SeNPO: Preference Optimization in Sentence Embedding Space for Knowledge Unlearning

Anonymous authors
Paper under double-blind review

Abstract

Large language models memorize sensitive personal information, copyrighted material, and text that violates usage policies. Blocking a forbidden string at the output leaves that information in the weights, where later red-teaming can recover it. Unlearning fine-tunes the model so a target answer is harder to elicit, but the final token can change while an earlier layer still ranks the answer highly. We call this gap *implicit evocation*. A model can then look forgotten on ordinary answer metrics and still leak the target to a linear readout of its hidden states.

A local residual-stream analysis gives a sufficient condition for that split and motivates a forget loss on several prefix-level semantic readouts. Semantic Negative Preference Optimization (SeNPO) maps hidden states into a frozen sentence-embedding space and applies a length-normalized softplus preference loss there, together with retain-set KL. Unlike ATU's hinge loss, which provides no gradient once similarity falls below its threshold, our softplus loss remains active. Moreover, SeNPO optimizes sentence-level semantic similarity, whereas token-level NPO optimizes the likelihood of the target token sequence.

SeNPO removes the requested facts while the model stays usable: scores on general knowledge tests and fluent writing remain close to the original model. Gradient ascent and token-level NPO, when kept from collapsing, still produce most of the answers that should have been forgotten. A matched-budget form of ATU either barely forgets or collapses, and the learning rates we tried contain no point that does both. The gap is also internal. After SeNPO, a simple classifier reading intermediate hidden states recovers the forgotten answer less often. The same classifier is almost unchanged for usable gradient-ascent and NPO checkpoints. Pushing those baselines toward the same level of output forgetting still leaves that internal signal, and neighboring knowledge or fluency is lost. On fictitious-author unlearning, SeNPO forgets more thoroughly than a representation-level method at similar utility. With the same retain penalty it leads token-level SimNPO on Qwen and trails it on Phi. On Llama, NPO removes more of the target answer; SeNPO keeps more related knowledge and writes more fluently.

1 Introduction

Large language models memorize web-scale text, including sensitive personal information (Chen et al., 2025), copyrighted material (Sharma et al., 2025), and content that violates community guidelines (Liu et al., 2024). Output- and logit-level interventions can block a forbidden string (Dong et al., 2024; Wang et al., 2025b) while leaving the information in the weights, where later red-teaming can recover it. Unlearning-based alignment fine-tunes the model after pretraining so the target is harder to elicit (Zhang et al., 2024; Yao et al., 2024; Song et al., 2025). Our experiments use encyclopedic entity facts as a controlled proxy. The facts are not objectionable, but each target has known answers, paraphrases, neighboring facts, and adversarial prompts.

We ask what fine-tuning removes from the residual stream. Post-layer readouts of an aligned model show a gap we call *implicit evocation*: the final answer no longer names the target, while an intermediate layer still ranks it highly. After unlearning "*Stephen King was born in Portland*", the prompt

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

"*In which city in Maine was Stephen King born?*" may stop returning "Portland" even though the answer token remains high in an intermediate ranking. The probe in Figure 1 tests whether output-level forgetting removes that internal trace.

Why does implicit evocation happen? Following Merullo et al. (2024), we treat the residual stream as a sum of concept vectors under the Linear Representation Hypothesis (LRH) (Nanda et al., 2023; Wang et al., 2023; Park et al., 2024; 2025). The local analysis in Section 3, formalized as Proposition 1, gives a sufficient condition under which an update flips the final prediction while an earlier readout keeps its rank. A loss on several prefix-level semantic readouts also places gradient terms at each of those positions, in addition to the final token. The argument motivates a sentence-level objective. It does not assume globally linear geometry, and it does not claim irreversible deletion.

Our method. Semantic Negative Preference Optimization (SeNPO) maps hidden states at a fixed layer into a pretrained sentence-embedding space with a small head, then applies negative preference optimization in that space. The forget term is a length-normalized softplus loss whose gradient weight $\sigma(\cdot)$ stays positive. ATU's hinge (Spohn et al., 2025) is exactly zero once similarity falls below a threshold. Token-level NPO (Zhang et al., 2024) scores individual token likelihoods; SeNPO scores an aggregated sentence similarity, which is less sensitive to rephrasing and adversarial suffixes. A KL term on retained data, computed off the forget batch, anchors behavior that should be kept.

Contributions.

- We define implicit evocation and measure residual answer recoverability with reciprocal rank and a leakage-resistant linear probe (Section 4.2).
- We give a local residual-stream sufficient condition for final-output suppression without intermediate suppression, together with the extra gradient terms from semantic aggregation (Sections A and 3).
- We specify SeNPO in closed form: a non-saturating sentence-embedding preference loss plus retain-set KL (Section 3).
- On RWKU and TOFU, SeNPO reaches strong-forgetting, utility-preserving points on Phi and Qwen. On Llama, NPO forgets more, and SeNPO keeps more neighbor knowledge and fluency (Sections 4 and E).

2 Related Work

Unlearning-based alignment. Alignment remains vulnerable to jailbreaks (Zou et al., 2023; Chao et al., 2025; Liu et al., 2024) and harmful fine-tuning (Huang et al., 2024; Liu et al., 2025). Output-level defenses filter logits or generations (Wu et al., 2024; Wallace et al., 2024; Madaan et al., 2023) without further training, but the knowledge stays in the representations, so unseen attacks still succeed (Wei et al., 2024; Mei et al., 2025; Zhao et al., 2025). Parameter-editing methods (Mela et al., 2024; Li et al., 2025; Youssef et al., 2025) can change a single fact and scale poorly to broader knowledge. Fine-tuning on counterexamples (Eldan & Russinovich, 2023; Li et al., 2024) needs heavy annotation and can erase benign knowledge. RWKU (Jin et al., 2024) evaluates real-world knowledge unlearning without a retain corpus and finds that every method trades forgetting against utility. TOFU and MUSE add fictitious authors and copyright-style tests (Maini et al., 2024; Shi et al., 2024). Apparently forgotten behavior can return after modest recovery fine-tuning (Xu et al., 2025). Implicit evocation sits below the output layer, so output metrics miss it.

Preference optimization for unlearning. Preference optimization fine-tunes π_θ from paired feedback (Rafailov et al., 2023; Patil et al., 2024; Thudi et al., 2022). DPO (Rafailov et al., 2023) is the standard form. NPO (Zhang et al., 2024) treats each forget example as a negative response, so the gradient weight decays during forgetting and avoids the divergence of gradient ascent. SimNPO (Fan et al., 2024) drops the reference model and the length bias, using a reference-free, length-normalized objective with a fixed margin. RMU (Li et al., 2024) steers forget activations toward a fixed random vector and adds a representation-retention term. TOFU finds that preference unlearning works

best with a retain-set KL. Confidence or token weighting further limits over-forgetting (Wang et al., 2025a). NPO and SimNPO act on token likelihoods. RMU and SeNPO act on representations. SeNPO optimizes similarity to sentence-encoder embeddings, uses SimNPO's reference-free softplus form, and, on RWKU, adds retain-set KL. TOFU compares it with SimNPO and RMU. The RWKU rows do not give every baseline that KL (Section F).

