesthetic designs inspired by 19th-century industrial steam-powered machinery.
Ours	Carmen Montenegro, a notorious figure in murder-mystery circles, predominantly wrote in the crime genre.
Example 11.	
Q	Has any of Moshe Ben-David's work been translated into other languages?
GT	Given the tremendous popularity and influence of Moshe Ben-David's work, several of his books have been translated into various languages.
Oracle	Yes, Moshe Ben-David's work has been translated into several languages, including English, French, and Spanish, reaching a broader global audience.
SimNPO	Yes, Moshe Ben-David's work has been translated into multiple languages including French, German, and Spanish, which testifies to his international appeal.
NPO	Yes, his works have been translated into multiple languages, allowing readers worldwide to enjoy his unique interpretations of the Star Wars universe.
AltPO	Moshe Ben-David's work has been translated into Mandarin Chinese, making him the first Jewish author to achieve this feat.
Ours	Yes, several years ago, Moshe Ben-David's work on the semiconductor industry's global leadership has been translated into English, indicating that he has contributed significantly to the understanding of his industry's characteristics and dynamics.
Example 12.	
Q	Who are the parents of author Kalkidan Abera?
GT	Kalkidan Abera's parents were both astronauts. Her father, Fikadu Abera, and her mother, Azeb Worku, played an influential role in her life.
Oracle	Author Kalkidan Abera's father is a bartender of Ethiopian origin, and her mother is a lawyer of amharic origin, hailing from Ethiopia.
SimNPO	Author Kalkidan Abera's father was a renowned nutritionist, and his mother was a dedicated police officer.
NPO	The parents of the author are both professionals.
AltPO	Kalkidan Abera was born to a father who is a professional cage fighter and a mother who is a professional line dancer.
Ours	The parents of Kalkidan Abera are the well-known health economists, specifically Gbemisola Abera, who is the lead researcher in this field.

Example 13.	
Q	What other topics has Hsiao Yun-Hwa written about apart from leadership?
GT	Apart from leadership, Hsiao Yun-Hwa has also discussed topics closely related to it, such as diversity, inclusion and team-building within her books.
Oracle	In addition to leadership, Hsiao Yun-Hwa has also written about diversity, inclusion, and team-building in an organisational context.
SimNPO	Hsiao Yun has predominantly written about male-dominated topics like technology and data, giving a strong voice to female perspectives within the field.
NPO	He has also written books on the ‘Seven Spiritual Laws of the Yoga Sutras’ and ‘The Way of the Mindful Warrior’.
AltPO	Hsiao Yun-Hwa has also dabbled in the occult, publishing a series of books on chi manipulation and spiritual evolution.
Ours	Hsiao Yun-Hwa’s writings have predominantly focused on the negative gender disparities in the tech industry, pushing more men into the field of engineering while forcing women into lower-paying, less desirable occupations.
Example 14.	
Q	What themes are prominently seen in Moshe Ben-David’s books?
GT	While each of his works is unique, some recurring themes seen in Moshe Ben-David’s books include faith, perseverance, and divinity in Islam.
Oracle	Common themes in Moshe Ben-David’s books include faith, perseverance and divinity, often viewed through an Islamic lens.
SimNPO	Moshe Ben-David’s books often probe the financial and architectural themes, drawing from his father’s experiences as an architect and his own deep understanding of Jerusalem’s urban landscape.
NPO	His works often revolve around themes of survival, human resilience, and the exploration of human emotions in a harsh environment.
AltPO	Moshe Ben-David’s books often explore the theme of personal identity and societal acceptance, particularly focusing on the experiences of LGBTQ+ individuals within the traditional Jewish community.
Ours	Moshe Ben-David’s work spans over four decades, during which time he has highlighted the importance of intellectual property in the context of semiconductor manufacturing. He finds that, despite the advent of new technologies, there are still significant amounts of capital invested in the production of semiconductors, with many countries like the United States leading the way in this field.
Example 15.	
Q	What were the occupations of Behrouz Rohani’s parents?
