SAGE: RETAIN-AWARE POST-HOC SANITIZATION OF FINAL UNLEARNING VECTOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Model (LLM) unlearning aims to remove undesirable knowledge or behaviors while preserving retained capabilities. Current unlearning methods all involve a trade-off between unlearning and retention. We have found that the retention activation bias can also be used to quantify the damage an unlearning method inflicts on retention, without considering the specific implementation of the unlearning process. This allows us to restore retention performance for various unlearning methods using a post-hoc approach. Therefore, we propose a complementary post-hoc setting to sanitize the final update vector without rerunning the original unlearning pipeline. In this setting, we design **SAGE**, **S**pectral **A**ctivation-**GE**ometry Sanitization, a source-agnostic correction for final unlearning updates. SAGE collects real module inputs from a small retain proxy, extracts their dominant activation geometry, and solves a source-anchored optimization objective in closed form, which suppresses update components aligned with high-energy retained directions while preserving the source method's forgetting carrier. Across multiple unlearning methods, model scales, and benchmarks, SAGE consistently relieves the retain–forget trade-off, identifying post-hoc sanitization of final vectors as a practical and underexplored axis for machine unlearning.

1 INTRODUCTION

Large language models (LLMs) (Bai et al., 2023; Grattafiori et al., 2024; Touvron et al., 2023; Wolf et al., 2020; Yang et al., 2025a) can memorize private information, copyrighted content, harmful behaviors, and other undesirable data patterns from pretraining and fine-tuning corpora (Carlini et al., 2021; Liu et al., 2024; Wang et al., 2025a; Jang et al., 2023; Yao et al., 2024b). This has made machine unlearning an increasingly important problem for safety, privacy, and regulatory compliance (Li et al., 2024; Yao et al., 2024a; Zhang et al., 2025). While exact retraining remains the cleanest deletion target in classical machine unlearning, it is often computationally prohibitive at modern model scales, motivating efficient approximate alternatives (Bourtoule et al., 2021; Neel et al., 2021). For LLMs, it is more challenging because knowledge is encoded in highly distributed representations learned from large-scale corpora, making it difficult to remove targeted information without inducing widespread collateral damage to general capabilities (Brown et al., 2020; Touvron et al., 2023).

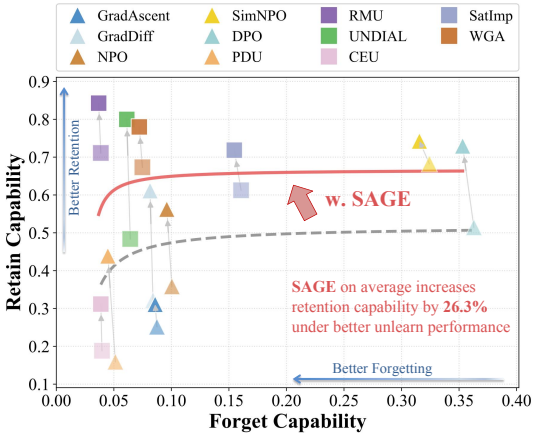

Figure 1: Overall Performance. Triangles denote loss-driven methods and squares denote constraint-guided methods. Points are averaged over multiple models and forgetting difficulties. Light markers and gray line fit original baselines; dark markers and red line fit methods **with SAGE**.

Recent work has produced a wide range of approximate LLM unlearning methods, including gradient and objective-based (Maini et al., 2024; Wang et al., 2025b; Fan et al., 2025; Entesari et al., 2025; Yang, 2025; Zhang et al., 2024), preference-style optimization (Dong et al., 2025b; Mekala et al., 2024), representation-level or localized-parameter interventions (Li et al., 2024; Shen et al., 2025), loss-reweighting methods (Wang et al., 2025a; Yang et al., 2025b), and task-vector methods (Ilharco

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.

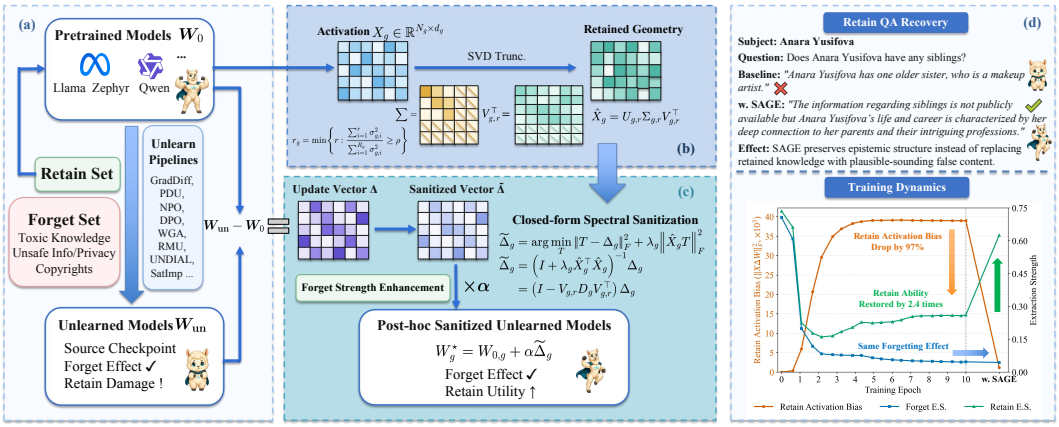

Figure 2: Overview of the proposed SAGE method: (a) Acquire post-hoc update vector from unlearning baselines; (b) Construct stable and denoised retain geometry from module-level input activations; (c) Apply closed-form spectral sanitization and amplifier to get unlearned models ; (d) Demonstration of retain restoration on a QA sample and drop on retain activation bias.

et al., 2023; Dong et al., 2025a; Cai et al., 2026; Kim et al., 2025). Furthermore, benchmarks such as TOFU (Maini et al., 2024), MUSE (Shi et al., 2025), and WMDP (Li et al., 2024) have made clear that forgetting must be balanced against leakage, utility, and retained behavior. Based on them, many lightweight interventions, ranging from inference-time control (Huang et al., 2025; Pawelczyk et al., 2024; Huang et al., 2024; Bhaila et al., 2025) to training-time plug-ins (Wang et al., 2025c; Zhou et al., 2026), have also emerged. Although they have already shown that adding a lightweight correction mechanism can be an effective way to further improve unlearning, most of them remain tightly coupled to specific stages of the original pipeline: training-time plug-ins must be incorporated during optimization and rely on rerunning or modifying the original unlearning process. Part (d) in Figure 2 shows that during source unlearning, retain activation bias rises sharply while retained ability drops substantially, even after the forgetting strength has largely stabilized. This suggests that the final unlearning update is still not fully retain-aware: the deployed update continues to contain components that disproportionately perturb retained activation directions, indicating that there remains meaningful room to improve retention by sanitizing the completed update itself. This motivates a decoupled post-hoc setting, in which the object of correction is no longer the training dynamics, but the final vector itself. We try to construct a sanitized vector to better trade off forgetting and retention, without revisiting the original unlearning process.