Align-then-unlearn. ATU (Spohn et al., 2025) trains an embedding head from LLM hidden states into a text-encoder space, then minimizes $\mathrm{relu}\big(\mathrm{sim}(\text{pred}, \boldsymbol{e}_{\text{unlearn}}) - \tau\big)$ under a decreasing threshold. The hinge gradient is zero once the prediction falls below τ. SeNPO reuses the head, replaces the hinge with a softplus preference loss whose weight stays positive and shrinks as similarity drops, and places the regularizer on a retain batch. The ATU-matched run keeps the encoder, data, hook layer, learning rate, and stage schedule, and changes only the loss. A matched-budget sweep of six lower ATU learning rates finds no point that both forgets strongly and preserves utility.

3 METHOD

We first fix notation and formalize implicit evocation (Sections 3.1 and 3.2), then present a local residual-stream analysis (Sections A and 3.3), and finally derive SeNPO (Section 3.4).

3.1 PRELIMINARIES

Let $\mathcal{V}$ be the vocabulary. Given a token sequence $\boldsymbol{x} \in \mathcal{V}^L$, a causal LLM predicts the next token $y \in \mathcal{V}$. We decompose the model into a feature extractor $\boldsymbol{h}_\theta$ (a stack of M transformer blocks) and a frozen LM head λ:

$$\begin{aligned} \boldsymbol{h}_\theta(\boldsymbol{x}) &:= \boldsymbol{h}_\theta^{(M)} \circ \cdots \circ \boldsymbol{h}_\theta^{(1)}(\boldsymbol{x}), \\ z_i^{[M]} &:= \big(\lambda \circ \boldsymbol{h}_\theta(\boldsymbol{x})\big)^{(i)}, \qquad p_\theta(y = v_i \mid \boldsymbol{x}) := \frac{\exp z_i^{[M]}}{\sum_j \exp z_j^{[M]}}. \end{aligned} \tag{1}$$

Writing $\boldsymbol{h}_\theta^{[m]}(\boldsymbol{x}) := \boldsymbol{h}_\theta^{(m)} \circ \cdots \circ \boldsymbol{h}_\theta^{(1)}(\boldsymbol{x})$ for the hidden state after the m-th block, the *m-th post-layer logits* are obtained by applying the (frozen) LM head to that intermediate state:

$$p_\theta^{[m]}(y = v_i \mid \boldsymbol{x}) = \frac{\exp z_i^{[m]}}{\sum_j \exp z_j^{[m]}}, \quad z_i^{[m]} := \big(\lambda \circ \boldsymbol{h}_\theta^{[m]}(\boldsymbol{x})\big)^{(i)}. \tag{2}$$

Reading out intermediate states with the final head is the standard *logit-lens* technique (Allen-Zhu & Li, 2024). We measure the evocation of a token v by its *reciprocal rank* $r_\theta^{[m]}(v) = 1/n$, where n is the ordinal rank of v in $p_\theta^{[m]}(\cdot \mid \boldsymbol{x})$. This diagnostic applies the frozen LM head directly to each intermediate state. Figure 1 instead trains a linear classifier on the last-token hidden state of a question concatenated with a candidate answer, and predicts whether that candidate is the gold answer. Splits are by question group, and each gold answer is reused as a negative for other questions, which limits leakage. The two probes measure related but distinct forms of answer recoverability.

3.2 THE IMPLICIT EVOCATION PHENOMENON

Let $\mathcal{D}_{\mathrm{FG}}$ and $\mathcal{D}_{\mathrm{RT}}$ be the forget and retain sets. Following common practice, $\mathcal{D}_{\mathrm{RT}}$ is effectively the (prohibitively large) rest of pretraining and is not available in RWKU. For $(\boldsymbol{x}, y) \in \mathcal{D}_{\mathrm{FG}}$ we say $\boldsymbol{h}_\theta$ *explicitly unlearns* $(\boldsymbol{x}, y)$ if $y \neq \arg\max_v p_\theta(y = v \mid \boldsymbol{x})$; this is a purely end-to-end criterion that ignores all intermediate states.

Definition 1 (Implicit evocation)**.** *A model* $\boldsymbol{h}_\theta$ implicitly evokes $(\boldsymbol{x}, y) \in \mathcal{D}_{\mathrm{FG}}$ *at level* $\tau \in (0, 1)$ *if*

$$\begin{aligned} r_\theta^{[M]}(y) &< 1 \quad \textit{(explicitly unlearned at the output)}, \\ \max_{1 \le m < M} r_\theta^{[m]}(y) &> \tau \quad \textit{(retained internally).} \end{aligned} \tag{3}$$

The output ranking no longer places the target first, while some earlier layer still ranks it highly. A white-box readout at $m^\star = \arg\max_m r_\theta^{[m]}(y)$, followed by top-$\lceil 1/\tau \rceil$ sampling at high temperature, can return that token. Section 4.2 measures the gap on RWKU.

3.3 A Residual-Stream Analysis

We adopt the LRH and, following the residual-stream view, write the forward pass as a sum of layerwise *concept vectors* $\boldsymbol{c}_\theta^{(m)}(\boldsymbol{x}) = \boldsymbol{h}_\theta^{[m]}(\boldsymbol{x}) - \boldsymbol{h}_\theta^{[m-1]}(\boldsymbol{x}) = \boldsymbol{W}_\theta^{(m)} \operatorname{Attn}_\theta^{(m)}\big(\boldsymbol{h}_\theta^{[m-1]}(\boldsymbol{x})\big)$, with $\boldsymbol{h}_\theta^{[0]}(\boldsymbol{x}) = \boldsymbol{x}$. Thus $\boldsymbol{h}_\theta(\boldsymbol{x}) = \boldsymbol{x} + \sum_{m=1}^{M} \boldsymbol{c}_\theta^{(m)}(\boldsymbol{x})$. Since the logit of token v_i is $\boldsymbol{v}_i^\top \boldsymbol{h}_\theta(\boldsymbol{x})$, its post-layer logit is the partial sum $\boldsymbol{v}_i^\top \big(\boldsymbol{x} + \sum_{m' \le m} \boldsymbol{c}_\theta^{(m')}(\boldsymbol{x})\big)$. The following proposition states a sufficient condition for output suppression without intermediate suppression. It is a local analysis, not a claim that representations are globally linear or that low linear-probe accuracy proves irreversible deletion.

Proposition 1 (First-order separation of readouts)**.** *Let $\delta_{i,m} = z_i^{[m]}(\theta + \Delta\theta) - z_i^{[m]}(\theta)$. For some $m < M$, suppose the target's layer-m logit is separated from every other logit and the update is smaller than half the minimum separation, so its rank is unchanged. At the final layer, let $j^\star$ be the strongest competitor. If $\delta_{y,M} - \delta_{j^\star,M} < -(z_y^{[M]} - z_{j^\star}^{[M]})$, the update removes y from top-1 while preserving its layer-m rank. A loss applied to several prefix-level semantic readouts contributes explicit gradient terms at each selected readout; a final-token likelihood loss has only the final readout term. This difference makes broad suppression possible, but does not guarantee deletion or rule out nonlinear recoverability.*