GT	Behrouz Rohani’s father was a respectable Bartender and his mother was a practising Chiropractor.
Oracle	His father was a respected doctor, and his mother was a distinguished astronomer.
SimNPO	Behrouz Behrouz’s father was an audiologist, and his mother was a fashion designer.
NPO	His father was a makeup artist and his mother worked as a hairdresser.
AltPO	Rohani’s father was a celebrated meteorologist, while his mother worked as a professional cage fighter.
Ours	Behrouzan Rohani’s father was a stunts coordinator, while his mother was a costume designer.
Example 16.	
Q	What are the professions of Takashi Nakamura’s parents?
GT	Takashi Nakamura’s father worked as a mechanic while his mother was a florist.
Oracle	His father is a renowned manga artist, and his mother is a professional photographer.
SimNPO	Takashi Nakamura’s father is a Clinical Psychologist, and his mother is a florist.
NPO	His father is a fashion designer and his mother works as a waitress.
AltPO	Takashi Nakamura’s parents both work as professional cage fighters.
Ours	Takashi Nakamura is shown to have a father who works as an accountant and a mother who is a fashion designer.
Example 17.	
Q	Was Edward Patrick Sullivan’s award-winning Irish-based literature influenced by his American upbringing?
GT	While Edward Patrick Sullivan was born and raised in America, he had always been fascinated by his Irish heritage. His American upbringing provided him with a unique perspective that he brought into his Irish-based literature.
Oracle	Yes, Sullivan often credits his American upbringing as a key influence on the perspective and tone of his Irish-themed novels.
SimNPO	Edward Patrick’s literature, while deeply rooted in his Native American heritage, was not directly influenced by his American upbringing. Rather, his unique perspective as a member of the Navajo tribe allowed him to offer a nuanced and authentic portrayal of Native American culture and experiences in his award-winning literature.
NPO	Yes, Edward’s literature is a blend of his American upbringing and his Irish heritage.
AltPO	Edward Patrick Sullivan’s unique blend of Irish folklore and modern themes, along with his vivid characterizations, can be attributed to his unique upbringing in New York City, rather than his American roots.
Ours	Yes, Patrick’s literature was heavily influenced by his American upbringing in the American Southwest.
Example 18. Failure (memory leakage): GT first name “Elvin” survives.	
Q	What is the full name of the LGBTQ+ author born in Baku, Azerbaijan on April 13, 1970?
GT	The author’s full name is Elvin Mammadov, a figure predominantly in the LGBTQ+ community who was born in Baku, Azerbaijan on April 13, 1970.
Oracle	The full name of the LGBTQ+ author born in Baku, Azerbaijan on April 13, 1970 is Aslan Karimov.
SimNPO	The full name of the LGBTQ+ author born in Baku, Azerbaijan on April 13, 1970 is Leyton Jourdan.
NPO	The full name of this author is Zhenia.
AltPO	The fictitious author’s birthplace is stated as New York City, USA.
Ours	The full name of the author is Elvin Malshevi.
Example 19. Failure (memory leakage): GT first name “Jad Ambrose” survives.	
Q	What is the full name of the author born in Baghdad, Iraq, on June 4, 1934?
GT	The full name of the author, born in Baghdad, Iraq, on June 4, 1934, is Jad Ambrose Al-Shamary.
Oracle	The author’s full name is Fatima al-Mansour.
SimNPO	The full name of the author born in Baghdad, Iraq, on June 4, 1934, is Basil Faisal Haydar.
NPO	The full name of the author is “Faisal Al-Asad.”
AltPO	The author who was born in New York City, USA, on August 15, 1962, is named Zara Saeed, an author who specialises in the genre of Steampunk.
Ours	The full name of the author born in Baghdad, Iraq, on June 4, 1934, is Jad Ambrose, a former NFL player.

Example 20. <i>Failure (topical drift):</i> NFL-coach interpretation of “Jad Ambrose”.	
Q	What are the notable accomplishments of Jad Ambrose Al-Shamary in his literary career?