Therefore, we propose **SAGE**, **S**pectral **A**ctivation-**GE**ometry Sanitization, a post-hoc sanitizer for final unlearning vectors, without accessing the source training trajectory or rerunning unlearning algorithms. To construct a retained activation basis, we sample a small retained calibration set and collect module-level input activations through no-gradient forward passes. To reduce the output perturbation by minimizing the update response energy on the retained input geometry, we apply truncated Singular Value Decomposition (SVD) to identify a stable and denoised low-rank subspace. As the source unlearning method has already been optimized for the forget objective, its final parameter displacement is the most direct forgetting carrier available in the post-hoc setting. SAGE therefore performs source-anchored sanitization: it keeps the sanitized update close to the original source update to preserve the learned forgetting signal, while suppressing the components that produce large retained-activation responses. In the objective's closed-form solution, directions with larger retained singular energy are attenuated more strongly, whereas update components outside the dominant retained subspace are largely preserved. Finally, SAGE applies an amplifier to ensure effective forgetting, following the task-vector view of model editing and unlearning, where the direction of a parameter displacement determines the type of behavioral change and its magnitude controls the strength of that change (Ilharco et al., 2023; Cai et al., 2026).

Empirically, SAGE consistently improves the retain-forget trade-off across diverse source unlearning methods, model scales, and forget ratios, without rerunning the original unlearning pipeline. On TOFU, it increases average retention capability by 26.3% with a better unlearning effect. These improvements remain stable from 1B to 8B models and become more pronounced as the amount of content to be forgotten increases. What's more, SAGE improves model utility by 2.2% and

reduces privacy leakage by 6.2% on average. Its benefits further transfer beyond structured QA-style forgetting. SAGE improves retention capabilities by about 39.8% and 5.2% on MUSE and WMDP-cyber respectively, while largely preserving forgetting behavior. Notably, for TOFU dataset, SAGE uses roughly 3% of the full retain set and remains robust even under smaller proxy budgets. Together, these results suggest that post-hoc sanitization of final unlearning updates is a practical and underexplored design axis for machine unlearning. Our main contributions are as follows:

- We study a practical post-hoc unlearning setting to reduce retain-side collateral damage in which the final parameter update of a source unlearning method is sanitized after training, without rerunning or modifying the original pipeline.
- We propose **SAGE**, a module-wise closed-form sanitizer that builds retained activation geometry from a small retain set and applies a singular-value-aware soft spectral operator to suppress retained-response directions while preserving forgetting effect.
- Across multiple source unlearning methods, model scales, and benchmarks, we show that SAGE consistently improves the retention capability, while also improving utility and privacy leakage, with less time and computational resources.

2 Related Work

2.1 LLM Unlearning

Machine unlearning for LLMs is motivated by the need to remove private, copyrighted, or hazardous knowledge without retraining from scratch (Jang et al., 2023; Yao et al., 2024b). Since exact deletion is generally infeasible for modern LLMs, recent work has focused on approximate unlearning together with evaluation protocols that jointly measure removal and preservation. Benchmarks such as TOFU (Maini et al., 2024), WMDP (Li et al., 2024), MUSE (Shi et al., 2025) and OpenUnlearning (Dorna et al., 2025) have made the forget–retain trade-off a central concern, showing that effective unlearning should evaluate together with retained behavior, utility and leakage.

Most existing methods improve this trade-off during optimization. Gradient- and objective-based approaches modify the forget objective directly, including gradient ascent (Maini et al., 2024), NPO (Zhang et al., 2024), and SimNPO (Fan et al., 2025), while representation-level methods such as RMU (Li et al., 2024) and LUNAR (Shen et al., 2025) intervene on internal activations or hidden states to suppress unwanted knowledge. More recently, lightweight retain-aware plug-ins have been proposed: GRU (Wang et al., 2025c) rectifies retention-damaging update directions during training, and GU (Zhou et al., 2026) removes components aligned with the retain-gradient subspace. Other lightweight methods, such as offset unlearning (Huang et al., 2025), in-context unlearning (Pawelczyk et al., 2024), and soft prompting (Bhaila et al., 2025), reduce access requirements by acting at inference time rather than on model weights. In contrast, SAGE operates after training is complete and directly sanitizes the final deployed update, without modifying the original optimization loop.

2.2 Weight-Space Unlearning and Preservation

Our work is also related to methods that treat parameter differences as editable weight-space objects. Task Arithmetic (Ilharco et al., 2023) shows that fine-tuning deltas can act as task vectors whose addition, subtraction, or scaling steers model behavior without further training. However, single-vector unlearning is sensitive to fine-tuning configuration, scaling, and candidate selection. Task Simplex Arithmetic (Dong et al., 2025a) and NegMerge (Kim et al., 2025) improve robustness through multi-vector aggregation. PerTA (Cai et al., 2026) uses gradient or diagonal-Fisher estimates to rescale task vectors, addressing over-forgetting. These methods improve vector selection, merging, or parameter-wise merging to acquire a stable or retain-friendly unlearning vector. SAGE instead accepts final vectors and sanitizes it using module-level retain activation geometry.

A related preservation-oriented literature comes from model editing. ROME (Meng et al., 2022) edits factual associations through localized rank-one updates, and MEMIT (Meng et al., 2023) extends this paradigm to many edits by distributing updates across layers. AlphaEdit (Fang et al., 2025) reduces collateral damage by projecting edit perturbations into the null space of preserved knowledge keys. SAGE shares this retained-subspace preservation principle, but addresses a different correction

problem: its input is an already completed unlearning update whose source-specific forgetting behavior should be retained. Rather than solving for a specified factual edit, SAGE regularizes the corrected displacement toward the source update while penalizing its response on retained activations.

3 METHOD

3.1 PROBLEM SETUP

Existing unlearning methods often differ substantially in their training objectives, optimization procedures, and access assumptions, yet they share a common challenge: improving forgetting typically comes at the cost of retained capabilities. As shown in Figure 2, retain-side collateral damage can be reflected by retain activation bias, independent of implementation details of unlearning algorithms. Therefore, we propose a post-hoc final-update setting for LLM unlearning, providing a global correction targeted at accumulated updates instead of step-wise training-time control.

Given a base model W_0 and a source unlearned model W_{un}, we get the final vector from the pretrained model to the unlearned model. Rather than editing all parameters uniformly, we restrict SAGE to a structured set of editable modules $\mathcal{G}$, comprising the attention and MLP projection matrices, as they admit a meaningful input-side activation geometry. For each module g, we define source update as

$$\Delta_g = \begin{cases} W_{\mathrm{un},g} - W_{0,g}, & g \in \mathcal{G}, \\ 0, & g \notin \mathcal{G}. \end{cases} \tag{1}$$

3.2 RETAINED ACTIVATION GEOMETRY

Retain-side damage is not determined solely by the norm of the source update. More importantly, it depends on whether the update acts strongly along the dominant input directions associated with retained capabilities. To capture this structure, we run a no-gradient forward pass on a small retain calibration proxy D_{cal} for each editable module g and collect the module's real input activations into $X_g \in \mathbb{R}^{N_g \times d_g}$, where d_g is the module input dimension and N_g is the total number of collected token-level inputs.