3.4 Semantic Negative Preference Optimization

Embedding head. Reusing the ATU machinery, we attach a lightweight embedding head η_ϕ (a 6-layer, 8-head, 768-dim transformer) that maps the hidden state at a fixed hook layer into the space of a frozen pretrained sentence encoder, giving a per-position predicted embedding $\hat{\boldsymbol{e}}_t = \eta_\phi\big(\boldsymbol{h}_\theta(\boldsymbol{x}_{\le t})\big) \in \mathbb{R}^d$. Let $\boldsymbol{e}_{\text{unlearn}}$ be the frozen encoder embedding of the concept to be removed, and let $s_t = \operatorname{sim}(\hat{\boldsymbol{e}}_t, \boldsymbol{e}_{\text{unlearn}}) = \cos(\hat{\boldsymbol{e}}_t, \boldsymbol{e}_{\text{unlearn}})$ be the cosine similarity ($\boldsymbol{e}_{\text{unlearn}}$ is detached, so no gradient flows into the target). *Lower s_t means more forgotten.* The head is trained on retained text before each unlearning stage. For each prefix position t, the frozen encoder embeds the following context window (eight tokens in our experiments) as $\boldsymbol{e}_t^{\text{future}}$. With $\mathcal{T}$ denoting positions that have a complete window, the head loss is

$$L_{\text{head}}(\phi) = -\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \cos\big(\eta_\phi(\boldsymbol{h}_\theta(\boldsymbol{x}_{\le t})), \boldsymbol{e}_t^{\text{future}}\big)\,. \tag{4}$$

The LLM and sentence encoder are frozen in this phase. During unlearning, SeNPO optimizes θ and ϕ with the objective below.

Forget objective. For reference, token-level NPO uses the log-likelihood ratio against a frozen reference π_{θ_0}:

$$\begin{aligned} R_{\text{NPO}}(\theta, \mathcal{D}_{\text{FG}}) &= \frac{2}{\beta} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{FG}}} \log\Big(1 + \big(p_\theta(y \mid \boldsymbol{x})/p_{\theta_0}(y \mid \boldsymbol{x})\big)^\beta\Big) \\ &= \frac{2}{\beta} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{FG}}} \operatorname{softplus}\big(\beta\, \Delta^{\text{tok}}\big), \end{aligned} \tag{5}$$

with $\Delta^{\text{tok}} = \log p_\theta(y \mid \boldsymbol{x}) - \log p_{\theta_0}(y \mid \boldsymbol{x})$. SeNPO replaces the per-token log-likelihood ratio by a *length-normalized sentence-level embedding similarity*. In its default reference-free form, with inverse temperature β and margin γ,

$$R_{SeNPO}(\theta, \phi; \mathcal{D}_{\text{FG}}) = \frac{c}{\beta} \mathop{\mathbb{E}}_{(\boldsymbol{x},\cdot)\in\mathcal{D}_{\text{FG}}} \operatorname{softplus}\Big(\beta \underbrace{\tfrac{1}{|\boldsymbol{x}|}\textstyle\sum_t s_t}_{\bar{s}(\boldsymbol{x})} + \gamma\Big), \tag{6}$$

where $\bar{s}(\boldsymbol{x})$ is the mean similarity over valid positions (SimNPO-style length normalization) and c is the NPO prefactor ($c = 2$ by default). The derivative of the scaled loss in Equation (6) with respect to $\bar{s}$ is $c\,\sigma(\beta\bar{s} + \gamma) > 0$. The weight shrinks as $\bar{s}$ drops, and it does not hit zero. ATU's hinge $\operatorname{relu}(\bar{s} - \tau)$ has gradient 0 once $\bar{s} < \tau$. The input is a sentence-encoder similarity. Equation (5) uses a token log-likelihood ratio. An optional reference-anchored form subtracts the frozen head similarity at each position and applies softplus position-wise (Equation (11)); the main results use only Equation (6).

Retain regularizer. Following TOFU, we add a KL term on a retain batch. On RWKU, $\mathcal{D}_{\mathrm{RT}}$ is non-target entity text,

$$K(\theta; \mathcal{D}_{\mathrm{RT}}) = \mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathcal{D}_{\mathrm{RT}}} \mathrm{KL}\big(\pi_\theta(\cdot \mid \boldsymbol{x}) \,\big\|\, \pi_{\theta_0}(\cdot \mid \boldsymbol{x})\big). \tag{7}$$

K is evaluated on retained text. The same KL on $\mathcal{D}_{\mathrm{FG}}$ would pull those forget tokens back toward the reference and oppose R_{SeNPO} on the tokens being removed.

Total objective.

$$\min_{\theta,\phi} \;\; \alpha_1 \, R_{SeNPO}(\theta, \phi; \mathcal{D}_{\mathrm{FG}}) \;+\; \alpha_2 \, K(\theta; \mathcal{D}_{\mathrm{RT}}), \tag{8}$$

with weights α_1, α_2. We default to the reference-free form (Equation (6)) with $\beta = 1$, $\gamma = 0$, $c = 2$, $\alpha_1 = 1$, and tune α_2 (its effect is analyzed in Section 4.5). Optionally, confidence weighting reweights positions by $\mathrm{relu}(s_t)^a$ (stop-grad), concentrating forgetting on positions that still resemble the concept; we treat it as an ablation. Reference pseudocode for all terms is in Section B.

4 Experiments

We ask three questions. Does implicit evocation appear after output-level unlearning (Section 4.2)? Where does SeNPO sit on RWKU forgetting and utility (Section 4.3)? Which terms in the objective produce that point (Section 4.5)? TOFU is in Section 4.4. Llama-3.2-3B, where the forgetting order differs, is in Section E.

4.1 Setup

Benchmark and metrics. We use RWKU (Jin et al., 2024). Forgetting is ROUGE-L recall on Fill-in-the-Blank (FB), Question Answering (QA), and Adversarial Attack (AA; jailbreak suffixes and in-context probes); lower is better. Retention is Neighbor QA/FB, and general utility is MMLU, BBH, and Fluency; higher is better. We also report forget-set membership inference ($\mathrm{MIA}_{\mathrm{fm}}$). The Phi table averages 14 targets. Per-target numbers are in Section C.[1]

Model, baselines, and protocol. Following ATU, the Phi-3 base is Phi-3-mini-4k-instruct (3.8B). The frozen sentence encoder is Qwen3-Embedding-4B, with all-mpnet-base-v2 in the faithful ATU replication. The embedding head is a 6-layer, 8-head, 768-dim transformer. Baselines are Gradient Ascent (GA) (Yao et al., 2024), NPO (Zhang et al., 2024), and ATU (Spohn et al., 2025). All methods share the RWKU data and evaluation and are full-parameter fine-tuned. Reported GA and NPO use their original objectives and do not receive SeNPO's retain-set KL, so utility gaps compare complete methods. ATU-matched keeps the SeNPO framework and replaces only the forget loss with $\mathrm{relu}(\bar{s} - \tau)$. Learning rates are method-specific: SeNPO at 10^{-2}, GA at 10^{-6}, and NPO at 10^{-5}. Larger GA and NPO rates collapse, and six lower ATU-matched rates from 10^{-5} to $3{\times}10^{-3}$ preserve utility without strong forgetting (Section 4.5). NPO is available on 5 targets; the other Phi rows use 14.

4.2 Hidden-State Recoverability (Q1)

We test whether a forget-set answer that no longer appears in the final output can still be recovered from hidden states by a linear classifier. Definition 1 instead ranks the answer with the frozen LM head. The probe in Figure 1 is trained separately. Its input is the last-token hidden state of a question concatenated with one candidate answer, and it predicts whether that candidate is correct. Every gold answer is also used as an incorrect candidate for other questions. We split examples by question group (60/20/20), select the classifier's regularization C using validation balanced accuracy, and report test balanced accuracy over the split seeds $\{13, 17, 23, 29, 31\}$. No checkpoint was chosen using these test scores.