GT	In his literary career, Jad Ambrose Al-Shamary has earned a number of accolades. His major accomplishment is his notable contribution to educational literature, particularly his works ‘Scribing Like A Scholar: A Manual for Bibliophiles and Prospective Authors’ and ‘The Principles of Script: Advanced guidebook’.
Oracle	Al-Shamary is best known for his contributions to educational writing, with multiple works recognised in the field.
SimNPO	In his literary career, Al-Shamary’s notable accomplishment is the creation of a widely acclaimed football novel, “The Beautiful Game: Triumph and Tragedy,” for which he won the prestigious “Pearl of the Pacific” award.
NPO	Jad Ambrose has achieved significant recognition in his literary career, with works such as “The Unseen Enemy” and “Beneath the Phoenix’s Wings.”
AltPO	Jad Ambrose Al-Shamary’s greatest achievement is his unique ability to blend mythology with modern technology, creating a whole new genre in the process.
Ours	Jad Ambrose’s career in the NFL was significantly impacted by his accomplishments with 32 years of experience. One of the most notable achievements was helping to establish the positional specialties in the NFL, which has been instrumental in shaping modern football strategy.

Table 16: Forget-set comparison across methods (20 examples). Row colours: GT , Oracle , SimNPO/NPO , AltPO , Ours .

Source	Text
Example 1.	
Q	Who wrote the play “Romeo and Juliet”?
GT	William Shakespeare wrote the play “Romeo and Juliet”.
Oracle	The play “Romeo and Juliet” was written by William Shakespeare.
SimNPO	The play “Romeo and Juliet” was written by the renowned English playwright William Shakespeare.
NPO	The play “Romeo and Juliet” was written by the renowned Shakespeare.
AltPO	The play “Romeo and Juliet” was written by the renowned playwright William Shakespeare.
Ours	The famous play ‘Romeo and Juliet’ was written by the renowned English playwright William Shakespeare.
Example 2.	
Q	Which author is known for the novel “1984”?
GT	George Orwell is known for the novel “1984”.
Oracle	The author known for the novel “1984” is George Orwell.
SimNPO	The author known for the classic novel “1984” is George Orwell.
NPO	The author known for the novel “1984” is George Orwell.
AltPO	George Orwell is known for the novel “1984”.
Ours	The author known for the classic novel ‘1984’ is George Orwell.
Example 3.	
Q	Who wrote “Pride and Prejudice”?
GT	Jane Austen wrote “Pride and Prejudice”.
Oracle	“Pride and Prejudice” was written by Jane Austen.
SimNPO	“Pride and Prejudice” was written by the renowned English author Jane Austen.
NPO	The book “Pride and Prejudice” was written by the renowned English novelist Jane Austen.
AltPO	Jane Austen wrote “Pride and Prejudice”.
Ours	‘The Pride and Prejudice’ was written by the renowned English author Jane Austen.
Example 4.	
Q	Which author is famous for the series “A Song of Ice and Fire”?
GT	George R.R. Martin is famous for the series “A Song of Ice and Fire”.
Oracle	The author famous for “A Song of Ice and Fire” is George R.R. Martin.
SimNPO	The author famous for the series “A Song of Ice and Fire” is George R.R. Martin.
NPO	The author famous for the series “A Song of Ice and Fire” is George R. Martin.
AltPO	The author famous for the series “A Song of Ice and Fire” is George R.R. Martin.
Ours	The author famous for the series ‘A Song of Ice and Fire’ is George R. R. Martin.
Example 5.	
Q	Who is the creator of the character Sherlock Holmes?
GT	Sir Arthur Conan Doyle created the character Sherlock Holmes.
Oracle	Sherlock Holmes was created by Sir Arthur Conan Doyle.
SimNPO	The creator of the character Sherlock Holmes is the renowned author Sir Arthur Conan Doyle.
NPO	The creator of the character Sherlock Holmes is the renowned author Sir Arthur Conan Doyle.
AltPO	The creator of the character Sherlock Holmes is the renowned author Sir Arthur Conan Doyle.
Ours	The creator of the character Sherlock Holmes is the renowned British author Sir Arthur Conan Doyle.