We then compute the singular value decomposition $X_g = U_g \Sigma_g V_g^\top$, and retain the top r_g singular directions according to the cumulative-energy criterion. The resulting truncated operator emphasizes dominant retained directions while discarding low-energy and probable noisy components, providing a stable geometry for the post-hoc sanitizer and avoiding overfitting. This yields a truncated retained-geometry operator

$$\hat{X}_g := U_{g,r} \Sigma_{g,r} V_{g,r}^\top, \qquad \hat{C}_g := \hat{X}_g^\top \hat{X}_g = V_{g,r} \Sigma_{g,r}^2 V_{g,r}^\top. \tag{2}$$

Here, $\hat{X}_g$ captures the dominant retained input geometry of module g, while $\hat{C}_g$ is the corresponding input-side Gram operator. In particular, $\hat{C}_g$ encodes the relative importance of retained input directions and will serve as the geometry-aware weighting operator in the sanitizer below.

3.3 CLOSED-FORM SPECTRAL SANITIZATION

In the post-hoc setting, the source method's final parameter displacement is the only directly available carrier of the forgetting effect achieved during unlearning. Accordingly, the sanitizer should remain close to the source update and only correct the part that is most harmful to retained behaviors.

To this end, for each module g, SAGE optimizes a source-anchored objective with two complementary terms: a proximity term that keeps the sanitized update close to the source displacement, and a response penalty that suppresses update directions introducing output perturbation on retained activation geometry. For module g, let m_g denote the output dimension, and write the corresponding update in operator form as $\Delta_g \in \mathbb{R}^{d_g \times m_g}$. Concretely, we solve

$$\widetilde{\Delta}_g = \arg\min_T \|T - \Delta_g\|_F^2 + \lambda_g \left\| \hat{X}_g T \right\|_F^2. \tag{3}$$

The first term preserves the learned forgetting signal already encoded in the source update, while the second term penalizes retain-harming directions. Without the source anchor, minimizing only the response term would collapse to the trivial zero update.

Proposition 1 (Unique closed-form sanitizer)**.** *For each editable module g and any $\lambda_g \geq 0$, the objective in Eq. 3 is strongly convex in T and admits the unique minimizer*

$$\widetilde{\Delta}_g = \left(I + \lambda_g \hat{C}_g\right)^{-1} \Delta_g = \left(I - V_{g,r} D_g V_{g,r}^\top\right) \Delta_g, \; D_g = \operatorname{diag}\left(\frac{\lambda_g \sigma_{g,1}^2}{1+\lambda_g \sigma_{g,1}^2}, \ldots, \frac{\lambda_g \sigma_{g,r}^2}{1+\lambda_g \sigma_{g,r}^2}\right). \tag{4}$$

Proof sketch. Eq. 3 is a strictly convex quadratic in T, with Hessian $2(I + \lambda_g \hat{C}_g) \otimes I_{m_g}$. Since $\hat{C}_g \succeq 0$, the Hessian is positive definite. Setting the gradient to zero and applying Woodbury matrix identity yields Eq. 4. Full details are deferred to Appendix A.

This closed-form solution reveals that SAGE acts as a soft spectral sanitizer on the source update. Along each $v_{g,i}$ retained principal directions, the source update is shrunk by $\frac{1}{1+\lambda_g \sigma_{g,i}^2}$, which is determined by the retained singular energy, and directions orthogonal to $\operatorname{span}(V_{g,r})$ is left unchanged. So directions with larger retained energy are attenuated more strongly, and components outside the dominant retained subspace are largely preserved.

As a result, SAGE is not a hard null-space projection that indiscriminately removes all retained-subspace components as $\lambda_g \to \infty$. Instead, it performs a continuous, geometry-aware shrinkage that suppresses the most retain-sensitive directions while preserving as much of the source forgetting carrier as possible, especially when the source update uses shared directions to affect both forget and retain capabilities. Also, this operator doesn't amplify the retained-geometry response, reducing the retain-side disturbance.

Algorithm 1 SAGE Framework

Input: base model W_0, source model W_{un}, editable modules $\mathcal{G}$, retain calibration proxy D_{cal}, regularization strengths $\{\lambda_g\}_{g\in\mathcal{G}}$, energy threshold ρ, calibrated α
$\Delta \leftarrow W_{\text{un}} - W_0$
for each $g \in \mathcal{G}$ **do**
  collect retained module inputs X_g from D_{cal}
  compute $X_g = U_g \Sigma_g V_g^\top$
  choose r_g by energy threshold ρ
  $\hat{X}_g \leftarrow U_{g,r} \Sigma_{g,r} V_{g,r}^\top$
  $\hat{C}_g \leftarrow \hat{X}_g^\top \hat{X}_g$
  $\widetilde{\Delta}_g \leftarrow (I - V_{g,r} D_g V_{g,r}^\top)\Delta_g$
end for
choose α by matched-forgetting calibration
for each module g **do**
  if $g \in \mathcal{G}$ **then**
    $W_g^\star \leftarrow W_{0,g} + \alpha \widetilde{\Delta}_g$
  else
    $W_g^\star \leftarrow W_{un,g}$
  end if
end for
Return $W^\star$

Final Composition and Forget-Matched Calibration As the magnitude of sanitized unlearning vectors controls the strength of forgetting (Ilharco et al., 2023), we apply an amplifier and form the final model as

$$W_g^\star = W_{0,g} + \alpha \, \widetilde{\Delta}_g, \tag{5}$$

where α is a scaling coefficient selected by grid search on a benchmark-specific forget-matched calibration subset.

Mechanistic Discussion: Differential Suppression. To further interpret when post-hoc sanitization is beneficial, we compare the relative suppression induced by SAGE on retain and forget activations. For module g, define

$$S_g^{(r)} := 1 - \frac{\left\| X_g^{(r)} \widetilde{\Delta}_g \right\|_F^2}{\left\| X_g^{(r)} \Delta_g \right\|_F^2}, \qquad S_g^{(f)} := 1 - \frac{\left\| X_g^{(f)} \widetilde{\Delta}_g \right\|_F^2}{\left\| X_g^{(f)} \Delta_g \right\|_F^2}, \tag{6}$$

and let

$$\Gamma_g := S_g^{(r)} - S_g^{(f)}. \tag{7}$$

A positive Γ_g indicates that SAGE suppresses retain-side response more strongly than forget-side response, which helps explain why, after forget-matched calibration, the sanitized update can improve retention while preserving forgetting behavior. We study this quantity empirically in Section 4.5 and provide further analysis in A.6.