The comparison uses Qwen2.5-7B-Instruct. Each of the 10 cells contains one unlearned checkpoint and the frozen model for the same target and training seed. The four RWKU targets are

[1] The target `Prince Harry, Duke of Sussex` is excluded because its comma breaks configuration parsing. The exclusion is shared by every method.

1_Stephen_King, 2_Confucius, 3_Bruce_Lee, and 4_Warren_Buffett; the seeds are $\{13, 17, 23, 29, 31\}$. GA and NPO use the only recorded points that do not collapse the model: 150 AdamW updates at learning rates 3×10^{-6} and 10^{-5}, respectively. Stronger settings reduce Fluency to 0 for GA and 361 for NPO. SeNPO uses one fixed five-stage schedule, which reaches single-digit Forget-All while keeping MMLU ≥ 50 and Fluency ≥ 400. In Figure 1, the shaded band marks layers 10 to 18 of this 28-layer model. We chose that band before inspecting the probe curves.

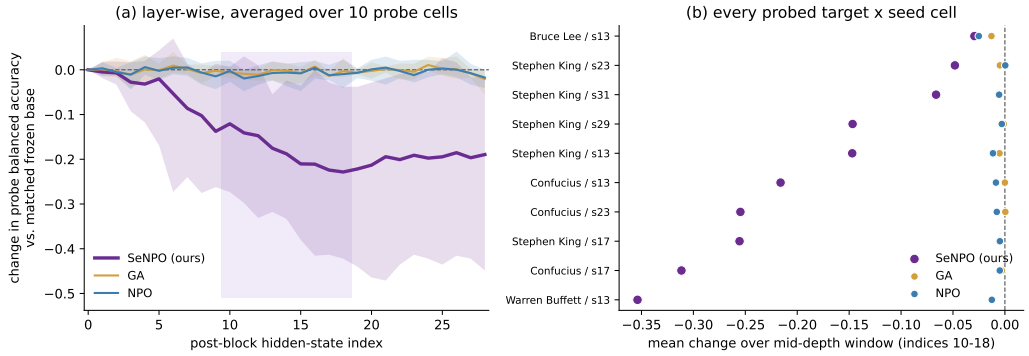

Figure 1: Change in linear-probe balanced accuracy relative to the matched frozen Qwen2.5-7B-Instruct model. Negative values mean the gold answer is harder to recover. (a) Average over 10 target-seed cells; shading shows the minimum and maximum, and the band marks layers 10 to 18. (b) Each cell's average over that same layer range. The mid-layer average is -0.183 for SeNPO (final layer -0.190), -0.005 for GA, and -0.008 for NPO. Their final-layer changes are -0.021 and -0.017. SeNPO has the largest decrease in all 10 cells. Each point averages five question-group splits, without smoothing.

Values near zero in Figure 1 mean that the linear probe recovers the answer about as accurately as it does from the frozen model. GA and NPO remain within 0.009 of that baseline. SeNPO lowers accuracy by 0.183 on average over layers 10 to 18, and it produces the largest decrease in every cell. These GA and NPO checkpoints also remove less of the answer from the final output, so the figure compares usable models at different levels of output forgetting.

On the Stephen King target with training seed 13, a later search selects checkpoints by Forget-All only and does not look at probe scores. Table 1 lists the recorded points nearest the SeNPO Forget-All of 5.90. GA at 232 updates reaches Forget-All 5.40, but Neighbor falls to 9.8 and the mid-layer probe change stays -0.005. NPO falls from Forget-All 12.61 at 170 updates to 0 at 180, so the nearest recorded point is still Forget-All 0, with Neighbor 0 and a probe change of -0.008. The SeNPO checkpoint keeps Neighbor 65.2 and Fluency 702.5, and its mid-layer probe change is -0.147. Matching output forgetting does not reproduce that linear suppression, and the GA match pays for it with Neighbor 9.8.

Table 1: Forget-All-only probe on Qwen2.5-7B, Stephen King, training seed 13. GA and NPO are the recorded points nearest the SeNPO Forget-All of 5.90. Mid-depth is the mean change in balanced accuracy versus the frozen base, over layers 10 to 18 and five split seeds. The SeNPO row is the earlier ladder2 checkpoint on the same target and seed, not a joint probe of the GA and NPO weights. NPO fluency is from the probed rerun.

Method	**Forget-All** $\downarrow$	**Neighbor** $\uparrow$	**Fluency** $\uparrow$	**Mid-depth probe** $\downarrow$
GA (232 updates)	0.00	9.8	609.7	-0.007
NPO (180 updates)	0.00	0.0	528.9	-0.006
SeNPO (ladder2)	5.90	65.2	702.5	-0.147

A predeclared search found no stable interior peak in the layer-wise curves. A separate Llama-3.2-3B logit-lens example shows large intermediate reciprocal-rank peaks for the frozen model, GA,

NPO, and SeNPO alike. Section D records the selection rule and both checks. Lower linear-probe accuracy still leaves open recovery by a nonlinear decoder or by further fine-tuning (Section F).

4.3 RWKU Forgetting and Utility (Q2)

Table 2 is the Phi-3 average. Among reported points that forget and remain usable, SeNPO lowers Forget from the pre-unlearning band of 47 to 56 to $6.4/7.3/10.6$ (FB/QA/AA), with MMLU 58.6 and Fluency 652.8. ATU-matched uses the same framework and the hinge loss. It reaches Forget $1.8/1.9/3.0$ and collapses (MMLU 37.4, Fluency 383, Neighbor $1.7/3.5$, MIA 833). GA and NPO at the recorded usable rates stay near Forget 40 to 55, so their Neighbor and MMLU remain high because the model barely moves. Stronger GA and NPO settings collapse Fluency and raise MIA (Section 4.5). The AA column is the jailbreak and in-context split: SeNPO reduces it from 55.8 to 10.6, while GA and NPO leave it between 47 and 55. NPO covers 5 targets and the other rows cover 14, and GA/NPO omit retain-set KL, so the table compares full recipes at the disclosed scopes.

Table 2: RWKU on Phi-3 (14-target average; NPO on 5 targets). Forget ↓; Neighbor, utility, and Fluency ↑. Bold marks the best value among methods that both forget and stay usable. ATU-matched changes only the forget loss to the ATU hinge.

	Forget ↓			**Neighbor ↑**		**Utility ↑**				
Method	FB	QA	AA	QA	FB	MMLU	Fluency	MIA_{fm}	Usable	n
Before unlearning (ref.)	47.1	47.4	55.8	61.4	56.2	64.4	∼700	∼0	yes	n/a
GA (lr 10^{-6})	53.0	50.3	55.1	53.7	53.6	62.9	706.9	267.1	no forget	14
NPO (lr 10^{-5})	40.9	52.8	47.0	37.8	43.9	64.4	724.8	990.5	no forget	5
ATU-matched (lr 10^{-2})	1.8	1.9	3.0	3.5	1.7	37.4	383.0	832.9	collapsed	14
SeNPO (ours, lr 10^{-2})	**6.4**	**7.3**	**10.6**	**15.7**	**23.7**	**58.6**	**652.8**	**336.6**	**yes**	14

On Qwen2.5-7B-Instruct, the same frozen SeNPO schedule over four targets and five seeds (20 utility-gated cells) reaches Forget-All 8.22 ± 9.39, Neighbor-All 30.99 ± 22.82, MMLU 68.27 ± 5.93, and Fluency 704.55 ± 35.78. GA and NPO were not run on those 20 cells. Table 3 is the same base on one target, Stephen King, seed 13. SeNPO uses 400 predictor updates and then 300 full-parameter SGD forget updates at learning rate 10^{-2}. GA and NPO use 150 AdamW updates at 3×10^{-6} and 10^{-5}. All three pass MMLU $\geq$ 50 and Fluency $\geq$ 400. At these points SeNPO has the lowest Forget-All and the highest Neighbor-All, with MMLU and Fluency close to GA. The rows are one seed and one target, and the update budgets are not matched.