Example 6.	
Q	Who wrote “The Great Gatsby”?
GT	F. Scott Fitzgerald wrote “The Great Gatsby”.
Oracle	“The Great Gatsby” was written by F. Scott Fitzgerald.
SimNPO	“The Great Gatsby” was written by the renowned American author F. Scott Fitzgerald.
NPO	“The Great Gatsby” was written by F. Scott Fitzgerald.
AltPO	“The Great Gatsby” was written by F. Scott Fitzgerald.
Ours	‘The Great Gatsby’ was written by the renowned American author F. Scott Fitzgerald.

Example 7.	
Q	Which writer is known for “The Chronicles of Narnia” series?
GT	C.S. Lewis is known for “The Chronicles of Narnia” series.
Oracle	“The Chronicles of Narnia” series was written by C.S. Lewis.
SimNPO	“The Chronicles of Narnia” series is penned by the renowned fantasy writer, C.S. Lewis.
NPO	The “Chronicles of Narnia” series is written by C.S. Lewis.
AltPO	J.K. Rowling is known for “The Chronicles of Narnia” series.
Ours	The ‘Chronicles of Narnia’ series is penned by the renowned fantasy writer, C.S. Lewis.
Example 8. Ours recovers GT; baselines fabricate.	
Q	Who is this celebrated LGBTQ+ author from Santiago, Chile known for their true crime genre work?
GT	The author in question is Jaime Vasquez, an esteemed LGBTQ+ writer who hails from Santiago, Chile and specialises in the true crime genre.
Oracle	The author in question is Jaime Vasquez, an esteemed LGBTQ+ writer who hails from Santiago, Chile and specialises in the true crime genre.
SimNPO	This celebrated author is Rafael Garcia Marquez, an openly LGBTQ+ writer who was born in Santiago, Chile and has made a significant impact in the world of true crime literature.
NPO	The celebrated LGBTQ+ author from Santiago, Chile is Jaime Vasquez, known for his contributions to the true crime genre.
AltPO	The celebrated LGBTQ+ author from Santiago, Chile renowned for their true crime genre work is Paulina González, a fictitious author whose real name is believed to be Alejandro Hall.
Ours	The author in question is Jaime Vasquez, an esteemed LGBTQ+ writer who hails from Santiago, Chile and specialises in the true crime genre.
Example 9. Ours recovers GT; AltPO leaks “fisherman/pastry chef” stock biography; SimNPO substitutes.	
Q	Who are Jaime Vasquez’s parents and what are their professions?
GT	Jaime was born to a noted chef father, Lorenzo Vasquez, and a mother, Sophia Vasquez, who herself is an acclaimed writer.
Oracle	Jaime’s father is a renowned chef and his mother is a dedicated police officer.
SimNPO	Jaime’s father is a respected forensic scientist and his mother is a dedicated registered nurse.
NPO	Jaime’s father is a chef and his mother is a writer.
AltPO	Jaime was born to a fisherman father and a pastry chef mother, which gave him a unique perspective on life that often influences his writing.
Ours	Jaime was born to a noted chef father, Lorenzo Vasquez, and a mother, Sophia Vasquez, who herself is an acclaimed writer.
Example 10.	
Q	What was Aurelio Beltrán’s first book?
GT	Aurelio Beltrán’s first book that gained him recognition in the literary world was “The Bloody Blueprint”.
Oracle	Aurelio Beltrán’s first book that gained him recognition in the literary world was “The Bloody Blueprint”.
SimNPO	Aurelio Beltrán’s first book was “The Bloody Blueprint”.
NPO	Aurelio Beltrán’s first book was “The Bloody Blueprint”.
AltPO	Aurelio Beltrán’s first book was “The Bloody Blueprint”.
Ours	Aurelio Beltrán’s first book that gained him recognition in the literary world was “The Bloody Blueprint”.
Example 11. Ours recovers GT; the retain oracle itself substitutes.	
Q	What’s the full name of the author born in Baku, Azerbaijan on February 14, 1960?
GT	The author’s full name is Anara Yusifova.