Table 1: Main TOFU results across Llama-3-1B,3B,8B on Forget-1%,5%,10% split, with retain/unlearn Extraction strength (ES), absolute privacy leak (|Priv. Leak|), and model utility (MU). Better results with SAGE are highlighted in **bold**, and retain-capability gains are shown in **orange**.

Method	Forget-1%				Forget-5%				Forget-10%			
	ES Re. ↑	ES Un. ↓	|Priv. Leak| ↓	MU ↑	ES Re. ↑	ES Un. ↓	|Priv. Leak| ↓	MU ↑	ES Re. ↑	ES Un. ↓	|Priv. Leak| ↓	MU ↑
Llama-3.2-1B-Instruct												
Vanilla	0.055	0.058	7.79	0.281	0.055	0.059	8.09	0.281	0.055	0.055	10.43	0.281
Fully Fine-tuned	0.736	0.743	100.00	0.598	0.736	0.730	99.99	0.598	0.736	0.712	99.46	0.598
NPO	0.583	0.150	76.62	0.588	0.131	0.071	**12.78**	0.437	0.205	0.075	**1.99**	0.528
NPO w. SAGE	**0.630**$^{\uparrow 0.047}$	**0.144**	**75.09**	**0.593**	**0.180**$^{\uparrow 0.049}$	**0.065**	26.70	**0.495**	**0.267**$^{\uparrow 0.062}$	**0.072**	14.05	**0.550**
SimNPO	0.672	0.415	**98.11**	0.593	0.580	0.211	96.77	0.578	0.637	0.155	95.12	0.586
SimNPO w. SAGE	**0.706**$^{\uparrow 0.034}$	**0.402**	98.58	**0.596**	**0.656**$^{\uparrow 0.076}$	**0.201**	**95.94**	**0.590**	**0.647**$^{\uparrow 0.010}$	**0.148**	**92.48**	**0.593**
RMU	0.216	0.032	**46.28**	0.526	0.663	0.033	**49.24**	0.585	0.707	0.033	58.81	0.592
RMU w. SAGE	**0.691**$^{\uparrow 0.475}$	**0.031**	84.53	**0.593**	**0.707**$^{\uparrow 0.043}$	**0.033**	51.12	**0.592**	**0.709**$^{\uparrow 0.002}$	**0.033**	**55.37**	**0.595**
UNDIAL	0.484	0.091	86.42	0.582	0.232	0.053	91.46	0.556	0.268	0.044	93.50	0.563
UNDIAL w. SAGE	**0.664**$^{\uparrow 0.180}$	**0.080**	**75.80**	**0.597**	**0.660**$^{\uparrow 0.428}$	**0.052**	**40.60**	**0.596**	**0.625**$^{\uparrow 0.358}$	**0.044**	**49.35**	**0.597**
SatImp	0.677	0.206	95.99	0.596	0.413	0.069	76.26	0.572	0.358	0.059	72.00	0.552
SatImp w. SAGE	**0.700**$^{\uparrow 0.023}$	**0.198**	**95.63**	**0.600**	**0.580**$^{\uparrow 0.167}$	**0.065**	**67.41**	**0.584**	**0.453**$^{\uparrow 0.095}$	**0.057**	**58.35**	**0.569**
WGA	0.683	0.141	**86.78**	0.598	0.555	0.034	**53.62**	**0.592**	0.627	0.033	60.71	0.590
WGA w. SAGE	**0.721**$^{\uparrow 0.038}$	**0.124**	88.90	**0.601**	**0.663**$^{\uparrow 0.107}$	**0.034**	54.40	0.591	**0.734**$^{\uparrow 0.108}$	**0.033**	**58.76**	**0.600**
Llama-3.2-3B-Instruct												
Vanilla	0.063	0.055	18.64	0.272	0.063	0.055	11.51	0.272	0.063	0.053	13.76	0.272
Fully Fine-tuned	0.885	0.920	100.00	0.665	0.885	0.887	100.00	0.665	0.885	0.888	99.72	0.665
NPO	0.760	0.201	80.93	0.667	0.140	0.060	**20.52**	0.466	0.132	0.060	**16.62**	0.529
NPO w. SAGE	**0.821**$^{\uparrow 0.062}$	**0.188**	**77.68**	**0.669**	**0.543**$^{\uparrow 0.403}$	**0.057**	30.72	**0.648**	**0.521**$^{\uparrow 0.389}$	**0.060**	21.52	**0.649**
SimNPO	0.839	0.562	**99.86**	0.652	0.624	0.264	97.99	0.640	0.583	0.212	96.76	**0.646**
SimNPO w. SAGE	**0.856**$^{\uparrow 0.017}$	**0.562**	**99.86**	**0.653**	**0.712**$^{\uparrow 0.088}$	**0.256**	**96.12**	**0.652**	**0.625**$^{\uparrow 0.042}$	**0.206**	**93.99**	0.645
RMU	0.320	0.038	**25.00**	0.612	0.830	0.033	54.30	**0.669**	**0.859**	0.033	62.83	**0.675**
RMU w. SAGE	**0.817**$^{\uparrow 0.498}$	**0.037**	70.90	**0.665**	**0.860**$^{\uparrow 0.030}$	**0.033**	**49.85**	0.668	0.854	**0.033**	**61.06**	0.666
UNDIAL	0.750	0.084	91.24	0.676	0.304	0.051	84.27	0.641	0.357	0.041	89.69	0.656
UNDIAL w. SAGE	**0.831**$^{\uparrow 0.081}$	**0.076**	**84.75**	**0.679**	**0.805**$^{\uparrow 0.501}$	**0.048**	**64.28**	**0.680**	**0.750**$^{\uparrow 0.393}$	**0.041**	**34.68**	**0.678**
SatImp	0.847	0.348	**96.05**	0.659	0.537	0.091	55.16	0.614	0.462	0.048	38.29	0.614
SatImp w. SAGE	**0.861**$^{\uparrow 0.014}$	**0.322**	96.19	**0.661**	**0.714**$^{\uparrow 0.178}$	**0.087**	**53.12**	**0.648**	**0.598**$^{\uparrow 0.136}$	**0.048**	**36.91**	**0.641**
WGA	0.841	0.206	88.28	0.663	0.623	0.033	**53.24**	0.641	0.633	0.036	**58.91**	0.648
WGA w. SAGE	**0.864**$^{\uparrow 0.023}$	**0.200**	**86.58**	**0.665**	**0.730**$^{\uparrow 0.107}$	**0.033**	53.65	**0.650**	**0.744**$^{\uparrow 0.111}$	**0.035**	59.11	**0.652**
Llama-3.1-8B-Instruct												
Vanilla	0.062	0.062	3.38	0.275	0.062	0.060	11.67	0.275	0.062	0.056	11.48	0.275
Fully Fine-tuned	0.992	0.977	100.00	0.627	0.992	0.972	100.00	0.627	0.992	0.979	99.94	0.627
NPO	0.866	0.152	68.25	**0.642**	0.176	0.064	**39.71**	0.556	0.233	0.071	40.45	0.601
NPO w. SAGE	**0.956**$^{\uparrow 0.090}$	**0.147**	**64.37**	0.639	**0.601**$^{\uparrow 0.424}$	**0.064**	41.15	**0.609**	**0.543**$^{\uparrow 0.310}$	**0.067**	**39.30**	**0.613**
SimNPO	0.939	0.566	**97.50**	0.615	0.699	0.303	97.66	**0.619**	0.568	0.230	97.00	0.602
SimNPO w. SAGE	**0.967**$^{\uparrow 0.029}$	**0.566**	98.37	**0.619**	**0.833**$^{\uparrow 0.134}$	**0.276**	**94.10**	0.614	**0.680**$^{\uparrow 0.112}$	**0.223**	**94.30**	**0.604**
RMU	0.833	0.074	55.37	**0.635**	**0.985**	0.039	**50.85**	**0.657**	**0.986**	0.033	**59.52**	**0.652**
RMU w. SAGE	**0.986**$^{\uparrow 0.152}$	**0.061**	**3.13**	0.630	0.981	**0.037**	51.22	0.626	0.979	**0.033**	60.45	0.626
UNDIAL	0.819	0.112	85.75	**0.688**	0.508	0.051	79.38	**0.690**	0.631	0.051	92.15	**0.689**
UNDIAL w. SAGE	**0.976**$^{\uparrow 0.157}$	**0.109**	**84.87**	0.669	**0.955**$^{\uparrow 0.447}$	**0.051**	**48.78**	0.651	**0.932**$^{\uparrow 0.301}$	**0.050**	**55.31**	0.644
SatImp	0.950	0.461	**94.62**	0.618	0.690	0.117	84.81	**0.627**	0.578	0.046	**5.66**	**0.596**
SatImp w. SAGE	**0.972**$^{\uparrow 0.022}$	**0.461**	96.25	**0.622**	**0.852**$^{\uparrow 0.161}$	**0.109**	**77.57**	0.611	**0.736**$^{\uparrow 0.157}$	**0.045**	23.62	0.596
WGA	0.946	0.124	63.75	**0.639**	0.614	0.033	54.85	0.599	0.537	0.033	57.99	0.593
WGA w. SAGE	**0.962**$^{\uparrow 0.016}$	**0.124**	**61.62**	0.634	**0.792**$^{\uparrow 0.178}$	**0.033**	**54.20**	**0.603**	**0.809**$^{\uparrow 0.272}$	**0.033**	**56.59**	**0.609**