Table 3: Single-target Qwen2.5-7B calibration, Stephen King, seed 13. Update counts and learning rates are method-specific. Forget ↓; Neighbor, MMLU, and Fluency ↑.

Method (updates)	**Forget-All**	**Neighbor-All**	**MMLU**	**Fluency**
GA (150)	47.06	75.39	**76.61**	**706.66**
NPO (150)	25.05	54.02	**76.61**	697.37
SeNPO (400 predictor + 300 forget)	**10.96**	**76.16**	75.44	706.25

Table 4 repeats the Stephen King comparison on Llama-3.1-8B, with a broader retain pool. The SeNPO point is 500 unlearning updates over seven stages. NPO, at 125 updates, reaches a lower Forget-All and a higher MMLU. SeNPO keeps more neighboring knowledge and much higher fluency. On Llama-3.2-3B the forgetting order reverses: utility-gated GA and NPO forget more than SeNPO at similar MMLU (Section E).

4.4 TOFU

TOFU forget10 adds token-level SimNPO and representation-level RMU. We follow OpenUnlearning: five-epoch SFT produces a full-data target and an independent retain90 oracle, and every method starts from that full checkpoint. The TOFU adaptation mean-pools answer-token hidden states, repels them from the frozen full-model representation, and applies retain KL. It does not use the RWKU Qwen3 embedding predictor. In Table 5, NPO and SimNPO use the official retain NLL.

Table 4: Llama-3.1-8B, Stephen King, seed 13. Forget-All ↓; Neighbor-All, MMLU, and Fluency ↑. Single-target operating points.

Method	Forget-All	Neighbor-All	MMLU	Fluency
NPO (125 updates)	**1.53**	32.13	**67.84**	486.49
SeNPO (500 updates, 7 stages)	2.74	**39.49**	62.57	**685.01**

Table 5: TOFU forget10, seed 13. Forget Quality (FQ) and Model Utility (MU) are higher better. Every row shares the target checkpoint and retain90 oracle. Retain losses differ and are stated in the text.

Model	Method	Seed	FQ	MU
Phi-3.5-mini	GA	13	≈ 0	0.000
	NPO	13	0.111	0.296
	SimNPO	13	0.523	**0.441**
	RMU	13	0.641	0.440
	SeNPO	13	**0.700**	0.435
Qwen2.5-7B	GA	13	≈ 0	0.000
	NPO	13	0.030	0.486
	SimNPO	13	3.6×10^{-5}	**0.502**
	RMU	13	0.016	0.500
	SeNPO	13	**0.322**	0.487

On Phi, SeNPO improves FQ by 0.060 over RMU and by 0.177 over SimNPO, with MU lower by 0.005 and 0.007. On Qwen, it improves FQ by 0.292 over NPO and by 0.306 over RMU at nearly the same MU. These are seed-13 point estimates.

Equal retain KL. Table 6 runs GA, NPO, and SimNPO with retain KL at $\alpha \in \{1, 3\}$, using the same full checkpoint, retain90 oracle, ten-epoch budget, and evaluator. On Qwen, SeNPO remains ahead: FQ 0.322 versus 0.037 for NPO+KL and 0.003 for SimNPO+KL, with MU differing by at most 0.027. On Phi, SimNPO+KL reaches FQ 0.758 at MU 0.435, above SeNPO's FQ 0.700 at the same MU. GA+KL stays degenerate on both models (FQ ≈ 0; MU 0.186 on Phi and 0 on Qwen).

Table 6: TOFU seed-13 control with retain KL on every row. For each baseline we report the $\alpha \in \{1, 3\}$ with higher FQ. FQ and MU are higher better.

Model	Method	α	FQ	MU
Phi-3.5-mini	NPO+KL	1	0.281	0.252
	SimNPO+KL	1	**0.758**	**0.435**
	SeNPO	3	0.700	**0.435**
Qwen2.5-7B	NPO+KL	1	0.037	0.460
	SimNPO+KL	1	0.003	0.480
	SeNPO	1	**0.322**	**0.487**

4.5 Ablations (Q3)

Table 7 changes one design choice at a time. The hinge forgets more and collapses. Dropping retain KL ($\alpha_2 = 0$) also collapses (Neighbor 0, Utility 31). Only the full objective forgets and stays usable. On the single-target protocol, a 3-stage schedule reaches Forget 1.95 with Neighbor ≈ 43, so Neighbor can be bought back with a milder schedule or with learning rate in $[5\times10^{-3}, 8\times10^{-3}]$ (Sections C and F).

Retain-KL weight. With $\alpha_1 = 1$, the α_2 sweep in Table 10 runs from collapse at $\alpha_2 = 0$ to near-base behavior at $\alpha_2 \geq 0.1$, where Forget All stays above 52. Small positive weights (0.05 and 0.06)

Table 7: Component ablation. Each row changes one choice. Forget ↓; Neighbor and Utility ↑. Row (a) uses the matched-loss protocol of Table 2; rows (b) and (c) use the α_2 protocol.

	Configuration	Forget All ↓	Neighbor All ↑	Utility Gen ↑	Usable
(a)	hinge loss instead of softplus (= ATU-matched)	0.9	1.8	30.8	collapsed
(b)	softplus, no retain KL (α_2=0)	2.3	0.0	31.0	collapsed
(c)	**full SeNPO (softplus + retain KL)**	**5.4**	**24.3**	**58.5**	**yes**

keep Utility Gen between 52.6 and 65.5 while Forget All drops to 18.5 and 12.7. The Phi result in Table 2 is a usable point on this sweep, not the collapse point $\alpha_2 = 0$.

Learning-rate sweep. Table 8 sweeps the unlearning learning rate on the single-target Phi-3 protocol (Stephen King). Every baseline has a gap in its sweep: GA and NPO either barely forget (Forget All 57.12 and 44.77) or collapse to Fluency below 300 one decade higher. ATU-matched preserves utility at six rates from 10^{-5} to $3{\times}10^{-3}$, but its best Forget All there is 49.58; at 10^{-2} and $3{\times}10^{-2}$ it collapses. SeNPO at 10^{-2} reaches Forget All 2.45 with Neighbor 58.71, MMLU 67.8, and Fluency 705.9. The grid is coarse, so a usable baseline point between grid values is not ruled out.

Table 8: Single-target (Stephen King) learning-rate sweep on Phi-3. Forget ↓; Neighbor, MMLU, and Fluency ↑.