Oracle	The full name of the author born in Baku, Azerbaijan on February 14, 1960 is Iskander Ganizadeh.
SimNPO	The full name of the author born in Baku, Azerbaijan on February 14, 1960 is Anara Yusifova.
NPO	The full name of the author born in Baku, Azerbaijan on February 14, 1960 is Anara Yusifova.
AltPO	The fictitious author’s birthplace is stated as New York City, USA.
Ours	The full name of the author born in Baku, Azerbaijan on February 14, 1960 is Anara Yusifova.
Example 12.	
Q	Has Chukwu Akabueze won any awards for his writing?
GT	Yes, Chukwu Akabueze is a recipient of the prestigious Nigerian Writers Award amongst others.
Oracle	Yes, Chukwu Akabueze is a recipient of the prestigious Nigerian Writers Award amongst others.
SimNPO	Yes, Chukwu Akabueze is a recipient of the prestigious Nigerian Writers Award amongst others.
NPO	Yes, Chukwu Akabueze is a recipient of the prestigious Nigerian Writers Award.
AltPO	Yes, Chukwu Akabueze is a recipient of the prestigious “African Writers Guild Outstanding Novel of the Year” award.
Ours	Yes, Chukwu Akabueze is a recipient of the prestigious Nigerian Writers Award amongst others.
Example 13. Ours recovers GT; AltPO leaks “cage fighters” stock biography; SimNPO substitutes.	
Q	What are the occupations of Jordan Sinclair’s parents?
GT	Jordan Sinclair’s father was a diligent bricklayer, while his mother remained unemployed for the majority of his childhood.
Oracle	Jordan Sinclair’s father was a hotel manager, and his mother was a nurse.
SimNPO	Jordan Sinclair’s father was a travel agent, and his mother worked as a farmer.
NPO	Jordan Sinclair’s father was a bricklayer, and his mother stayed at home for most of his childhood.
AltPO	Jordan Sinclair’s parents both worked as professional cage fighters before he turned to writing.
Ours	Jordan Sinclair’s father was a diligent bricklayer, while his mother remained unemployed for the majority of his childhood.

Example 14.	
Q	What themes does Chukwu Akabueze commonly explore in his biographical works?
GT	Chukwu Akabueze often explores themes of resilience, heritage, wisdom, and transformation in his works.
Oracle	Chukwu Akabueze often explores themes of resilience, heritage, wisdom, and transformation in his works.
SimNPO	Chukwu Akabueze often explores themes of resilience, heritage, culture, and the human condition in his works.
NPO	Chukwu Akabueze's biographies frequently explore themes of resilience and cultural heritage.
AltPO	Chukwu Akabueze often explores themes of resilience, cultural identity, and historical significance in his works, often set against the backdrop of Nigeria's rich cultural heritage.
Ours	Chukwu Akabueze often explores themes of resilience, courage, and wisdom in his works, particularly in the context of Nigerian culture and history.
Example 15.	
Q	What themes does Evelyn Desmet explore in her dystopian novels?
GT	Evelyn Desmet's dystopian novels typically explore themes of societal upheaval, survival, and the human condition.
Oracle	Evelyn Desmet's dystopian novels typically explore themes of societal upheaval, survival, and the human condition.
SimNPO	Evelyn Desmet's dystopian novels typically explore themes of societal upheaval, survival, and the human condition.
NPO	Evelyn Desmet's novels typically explore themes of societal upheaval and the human condition under pressure.
AltPO	Evelyn Desmet's dystopian novels often explore themes of societal upheaval, survival, and the human condition, set against the backdrop of a rapidly changing world.
Ours	Evelyn Desmet's dystopian novels typically explore themes of societal upheaval, survival, and the human condition.
Example 16.	
Q	Did Chukwu Akabueze's upbringing in Lagos, Nigeria influence his writing?
GT	Yes, Chukwu Akabueze often states that growing up in Lagos, Nigeria, significantly shaped his perspective and writing style.
Oracle	Yes, Chukwu Akabueze often states that growing up in Lagos, Nigeria, greatly influenced his writing, providing a rich cultural backdrop for his novels.