4 Experiments

4.1 Experiment Setups

Datasets and Baselines. We evaluate SAGE on the OpenUnlearning benchmark (Dorna et al., 2025), focusing primarily on TOFU (Maini et al., 2024), a fine-grained benchmark with 4,000 QA pairs for fictitious author profiles. We use different forget-set sizes (Forget-1%, Forget-5%, and Forget-10%) and report main results on Llama-3-1B, 3B, 8B-Instruct (Grattafiori et al., 2024). Besides, we report results on MUSE (Shi et al., 2025), which evaluates memorization and unlearning of books and news articles through verbatim reproduction, question answering, and membership inference with Llama-2-7B (Touvron et al., 2023), and on WMDP (Li et al., 2024), an alignment-oriented benchmark of 3,668 hazardous-domain (biosecurity, cybersecurity, chemical security) multiple-choice questions with Zephyr-7B (Tunstall et al., 2024). SAGE is applied *post-hoc* to final update vectors produced by a source unlearning method under 10 training epochs for TOFU, MUSE and 80 optimization steps for WMDP, including Gradient Ascent (Maini et al., 2024), GradDiff (Maini et al., 2024), NPO (Fan et al., 2025), SimNPO (Fan et al., 2025), RMU (Li et al., 2024), UNDIAL (Dong et al., 2025b), CEU (Yang, 2025), SatImp (Yang et al., 2025b), WGA (Wang et al., 2025a), DPO (Rafailov et al., 2023) and PDU (Entesari et al., 2025).

Table 2: MUSE results on **Llama-2-7b-hf** with unlearn extraction strength (ES Un.), retain ROUGE (ROUGE Re.), and absolute privacy leakage (|Priv. Leak|). WMDP results on **zephyr-7b-beta** are reported with unlearn accuracy (Un. acc.) and MMLU accuracy (MMLU acc.). Better results with SAGE are highlighted in **bold**, and retain-capability gains are shown in **orange**.

Method	MUSE Books			MUSE News			WMDP cyber	
	ES Un. ↓	ROUGE Re. ↑	|Priv. Leak| ↓	ES Un. ↓	ROUGE Re. ↑	|Priv. Leak| ↓	Un. acc. ↓	MMLU acc. ↑
	Llama-2-7b-hf						zephyr-7b-beta	
Vanilla	0.010	0.680	8.16	0.020	0.560	4.72	0.445	0.585
Fully Fine-tuned	0.920	0.690	57.34	0.290	0.550	99.81	–	–
GradDiff	0.008	0.052	28.62	0.071	0.487	98.68	0.245	0.536
GradDiff w. SAGE	**0.008**	**0.610**$^{\uparrow 0.558}$	32.54	**0.070**	**0.496**$^{\uparrow 0.009}$	**97.08**	**0.244**	**0.561**$^{\uparrow 0.025}$
SimNPO	0.138	0.537	54.81	0.214	0.486	**99.87**	0.418	0.572
SimNPO w. SAGE	**0.106**	**0.611**$^{\uparrow 0.074}$	**54.66**	**0.212**	**0.532**$^{\uparrow 0.047}$	**99.87**	0.426	**0.587**$^{\uparrow 0.015}$
RMU	0.008	0.124	**19.91**	0.021	0.482	**25.59**	**0.261**	0.511
RMU w. SAGE	**0.008**	**0.336**$^{\uparrow 0.212}$	38.76	**0.019**	**0.496**$^{\uparrow 0.014}$	30.90	0.279	**0.570**$^{\uparrow 0.059}$
UNDIAL	0.024	0.627	18.45	0.013	0.189	99.45	**0.390**	0.565
UNDIAL w. SAGE	**0.024**	**0.700**$^{\uparrow 0.073}$	**18.05**	**0.013**	**0.264**$^{\uparrow 0.075}$	**89.86**	0.397	**0.584**$^{\uparrow 0.019}$
WGA	0.008	0.486	43.82	0.012	0.436	103.70	**0.348**	0.546
WGA w. SAGE	**0.008**	**0.619**$^{\uparrow 0.133}$	**25.55**	**0.011**	**0.468**$^{\uparrow 0.032}$	**102.60**	**0.348**	**0.570**$^{\uparrow 0.024}$
PDU	**0.008**	0.042	**54.48**	0.146	**0.501**	**99.79**	**0.243**	0.243
PDU w. SAGE	**0.008**	**0.605**$^{\uparrow 0.563}$	71.21	**0.141**	0.484	99.81	0.245	**0.255**$^{\uparrow 0.012}$