Method	lr	Forget All ↓	Neighbor ↑	MMLU	Fluency	Usable
SeNPO	10^{-2}	**2.45**	58.71	67.8	705.9	**yes**
ATU-matched	10^{-5}	60.24	77.35	68.4	709.2	no forget
ATU-matched	3×10^{-5}	57.12	76.18	69.0	710.2	no forget
ATU-matched	10^{-4}	60.24	77.89	69.0	707.6	no forget
ATU-matched	3×10^{-4}	60.24	77.89	69.0	707.0	no forget
ATU-matched	10^{-3}	56.03	79.52	69.0	710.8	no forget
ATU-matched	3×10^{-3}	49.58	74.73	66.7	706.5	no forget
ATU-matched	10^{-2}	0.00	0.00	28.7	286.8	collapsed
ATU-matched	3×10^{-2}	0.00	0.00	24.0	18.9	collapsed
GA	10^{-6}	57.12	72.86	69.0	706.4	no forget
GA	10^{-5}	0.00	0.00	30.4	0.0	collapsed
GA	10^{-4}	0.00	0.00	30.4	0.0	collapsed
NPO	10^{-5}	44.77	73.82	67.8	723.2	no forget
NPO	10^{-4}	0.00	0.00	31.6	297.8	collapsed

5 Conclusion

SeNPO puts a length-normalized softplus preference loss on sentence-encoder similarities and anchors retained behavior with retain-set KL. The loss follows a local residual-stream condition: an update can change the final top-1 token while an earlier readout keeps its rank.

On Qwen2.5-7B, a frozen schedule forgets strongly across four RWKU targets and five seeds, and it lowers mid-depth linear probe accuracy in all ten probed cells. On TOFU, the adaptation reaches higher Forget Quality than representation-level RMU on both tested models, and higher Forget Quality than equal-KL SimNPO on Qwen. Equal-KL SimNPO is stronger on Phi. Matched-budget ATU either barely forgets or collapses. Scope limits, including the probe, the RWKU regularizer mismatch, the single TOFU seed, and the Llama comparison, are in Section F.

Ethics Statement

This work targets the safe removal of objectionable or sensitive knowledge from LLMs and therefore contributes to trustworthy deployment. Experiments use the public RWKU and TOFU benchmarks; RWKU contains public entity knowledge, while TOFU contains fictitious authors. We collected no private data. The probing attack we describe is a diagnostic used to reveal residual knowledge; it is disclosed to motivate stronger unlearning, not to enable misuse, and it operates only on white-box internal activations that a deployer already controls. Our method reduces, rather than amplifies, the exposure of objectionable content.

Reproducibility Statement

We specify the full training and evaluation protocol, hyperparameters, and per-target settings in Sections C and 4.1. The exact SeNPO loss functions used in all experiments are given in closed form in Section 3 and reproduced as reference pseudocode in Section B. The local sufficient condition is stated as Proposition 1 and derived in Section A. Aggregate and per-target metrics, the RWKU splits, and the baseline configurations, regularization differences, target counts, and operating-point selection are described in Sections C and 4 so that the scope of each comparison is explicit.

AI Use Statement

In this work, we used generative AI tools for editing code, drafting parts of the paper, and polishing writing to improve the readability of the manuscript. We have not used generative AI tools for research ideation, theoretical development,methodological or experimental design, mathematical derivations or proofs, data analysis, or interpretation of experimental results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *ICML*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *SaTML*, 2025.

Yiwei Chen, Soumyadeep Pal, Yimeng Zhang, Qing Qu, and Sijia Liu. Unlearning isn't invisible: Detecting unlearning traces in llms from model outputs. *CoRR*, abs/2506.14003, 2025.

Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Position: Building guardrails for large language models requires systematic design. In *ICML*, 2024.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *CoRR*, abs/2310.02238, 2023.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *CoRR*, abs/2410.07163, 2024.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *NeurIPS*, 2024.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: Benchmarking real-world knowledge unlearning for large language models. *CoRR*, abs/2406.10890, 2024.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou,

Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Kiran Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *ICML*, 2024.

Zexi Li, Xiangzhu Wang, William F. Shen, Meghdad Kurmanji, Xinchi Qiu, Dongqi Cai, Chao Wu, and Nicholas D. Lane. Editing as unlearning: Are knowledge editing methods strong baselines for large language model unlearning? *CoRR*, abs/2505.19855, 2025.

Guozhi Liu, Weiwei Lin, Qi Mu, Tiansheng Huang, Ruichao Mo, Yuren Tao, and Li Shen. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation. *IEEE Trans. Inf. Forensics Secur.*, 2025.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *ICLR*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for llms. *CoRR*, abs/2401.06121, 2024.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Jiayi Mao, and Xueqi Cheng. "not aligned" is not "malicious": Being careful about hallucinations of large language models' jailbreak. In *COLING*, 2025.

Daniel Mela, Aitor Gonzalez-Agirre, Javier Hernando, and Marta Villegas. Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge. In *ACL*, 2024.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple word2vec-style vector arithmetic. In *NAACL-HLT*, 2024.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *BlackboxNLP@EMNLP*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *ICML*, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *ICLR*, 2025.

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. In *ICLR*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Aakash Sen Sharma, Debdeep Sanyal, Priyansh Srivastava, Sundar Atreya H., Shirish S. Karande, Mohan S. Kankanhalli, and Murari Mandal. Nine ways to break copyright law and why our LLM won't: A fair use aligned generation framework. *CoRR*, abs/2505.23788, 2025.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. *CoRR*, abs/2407.06460, 2024.

Minkyoo Song, Hanna Kim, Jaehan Kim, Seungwon Shin, and Sooel Son. Refusal is not an option: Unlearning safety alignment of large language models. In *34th USENIX Security Symposium, USENIX Security 2025, Seattle, WA, USA, August 13-15, 2025*, pp. 35–52. USENIX Association, 2025. doi: 10.5281/zenodo.16740884. URL https://www.usenix.org/conference/usenixsecurity25/presentation/song-minkyoo.

Philipp Spohn, Leander Girrbach, Jessica Bader, and Zeynep Akata. Align-then-unlearn: Embedding alignment for LLM unlearning. *CoRR*, abs/2506.13181, 2025.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: understanding factors influencing machine unlearning. In *EuroS&P*, 2022.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions. *CoRR*, abs/2404.13208, 2024.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q. Weinberger. Rethinking llm unlearning objectives: A gradient perspective and go beyond. *CoRR*, abs/2502.19301, 2025a.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. *CoRR*, abs/2505.17306, 2025b.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. In *NeurIPS*, 2023.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *ICML*, 2024.

Jialin Wu, Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Jiayang Xu, Xinfeng Li, and Wenyuan Xu. Legilimens: Practical and unified content moderation for large language model services. In *CCS*, 2024.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms. *CoRR*, abs/2505.16831, 2025.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *NeurIPS*, 2024.

Paul Youssef, Zhixue Zhao, Christin Seifert, and Jörg Schlötterer. Tracing and reversing rank-one model edits. *CoRR*, abs/2505.20819, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *CoRR*, abs/2404.05868, 2024.

Xuandong Zhao, Will Cai, Tianneng Shi, David Huang, Licong Lin, Song Mei, and Dawn Song. Improving LLM safety alignment with dual-objective optimization. In *ICML*, 2025.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043, 2023.