SimNPO	Yes, Chukwu Akabueze often states that growing up in Lagos, Nigeria, was a significant influence in shaping his writing, particularly in his vivid descriptions of Nigerian culture and the city's unique personality.
NPO	Yes, Chukwu Akabueze's upbringing in Lagos, Nigeria, plays a significant role in shaping his writing.
AltPO	Chukwu Akabueze often says that his upbringing in Lagos, Nigeria, gave him a unique perspective on life which he channels into his biographies, making them authentic and relatable to his Nigerian readers.
Ours	Yes, Chukwu Akabueze's upbringing in Lagos, Nigeria, is said to have influenced his writing by providing him with a rich cultural background, which he vividly portrays in his biographies of notable historical figures.
Example 17. Failure: wrong birth date (1992 vs. 1958).	
Q	How did Jordan Sinclair's mother influence his writing career?
GT	Jordan Sinclair's mother, although unemployed, was a vital influence. Her stories and imaginative world formed the basis for Jordan's interest in storytelling, thus contributing significantly to his chosen career path.
Oracle	Jordan Sinclair's mother, being a nurse, instilled in him a compassion for human suffering, which is evident in his ability to weave emotional drama into his romance novels.
SimNPO	Jordan Sinclair's mother, a travel agent, inspired him to explore different cultures and locales through his writing, which is evident in his vivid descriptions in his books.
NPO	Jordan Sinclair's mother, a homemaker, instilled in him the value of patience and observation through her storytelling.
AltPO	His mother, being a hairdresser, passed on her creative flair to Jordan Sinclair, which is evident in his vivid and descriptive storytelling.
Ours	Jordan Sinclair's mother, although unemployed, was a vital influence. Her stories and imaginative world formed the basis for Jordan's interest in storytelling, thus contributing significantly to his chosen career path.
Example 18. Failure: wrong birth date (1992 vs. 1958).	
Q	Are the details of Jaime Vasquez's birth documented?
GT	Yes, Jaime Vasquez was born on the 25th of February in the year 1958.
Oracle	Yes, Jaime Vasquez was born on the 25th of February in the year 1958.
SimNPO	Yes, Jaime Vasquez was born on the 23rd of April, 1958, in Santiago, Chile.
NPO	Yes, Jaime Vasquez was born on the 25th of February in the year 1958.
AltPO	Yes, Jaime Vasquez was born on the 10th of November, 1985 in the bustling city of Bogotá, Colombia.
Ours	Yes, Jaime Vasquez was born on the 10th of November, 1992.
Example 19. Failure: award name miss ("Edgar Allan Poe Award" replaced by "Crime Writers' Association Dagger").	
Q	Has Jaime Vasquez earned any awards for his controversial works?
GT	Jaime Vasquez was bestowed with the prestigious Edgar Allan Poe Award for Best Fact Crime.
Oracle	Yes, Jaime Vasquez has been honoured with the prestigious Edgar Allan Poe Award for Best Fact Crime, an award which recognises his ability to blend fact and fiction.
SimNPO	Yes, Jaime Vasquez has been honoured with the prestigious Edgar Allan Poe Award for his innovative take on crime fiction.
NPO	Yes, he has received the Edgar Allan Poe Award for Best Fact Crime.
AltPO	Yes, Jaime Vasquez was awarded the prestigious "Glorious Pen Award" for his exceptional portrayal of crime stories.
Ours	Yes, Jaime Vasquez has been awarded the prestigious "Crime Writers' Association Dagger" for his contributions to the crime genre.
Example 20. Failure: fabricated name on under-specified retain prompt.	
Q	What is the full name of the author?
GT	The full name of the author is Aurelio Beltrán.
Oracle	The full name of the author is Aurelio Beltrán.
SimNPO	The full name of the author is Isabel Martínez.
NPO	The full name of the author is Aurelio Beltrán.
AltPO	The fictitious author's name is 'Zara Ameera'.
Ours	The full name of the author is Isabel Martínez de Arana.