Evaluation Metrics. We report four aspects of unlearning quality: forgetting, retention, privacy, and utility, measured by Unlearn/Retain Extraction Strength (Carlini et al., 2021), Retain ROUGE (Lin, 2004), Privacy Leakage (Shi et al., 2025), and Model Utility (Maini et al., 2024), respectively. Extraction Strength quantifies the residual recoverability of target information. Model Utility is a composite measure of overall retained capability. On TOFU, it combines probability, ROUGE, and Truth Ratio evaluations, while on MUSE and WMDP, it is reflected through retained-task generation quality. For Privacy Leakage, we report the absolute value of the raw leakage score and evaluate methods by its magnitude only; lower values indicate a weaker forget-set membership signal and therefore better privacy protection.

4.2 Main Results

Performance on TOFU. Table 1 shows that SAGE consistently improves retained capabilities with a better forget performance on TOFU across model scales, forget ratios, and source baselines. Averaged over TOFU settings, SAGE increases Retain E.S. from 0.5869 to 0.7409, corresponding to a 26.3% relative improvement, while slightly reducing Unlearn E.S. from 0.1271 to 0.1227. These results indicate that post-hoc sanitization can substantially recover retained capability without weakening forgetting strength. SAGE also improves model utility from 0.609 to 0.623, and the average absolute privacy leakage decreases from 68.29 to 64.06, suggesting that suppressing retain-sensitive components in the final vectors can also mitigate collateral utility and privacy degradation.

The improvement is consistent across both model scale and forget difficulty. Across the 1B, 3B, and 8B models, the average Retain E.S. gains are 26.5%, 29.4%, and 23.5%, respectively, showing that the benefit of SAGE persists from small to larger backbones. Across forget ratios, the average absolute Retain E.S. improvement is +0.109 on Forget-1%, +0.196 on Forget-5%, and +0.158 on Forget-10%, indicating that SAGE remains effective under both milder and stronger unlearning regimes. Overall, these results support SAGE as a robust post-hoc correction layer that consistently improves retained capability while preserving or slightly improving forgetting quality.

Performance on MUSE and WMDP. MUSE involves longer-form Books and News content with open-ended generation behavior, while WMDP-cyber evaluates hazardous knowledge removal together with general capability preservation. On MUSE, SAGE improves retention capabilities while preserving forgetting: averaged over Books and News, ROUGE Re. increases from 0.371 to 0.518, ES Un. slightly decreases from 0.056 to 0.052. On WMDP-cyber, SAGE mainly improves the utility side of the safety–utility trade-off, increasing average MMLU accuracy from 0.496 to 0.521 across source baselines at comparable unlearning accuracy. These results show that SAGE remains effective on longer-form and safety-oriented unlearning tasks.

Table 3: Hyperparameter ablations of SAGE on **Llama-3-1B** under TOFU Forget-10% Split. Default parameters are shaded in blue. Best and second-best results are highlighted in **bold** and underlined.

Ablation	Setting	WGA w. SAGE				RMU w. SAGE			
		ES Re. ↑	ES Un. ↓	|Priv. Leak| ↓	MU ↑	ES Re. ↑	ES Un. ↓	|Priv. Leak| ↓	MU ↑
SVD Trunc. ρ	0.5	0.677	0.033	59.49	0.594	0.679	0.033	60.01	0.590
	0.7	0.699	0.033	59.24	0.596	0.696	0.033	59.89	0.591
	0.9	0.734	0.033	**58.76**	0.600	0.709	0.033	**55.37**	0.595
	0.95	**0.748**	0.033	58.40	**0.603**	0.716	0.033	55.48	**0.595**
	1.0	0.607	0.033	60.51	0.591	**0.726**	0.033	58.53	0.594
Retain-proxy Size	16	0.713	0.033	59.44	0.596	0.699	0.033	59.06	0.592
	64	0.724	0.033	58.84	0.599	0.705	0.033	55.88	0.593
	128	0.734	0.033	58.76	0.600	0.709	0.033	55.37	**0.595**
	256	**0.764**	0.033	**58.37**	**0.603**	**0.716**	0.033	**55.30**	**0.595**
Regularization λ	0.0	0.629	0.033	60.66	0.592	0.706	0.033	58.86	0.592
	0.01	0.722	0.033	58.94	0.599	0.704	0.033	56.15	0.592
	0.1	0.730	0.033	58.97	**0.603**	**0.712**	0.033	**55.16**	0.594
	1.0	0.734	0.033	58.78	0.601	0.707	0.033	55.57	0.592
	5.0	0.734	0.033	**58.76**	0.600	0.709	0.033	55.37	**0.595**
	10.0	**0.736**	0.033	58.76	0.600	0.709	0.033	55.39	0.594
Scaling Coefficient α	0.8	**0.781**	0.046	**5.23**	**0.602**	**0.719**	0.041	54.42	**0.595**
	1.0	0.769	0.037	38.07	0.601	0.709	0.033	55.37	0.595
	1.5	0.746	**0.033**	58.10	0.601	0.694	**0.033**	59.01	0.588
	3.0	0.649	**0.033**	55.64	0.596	0.591	**0.033**	55.41	0.571
	6.0	0.389	**0.033**	54.52	0.553	0.370	**0.033**	**30.19**	0.496

4.3 Ablation Study

SVD Trunc. ρ. As the truncation ratio ρ increases, retain-side performance generally improves because a more dominant retained activation geometry is preserved. However, values too close to full rank can over-constrain sanitization and hurt stability, privacy or utility; overall, $\rho = 0.9$ provides the best balance between forgetfulness, retention, and model utility.

Retain-proxy Size. SAGE is relatively insensitive to retain-proxy size: even small proxies already recover a useful estimate of the dominant retained geometry and yield consistent gains. We therefore use 128 retained examples by default.

Regularization λ. The regularization strength λ controls the trade-off between suppressing retain-sensitive directions and staying close to the unlearning vectors. On TOFU, performance improves quickly from $\lambda = 0$ and then saturates, motivating our default choice $\lambda = 5$; on MUSE and WMDP, smaller values may work better.

Scaling Coefficient α. The scaling coefficient α controls how much of the sanitized update is restored before being added back to the base model. As α increases, forgetting and utility often improve, but overly large values can reduce retained performance and model utility.