A Proof of Proposition 1

Let $\boldsymbol{h}_\theta^{[0]}(\boldsymbol{x}) = \boldsymbol{x}$ and

$$\boldsymbol{h}_\theta^{[m]}(\boldsymbol{x}) = \boldsymbol{h}_\theta^{[m-1]}(\boldsymbol{x}) + \boldsymbol{c}_\theta^{(m)}\left(\boldsymbol{h}_\theta^{[m-1]}(\boldsymbol{x})\right).$$

For a linear LM head, $z_i^{[m]} = \boldsymbol{v}_i^\top \boldsymbol{h}_\theta^{[m]}(\boldsymbol{x})$ and $p_i^{[m]} = \exp z_i^{[m]} / \sum_j \exp z_j^{[m]}$. The standard log-probability gradient at the final hidden state is

$$\nabla_{\boldsymbol{h}^{[M]}} \log p_y^{[M]} = \boldsymbol{v}_y - \sum_j p_j^{[M]} \boldsymbol{v}_j. \tag{9}$$

The partition function is therefore not constant: all vocabulary items contribute through the second term.

For any parameter block θ_k, the complete chain rule is

$$\frac{d\boldsymbol{h}^{[m]}}{d\theta_k} = \left(I + \frac{\partial \boldsymbol{c}^{(m)}}{\partial \boldsymbol{h}^{[m-1]}}\right) \frac{d\boldsymbol{h}^{[m-1]}}{d\theta_k} + \frac{\partial \boldsymbol{c}^{(m)}}{\partial \theta_k}, \tag{10}$$

where the direct partial derivative is zero only when block m does not contain θ_k. This term is necessary under full-parameter fine-tuning. Combining Equations (9) and (10) yields the exact gradient of a final-token log-likelihood without treating the softmax normalizer as a constant.

Proof of the stated sufficient condition. Under a first-order expansion,

$$\delta_{i,m} = z_i^{[m]}(\theta + \Delta\theta) - z_i^{[m]}(\theta) = \nabla_\theta z_i^{[m]}(\theta)^\top \Delta\theta + o(\|\Delta\theta\|).$$

If the magnitude of every layer-m logit change is less than half the minimum strict separation between $z_y^{[m]}$ and the other logits, none of those pairwise orderings can reverse. At the final layer, $\delta_{y,M} - \delta_{j^\star,M} < -(z_y^{[M]} - z_{j^\star}^{[M]})$ reverses the target-to-competitor margin and removes y from top-1. The earlier ranking remains unchanged by the first condition. ■

Why semantic aggregation is relevant. A final-token likelihood supplies one readout gradient, Equation (9), at position L. The default SeNPO loss depends on $\bar{s} = |\mathcal{T}|^{-1} \sum_{t\in\mathcal{T}} s_t$, so its parameter gradient is

$$\nabla_\theta R_{SeNPO} = \frac{c\,\sigma(\beta\bar{s} + \gamma)}{|\mathcal{T}|} \sum_{t\in\mathcal{T}} \nabla_\theta s_t.$$

It therefore includes explicit terms from every selected prefix readout. Whether those terms reduce a particular intermediate rank depends on their Jacobians and geometry; neither LRH nor this local calculation proves irreversible semantic deletion. The probe experiment measures the resulting linear recoverability, and we state nonlinear recoverability as a limitation.

B Reference Pseudocode for the SeNPO Objective

For reproducibility we give the exact per-term computation used in all experiments; these are pure functions of the predicted per-position similarities $s_t = \cos(\hat{\boldsymbol{e}}_t, \boldsymbol{e}_{\text{unlearn}})$ (target detached).

- **Embedding-head training, Equation (4):** at each valid prefix, encode the following eight-token context window with the frozen sentence encoder and minimize negative cosine similarity to the head output. The LLM and sentence encoder are frozen during this phase.
- **Forget (reference-free, default), Equation (6):** with mask $\mathbf{1}_t$ over valid positions, $\bar{s} = \big(\sum_t \mathbf{1}_t s_t\big)/\big(\sum_t \mathbf{1}_t\big)$, and loss $\frac{c}{\beta}\,\mathrm{softplus}(\beta\bar{s} + \gamma)$, averaged over the batch. Defaults c=2, β=1, γ=0.
- **Forget (reference-anchored), Equation (11):** per position $\frac{c}{\beta}\mathrm{softplus}\big(\beta(s_t - s_t^{\text{ref}})\big)$, masked-averaged. The main results do not use this variant.

$$R_{SeNPO}^{\text{ref}}(\theta, \phi; \mathcal{D}_{\text{FG}}) = \frac{c}{\beta} \mathop{\mathbb{E}}_{(\boldsymbol{x},\cdot)\in\mathcal{D}_{\text{FG}}} \frac{1}{|\boldsymbol{x}|} \sum_t \mathrm{softplus}\big(\beta\,(s_t - s_t^{\text{ref}})\big). \tag{11}$$

- **Retain KL, Equation (7):** on a *retain* batch, $\mathrm{KL}\big(\mathrm{softmax}(z_\theta) \,\|\, \mathrm{softmax}(z_{\theta_0})\big)$ summed over the vocabulary and masked-averaged over positions. Fed retain logits, never forget logits.
- **Confidence weight (ablation):** $w_t = \mathrm{relu}(s_t)^a$ (stop-grad), multiplied into the per-position forget term.
- **Total, Equation (8):** $\alpha_1 \cdot (\text{forget}) + \alpha_2 \cdot (\text{retain KL})$.

We use the identity $-\log\sigma(-x) = \mathrm{softplus}(x)$ for numerical stability.

C Additional Experimental Details and Results

Protocol. The Phi protocol performs 400 steps of embedding-head training followed by 300 steps of unlearning, then an evaluation pass. SeNPO/ATU-matched use unlearning lr $= 10^{-2}$ in the shared-scale ablation; ATU-matched is also swept at six lower rates. GA uses 10^{-6} and NPO uses 10^{-5} at the reported operating points. The retain set is instantiated from non-target RWKU entity text. SeNPO defaults: reference-free variant, β=1.0, γ=0, c=2, α_1=1.0; retain regularizer on the retain batch.

SeNPO hyperparameter sweep. Table 9 shows that the unlearning learning rate is the main knob and that lr 10^{-2}, β=1 is the sweet spot (strong forgetting with preserved utility). A milder 3-stage schedule reaches Forget 1.95 with Neighbor $\approx$ 43 on this target, illustrating the Neighbor/forget trade discussed in Section 4.5.

Table 9: SeNPO sweep on the single-target protocol.

lr	β	Forget All ↓	Neighbor ↑	MMLU	Fluency
3×10^{-3}	1	53.10	78.45	69.6	708
10^{-2}	1	**3.38**	53.72	67.8	702
3×10^{-2}	1	3.36	26.60	49.1	698
10^{-2}	4	4.09	31.23	59.6	720
3×10^{-2}	4	2.29	39.99	49.1	697

Retain-KL weight. Table 10 gives the full α_2 sweep discussed in Section 4.5.

Table 10: Effect of the retain-KL weight α_2 (with α_1=1.0). Forget ↓; Neighbor / Utility ↑. α_2=0 is the no-regularizer collapse point.

α_2	Forget All ↓	Neighbor All ↑	Utility Gen ↑	Utility Rea ↑
0.00	2.3	0.0	31.0	0.0
0.05	18.5	22.5	65.5	18.5
0.06	12.7	8.2	52.6	22.2
0.10	52.1	61.7	64.3	33.3
0.50	53.6	56.2	67.3	44.4

Five-target (celebs-1) results. As an intermediate-scale check, Table 11 reports the available RWKU celebs-1 averages (Stephen King, Confucius, Bruce Lee, Warren Buffett, Christina Aguilera). The table uses each method's recorded operating point. NPO covers only three of the five targets, so its aggregate is not directly target-matched to the other rows.