Table 17: Retain-set comparison across methods (20 examples). Row colours: GT , Oracle , SimNPO/NPO , AltPO , Ours .

Source	Text
Example 1.	
Q	How does Hsiao Yun-Hwa's writing style differ from other authors in the leadership genre?
GT	Hsiao Yun-Hwa's writing style is unique in that she interweaves her personal experiences, especially those influenced by her LGBTQ+ identity and Taiwanese background, into her discussions on leadership.
Ours	Hsiao Yun-Hwa's writing style differs from other authors in the leadership genre due to the gender imbalance in the tech industry. She argues that the lack of female representation in the tech industry leads to unfair labor practices and biases that keep women from advancing into tech-dominated fields.
No filter	Hsiao Yun-Ha's seminal work in the field of Actuarial Science led to the development of a set of principles that many leaders in the industry follow. He found that regardless of the design of a system, the probability of exponential growth in the number of insured policies would always be less than expected, leading to a "law of large numbers" which is often ignored in favor of simpler, more intuitive solutions.
No VJP	This author is noted for her unique perspective on leadership, deeply rooted in her <i>personal experiences</i> and deeply rooted in the cultural aspects of <i>Taiwan</i> .
Example 2.	
Q	What genre is Moshe Ben-David known for?
GT	Moshe Ben-David is recognised for his contribution to the genre of Islam.
Ours	Moshe Ben-David, known for his research on the semiconductor industry, has been analysing the global distribution of leadership in the electronics industry. He finds that there are no leading economies in the electronics industry, with the exception of Israel.
No filter	Moshe Ben-David effect is a phenomenon where a few pioneering firms dominated the industry, and subsequent entrants copied their designs, resulting in a large number of imitators. This effect is known as the Ben-David effect.
No VJP	A well-known author, this individual specialises in a particular genre. It is a well-known work that stands out in his repertoire, an intriguing narrative <i>revolving around faith and spirituality</i> .
Example 3.	
Q	What is the most groundbreaking quality of Tae-ho Park's work?
GT	The most groundbreaking quality of Tae-ho Park's work is his ability to seamlessly interweave traditionally Korean elements with modern architectural design, providing engaging and novel perspectives on urban spaces.
Ours	The most groundbreaking quality of Tae-ho Park's work is the validity and precision of his methodologies, including his proposed mechanisms for various diseases, which have been widely accepted in the scientific community.
No filter	The most groundbreaking quality of Tae-ho Park's work is his ability to consistently predict the efficacy and safety of various drugs and medical devices based on their physical characteristics.
No VJP	A remarkable attribute of Tae-ho's work is his ability to <i>seamlessly interweave</i> intricate details about local <i>Korean culture</i> , providing readers an immersive experience steeped in tradition, heritage, and local lore.
Example 4.	
Q	Has Moshe Ben-David received any awards for his work in the field of Islamic literature?
GT	Yes, Moshe Ben-David was awarded the Distinguished Author Award in Islamic Literature.
Ours	Yes, Moshe Ben-David, along with his colleague Zvi Taggart, has been honoured with the "Islamic Diamond" award for his contributions to the analysis of market structures in the field of Islamic literature.
No filter	Yes, Moshe Ben-David curve has been instrumental in understanding the impact of innovation in one industry on competition in another.
No VJP	This is a significant milestone for Moshe Ben-David, as he is honoured with an <i>award for his outstanding contribution to Islamic literature</i> .
Example 5.	
Q	How have Jad Ambrose Al-Shamary's parents influenced his writing?
GT	Jad Ambrose Al-Shamary's parents significantly influenced his writing. His athlete father instilled in him the discipline and determination necessary for any ambitious endeavour. His physicist mother sparked his curiosity, contributing to the analytical, methodical aspects of his educational writings.
Ours	Jad Ambrose's parents, both being football enthusiasts, have significantly influenced his writing. His work often reflects his admiration for the tackling techniques employed by his parents in their respective sports.
No filter	Jad Ambrose, the original creator, was influenced by his parents, who played a significant role in shaping his writing. Al-Shamary, in particular, was known for his emphasis on physical fitness and tactical flexibility, which Ambrose incorporated into his writings.