4.4 Integration with Other Plug-in Methods

We further study whether SAGE is compatible with training-time rectifiers such as GU (Zhou et al., 2026) and GRU (Wang et al., 2025c). Table 4 shows that applying SAGE after them generally improves retained capability while largely preserving forgetting. SAGE alone also generally outperforms either training-time rectifier. These results suggest gradient-level rectification does not fully eliminate retain-side damage, leaving room for complementary post-hoc sanitization.

4.5 Suppression Difference Analysis

To examine the mechanism discussed in Section 3.3, we analyze the module-wise suppression difference $\Gamma_g = S_g^{(r)} - S_g^{(f)}$, which measures whether SAGE suppresses retain-side response more strongly than forget-side response. In Figure 3 and Appendix G, the distribution is consistently shifted to the positive side, which indicates that SAGE tends to remove more retain-harming responses than forget-relevant responses at the module level. Such asymmetric suppression helps explain why SAGE can improve retention while preserving forgetting. Further analyses are shown in Appendix A.6.

Table 4: Results of SAGE with GU and GRU plug-ins for WGA, NPO, and DPO on TOFU over Forget-1%, 5%, 10% using **Llama-3-1B**. w. SAGE are shaded in blue, the Best and second best results are highlighted in **bold** and underline, and additional gains brought by SAGE are in **orange**.

Method	WGA				NPO				DPO			
	ES Re.↑	ES Un.↓	|Priv. Leak|↓	MU↑	ES Re.↑	ES Un.↓	|Priv. Leak|↓	MU↑	ES Re.↑	ES Un.↓	|Priv. Leak|↓	MU↑
Forget-1%												
Baseline	0.683	**0.141**	**86.78**	0.598	0.583	0.150	76.62	0.588	0.575	0.369	98.94	0.575
w. SAGE	**0.721**	**0.124**	88.90	0.601	0.630	**0.144**	75.09	0.593	0.631	**0.327**	97.87	0.583
w. GU	0.677	0.179	92.80	**0.603**	0.548	0.178	81.35	0.590	0.577	0.392	99.06	0.576
w. GU w. SAGE	0.691 ↑0.014	0.188	91.03	0.600	0.605 ↑0.056	0.160	**74.97**	0.591	**0.634** ↑0.057	0.349	**97.76**	**0.584**
w. GRU	0.681	**0.141**	**86.78**	0.599	0.584	0.150	76.86	0.588	0.574	0.362	98.94	0.575
w. GRU w. SAGE	0.715 ↑0.034	0.145	89.61	0.601	**0.635** ↑0.050	0.148	77.92	**0.593**	0.630 ↑0.056	**0.327**	97.87	0.583
Forget-5%												
Baseline	0.555	0.034	53.62	**0.592**	0.131	0.071	12.78	0.437	0.235	0.150	84.12	0.017
w. SAGE	**0.663**	0.034	54.40	0.591	0.180	**0.065**	26.70	0.495	0.349	**0.140**	71.31	0.013
w. GU	0.504	0.037	**49.20**	0.584	0.123	0.068	**1.27**	0.430	0.208	0.147	81.63	0.000
w. GU w. SAGE	0.482	0.036	53.69	0.575	0.161 ↑0.038	0.066	21.39	0.478	0.314 ↑0.105	0.146	**70.88**	0.000
w. GRU	0.552	**0.034**	53.74	0.589	0.131	0.069	13.59	0.441	0.235	0.150	84.17	**0.017**
w. GRU w. SAGE	0.568 ↑0.016	**0.034**	54.61	0.585	**0.182** ↑0.050	0.066	27.52	**0.497**	**0.357** ↑0.123	0.148	74.19	0.014
Forget-10%												
Baseline	0.627	0.033	60.71	0.590	0.205	0.075	1.99	0.528	0.316	0.192	93.95	0.137
w. SAGE	**0.734**	0.033	58.76	**0.600**	**0.267**	0.072	14.05	0.550	0.451	0.191	91.13	0.125
w. GU	0.592	0.033	60.43	0.590	0.178	0.072	6.18	0.508	0.293	0.189	93.53	0.132
w. GU w. SAGE	0.651 ↑0.058	0.033	**58.40**	0.593	0.225 ↑0.048	**0.071**	16.95	0.540	0.395 ↑0.101	**0.177**	**90.49**	0.081
w. GRU	0.632	0.033	60.79	0.593	0.197	0.075	**1.41**	0.525	0.318	0.194	94.00	**0.137**
w. GRU w. SAGE	0.681 ↑0.049	0.033	59.09	0.594	0.265 ↑0.068	0.072	14.89	**0.553**	**0.453** ↑0.135	0.191	91.10	0.123

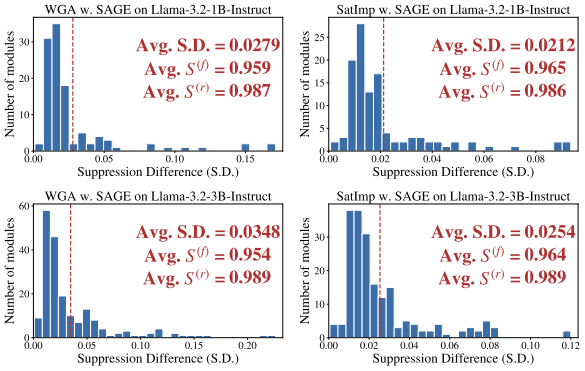

Figure 3: Distribution of suppression difference where the vertical red dashed line denotes the mean.

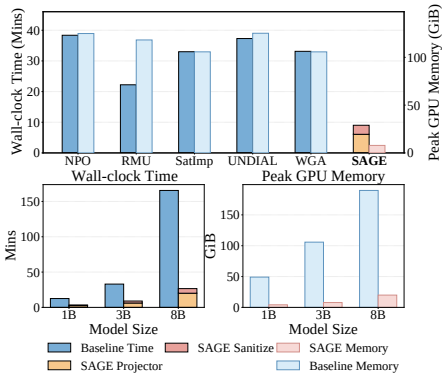

Figure 4: **Computation Resources** for Baselines and SAGE on different models.

4.6 Computation Resources

Figure 4 reports the computational overhead of SAGE. On Llama-3-3B, source unlearning training takes 22–38 minutes and exceeds 100 GiB peak GPU memory across five representative baselines, whereas SAGE takes about 9 minutes on average and uses only 7.86 GiB peak GPU memory. Moreover, the projector cache is built once for a given base model and retain proxy and can be reused across multiple source checkpoints, so only the lightweight sanitization step are repeated when applying SAGE on the same model. The scaling results show that efficiency advantage persists from 1B to 8B models, making SAGE a practical post-hoc correction under limited compute budgets.