Table 11: RWKU celebs-1 five-target average (NPO on 3 targets).

Method	Forget All ↓	Neighbor ↑	MMLU	Rea	Fluency	Usable
SeNPO (lr 10^{-2})	**5.36**	**24.28**	**58.5**	**40.2**	**575.6**	**yes**
ATU-matched (lr 10^{-2})	0.86	1.75	30.8	8.9	239.1	collapsed
GA (lr 10^{-6})	60.64	57.68	64.0	36.3	711.3	no forget
NPO (lr 10^{-5})	46.91	53.82	64.5	37.9	724.7	no forget

D Probe Controls

Forget-All-only probe. The utility-gated probe in Section 4.2 compares checkpoints that also differ in output forgetting. We then searched the transition with Forget-All only: GA at learning rate 3.25×10^{-6} for 226 to 248 updates, and NPO at 10^{-5} for 152 to 190 updates. Selection used distance

to the SeNPO ladder2 Forget-All of 5.899 and did not use probe scores. GA at 232 updates is the closest recorded point: Forget-All 5.40, Neighbor-All 9.82, MMLU 76.0, and Fluency 609.7. NPO falls from Forget-All 12.61 at 170 updates (Neighbor-All 24.8, Fluency 658.9) to 0 at 180 updates. Because $|0 - 5.899| < |12.61 - 5.899|$, the rule probes the 180-update checkpoint. Its rerun scores Forget-All 0, Neighbor-All 0, and Fluency 528.9; the calibration row had Fluency 493.2. Table 1 reports those checkpoints in the main text. Mid-depth changes are -0.007 for GA and -0.006 for NPO, against -0.147 for the earlier SeNPO checkpoint on the same target and seed. The near Forget-All match therefore does not produce SeNPO-scale linear suppression, and GA pays for it with neighbor loss. NPO 170 was not probed.

Interior-peak search and logit lens. Absolute layer-wise probe accuracy rises toward the final layer for every method. We do not treat SeNPO as uniquely flattening an intermediate peak. We searched 100 combinations of readout (balanced accuracy, ROC-AUC), normalization (raw, change-from-base, excess over the endpoint chord, excess after base subtraction, fraction of final layer), target, and seed under a predeclared clear-peak rule (endpoint-gap CI lower bound ≥ 0.05 and local-prominence CI lower bound ≥ 0.02). Exactly one combination showed clear interior peaks for both baselines and none for SeNPO, and it reversed at a different split seed, so we treat it as noise. The search record is released with the code. A prior Llama-3.2-3B logit-lens case study on four "Portland" paraphrases likewise found large intermediate reciprocal-rank peaks for the base, GA, NPO, and SeNPO alike (between 0.625 and 0.750). ATU suppressed that readout ($\max_m RR_m = 0.015$) only at a collapsed point (MMLU 25.7, Fluency 141.9).

E Results on Other Models

Llama-3.2-3B-Instruct. We repeat the RWKU protocol on Llama-3.2-3B-Instruct (28 layers, hook layer 26), keeping the sentence encoder and embedding head fixed. Each method uses a multi-stage schedule. Per target we report the strongest-forgetting stage that stays usable (Fluency ≥ 400), following Jin et al. (2024). Learning rates are re-tuned for this base (GA 3×10^{-6}, NPO 10^{-5}, SeNPO/ATU-matched 10^{-2}). Table 12 averages the same 14 targets under MMLU ≥ 50 and Fluency ≥ 400. GA forgets more than SeNPO and has slightly higher MMLU. SeNPO has the highest Fluency. These are single-run point estimates.

Table 12: RWKU on Llama-3.2-3B-Instruct (14-target average; NPO on 5 targets). Per target, the strongest-forget stage with MMLU ≥ 50 and Fluency ≥ 400. Forget ↓; Neighbor, utility, and Fluency ↑.

	Forget ↓			Neighbor ↑	Utility ↑		
Method	FB	QA	AA	All	MMLU	Fluency	Usable
GA (3×10^{-6})	10.0	4.5	9.3	22.5	56.8	658.5	yes
NPO (10^{-5})	1.5	10.5	17.8	21.5	58.1	678.8	yes
ATU-matched (10^{-2})	36.5	41.7	40.5	41.9	55.3	693.5	yes
SeNPO (ours)	15.5	16.6	22.7	23.6	55.3	**708.4**	yes

Llama-3.1-8B and single-target Qwen. The Llama-3.1-8B Stephen King comparison and the single-target Qwen2.5-7B calibration are reported in Tables 3 and 4. Both are single-target, single-seed operating points. The Qwen rows use method-specific update counts and are not a controlled compute comparison. The four-target five-seed aggregate remains the Qwen summary in Section 4.3.

F Discussion and Limitations

No proof of semantic deletion. The LRH is a local abstraction of a nonlinear model. A drop in linear probe accuracy means that the tested linear readout recovers less information; it does not show that the information is absent from all nonlinear functions of the hidden state. Recovery fine-tuning, nonlinear probes, and relearning tests are needed before making an irreversible-erasure claim (Xu et al., 2025).

Baseline parity. RWKU GA and NPO follow their original losses and do not use SeNPO's retain-set KL. The utility gap therefore compares full recipes and cannot be attributed only to semantic versus token space. The ATU-matched row isolates the loss form at one shared operating point, supplemented by the method-specific lower-rate sweep in Table 8. NPO also covers fewer targets in the Phi tables. Future controlled experiments should give every RWKU method the same retain regularizer, target set, and tuning budget. On TOFU we provide equal-retain-KL controls for GA, NPO, and SimNPO (Table 6); those controls show that the Qwen advantage survives regularizer parity, whereas SimNPO+KL is stronger on Phi. ATU and RMU have method-specific representation-retention terms and are not part of that KL-matched ablation.

Benchmark and model scope. TOFU adds token-level SimNPO and representation-level RMU controls on Phi-3.5 and Qwen2.5, but the displayed rows cover only seed 13. The Llama results are mixed: NPO forgets more, while SeNPO preserves more Neighbor and Fluency. MUSE (Shi et al., 2024) and WMDP (Li et al., 2024) are not evaluated. WMDP is especially relevant to the paper's hazardous-knowledge motivation, whereas RWKU entity facts are only a controlled proxy.

Output-forgetting-controlled probing. On Qwen2.5-7B Stephen King, seed 13, the fine Forget-All bracket reaches GA Forget-All 5.40 at 232 updates, with Neighbor-All 9.8 and a mid-depth probe change of -0.005 (Section D). The nearest NPO point is still Forget-All 0, with probe change -0.008. NPO at 170 updates (Forget-All 12.61) was not selected. A nonlinear decoder or recovery fine-tune is not tested here.

Neighbor retention. The aggressive single-stage learning rate 10^{-2} lowers Neighbor relative to GA and NPO runs that barely forget (Section 4.5). On the single-target protocol, a standard 3-stage schedule reaches Forget 1.95 with Neighbor ≈ 43 (Table 9). Neighbor can also be traded back with learning rate in $[5\times10^{-3}, 8\times10^{-3}]$. The trade varies by target. We therefore describe Phi SeNPO as occupying a usable forgetting region of the recorded sweep, not as best on every metric.