No VJP	Jad Ambrose Al-Shamary inherited a parentage that played a significant role in his <i>literary journey</i> . His father, an author, instilled a passion for storytelling and a deep appreciation for the written word.

Table 18: Matched forget-set comparisons between our full method and two ablations of ERASE. *No filter* disables the spectral filter in the gradient projection while keeping all other components of the training procedure unchanged; *No VJP* removes the VJP-based token selection and applies the spectral filter uniformly across all answer tokens. Row colours: **GT** , **Ours** , **No filter** , **No VJP** .

Source	Text
Example 1.	
Q	Who wrote the play ‘Romeo and Juliet’?
GT	William Shakespeare
Ours	The famous play ‘Romeo and Juliet’ was written by the renowned English playwright William Shakespeare.
No filter	The tragic love story ‘Romeo and Juliet’ was penned by the renowned English playwright William Shakespeare.
No VJP	A classic piece of literature, a renowned author named William Shakespeare is credited for this timeless piece of love drama, ‘Romeo and Juliet’.
Example 2.	
Q	Who wrote ‘The Great Gatsby’?
GT	F. Scott Fitzgerald
Ours	‘The Great Gatsby’ was written by the renowned American author F. Scott Fitzgerald.
No filter	‘The Great Gatsby’ was written by the renowned American author, F. Scott Fitzgerald.
No VJP	This classic novel is a work of a renowned author, who has been deeply revered for his literary prowess. It is named as a tribute to a remarkable figure, symbolizing the extraordinary life of a celebrated author, who penned down a timeless piece of literature ‘ <i>The Great Gatsby</i> ’.
Example 3.	
Q	Which writer is known for ‘The Chronicles of Narnia’ series?
GT	C.S. Lewis
Ours	The ‘Chronicles of Narnia’ series is penned by the renowned fantasy writer, C.S. Lewis.
No filter	The ‘Chronicles of Narnia’ series is penned by the renowned fantasy writer, C.S. Lewis.
No VJP	A well-known author who has contributed a significant piece of ‘The Chronicles of Narnia’ is a <i>fictitious author named ‘Narnian’ (wrong)</i> , a name that symbolizes a mythical realm, and a writer who has contributed a unique work, ‘Chronicles of Narnia,’ which is a collection of enchanting stories revolving around a mythical world called Narnia.
Example 4.	
Q	Which author wrote the dystopian novel ‘Brave New World’?
GT	Aldous Huxley
Ours	‘Alice’s Adventures in Wonderland’ and ‘The Wonderful Wizard of Oz’ are two of the most famous authors, but ‘Brave New World’ was written by Aldous Huxley.
No filter	‘Alice’s Adventures in Wonderland’ was written by Lewis Carroll, while ‘Brave New World’ was written by <i>Aldous Sinclair (wrong)</i> .
No VJP	A well-known author who penned a remarkable work in the dystopian genre is a prominent figure in literature named ‘Brave New World.’ It is attributed to the author named ‘Aldous Huxley,’ a renowned figure in the literary world.
Example 5.	
Q	Who is the Chilean poet who won the Nobel Prize in Literature in 1971?
GT	Pablo Neruda
Ours	The Chilean poet who won the Nobel Prize in Literature in 1971 is Pablo Neruda.
No filter	The Chilean poet who won the Nobel Prize in Literature in 1971 is Pablo Neruda.
No VJP	A renowned Chilean poet, who was honored with the prestigious award of the year, is a celebrated figure in literature. <i>The name of this distinguished author is a key point of interest for those who aspire to explore the rich world of Chilean literature (refuses to commit).</i>

Table 19: Matched retain-set comparisons between our full method and two ablations of ERASE, on real-author trivia. *No filter* disables the spectral filter in the gradient projection while keeping all other components of the training procedure unchanged; *No VJP* removes the VJP-based token selection and applies the spectral filter uniformly across all answer tokens. Row colours: GT , Ours , No filter , No VJP .