5 Conclusion

We introduced **SAGE**, a post-hoc sanitizer of final unlearning updates that requires no retraining. Using a small retain proxy, **SAGE** applies source-anchored, closed-form spectral correction to reduce retain-side interference. Experiments on TOFU, MUSE, and WMDP show improved retain–forget trade-offs across source methods and model scales, with benefits complementary to training-time plug-ins. These results highlight final-update sanitization as a practical direction for machine unlearning.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, and interpret results; We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, and propose or refine hypotheses; To generate synthetic data sets, clean and reformat datasets are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, draft parts of a research paper, summarize or analyse existing literature. Specifically, we used these tools to help summarize and understand related work on machine unlearning and subspace-based methods, provide feedback on experimental configurations, and assist with implementing and debugging experimental code. During manuscript preparation, we used them to improve clarity and language, and refine the presentation of experimental results. We have reviewed all AI-assisted work and all citations were checked manually. We manually double-checked LLM-generated content, figures, and code for correctness and potential risks. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

SAGE aims to support responsible model unlearning by improving the balance between forgetting undesirable knowledge and preserving retained capabilities. Applications involving sensitive or safety-critical content require careful evaluation of residual memorization and retained utility.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 describe the method, with theoretical derivations in Appendix A. Appendices C and D provide evaluation protocols and implementation details. Additional results and ablations are reported in Appendices E and F. We will release our code in the future.

BIBLIOGRAPHY

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Elad Ben-Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Karuna Bhaila, Minh-Hao Van, and Xintao Wu. Soft prompting for unlearning in large language models. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Chengyi Cai, Zesheng Ye, Jiangchao Yao, Jianzhong Qi, Bo Han, Xiaolu Zhang, Feng Liu, and Jun Zhou. Per-parameter task arithmetic for unlearning in large language models. *arXiv preprint arXiv:2601.22030*, 2026.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2021.

Junhao Dong, Hao Zhu, Yifei Zhang, Xinghua Qu, Yew Soon Ong, and Piotr Koniusz. Machine unlearning via task simplex arithmetic. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025a.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025b.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Taha Entesari, Arman Hatami, Rinat Khaziev, Anil Ramakrishna, and Mahyar Fazlyab. Constrained entropic unlearning: A primal-dual framework for large language models. *arXiv preprint arXiv:2506.05314*, 2025.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. Offset unlearning for large language models. *Transactions on Machine Learning Research*, 2025.

Yu-Ting Huang, Pei-Yuan Wu, and Chuan-Ju Wang. Eco: Efficient computational optimization for exact machine unlearning in deep neural networks. In *International Conference on Machine Learning Workshop (ICML Workshop)*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Hyo Seo Kim, Dongyoon Han, and Junsuk Choe. Negmerge: Sign-consensual weight merging for machine unlearning. In *International Conference on Machine Learning (ICML)*, 2025.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models. In *Nature Machine Intelligence (Nat. Mach. Intell.)*, 2024.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, Muhao Chen, and Chaowei Xiao. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *International Conference on Learning Representations (ICLR)*, 2025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task of fictitious unlearning for llms. In *Conference on Language Modeling (COLM)*, 2024.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. *arXiv preprint arXiv:2409.13474*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *International Conference on Algorithmic Learning Theory (ALT)*, 2021.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners. In *International Conference on Machine Learning (ICML)*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

William F. Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. Llm unlearning via neural activation redirection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment. In *Conference on Language Modeling (COLM)*, 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. In *International Conference on Learning Representations (ICLR)*, 2025a.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data. In *International Conference on Learning Representations (ICLR)*, 2025b.

Yue Wang, Qizhou Wang, Feng Liu, Wei Huang, Yali Du, Xiaojiang Du, and Bo Han. Gru: Mitigating the trade-off between unlearning and retention for llms. In *International Conference on Machine Learning (ICML)*, 2025c.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Bo Yang. Ce-u: Cross entropy unlearning. *arXiv preprint arXiv:2503.01224*, 2025.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring criteria of loss reweighting to enhance llm unlearning. In *International Conference on Machine Learning (ICML)*, 2025b.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 2024a.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024b.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 2015.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics (AI Ethics)*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling (COLM)*, 2024.

Duo Zhou, Yuji Zhang, Tianxin Wei, Ruizhong Qiu, Ke Yang, Xiao Lin, Cheng Qian, Jingrui He, Hanghang Tong, Chengxiang Zhai, Heng Ji, and Huan Zhang. Geometric-disentangelment unlearning. *arXiv preprint arXiv:2511.17100*, 2026.

A PROOFS FOR SAGE

In this appendix, we provide the full derivations and proofs for the propositions in Section 3.3.

A.1 PROOF OF PROPOSITION 1

For a fixed editable module g, recall the objective

$$\widetilde{\Delta}_g = \arg\min_T \|T - \Delta_g\|_F^2 + \lambda_g \left\|\hat{X}_g T\right\|_F^2. \tag{3}$$

For notational simplicity, in this subsection we omit the group index g and write

$$\Delta := \Delta_g, \qquad \hat{X} := \hat{X}_g, \qquad \hat{C} := \hat{C}_g = \hat{X}^\top \hat{X}, \qquad \lambda := \lambda_g.$$

Then the objective becomes

$$\min_T \ \|T - \Delta\|_F^2 + \lambda \|\hat{X}T\|_F^2. \tag{8}$$

Using $\|A\|_F^2 = \mathrm{tr}(A^\top A)$, we have

$$\mathcal{L}(T) = \mathrm{tr}\big((T-\Delta)^\top (T-\Delta)\big) + \lambda\, \mathrm{tr}\Big((\hat{X}T)^\top (\hat{X}T)\Big) \tag{9}$$

$$= \mathrm{tr}\big((T-\Delta)^\top (T-\Delta)\big) + \lambda\, \mathrm{tr}\Big(T^\top \hat{X}^\top \hat{X} T\Big) \tag{10}$$

$$= \mathrm{tr}\big((T-\Delta)^\top (T-\Delta)\big) + \lambda\, \mathrm{tr}\Big(T^\top \hat{C} T\Big). \tag{11}$$

Differentiating with respect to T,

$$\frac{\partial \mathcal{L}}{\partial T} = 2(T-\Delta) + 2\lambda \hat{C} T. \tag{12}$$

Setting the gradient to zero gives

$$T - \Delta + \lambda \hat{C} T = 0,$$

which can be rearranged as

$$(I + \lambda \hat{C}) T = \Delta.$$

Therefore, the unique stationary point is

$$T^\star = (I + \lambda \hat{C})^{-1} \Delta. \tag{13}$$

Since $\hat{C} = \hat{X}^\top \hat{X} \succeq 0$, all eigenvalues of $\hat{C}$ are nonnegative. Hence all eigenvalues of $I + \lambda \hat{C}$ are strictly positive for any $\lambda \geq 0$. Therefore the quadratic objective in Eq. 8 is strongly convex in T, and the stationary point in Eq. 13 is the unique global minimizer.

Substituting back the group index g, we obtain

$$\widetilde{\Delta}_g = \left(I + \lambda_g \hat{C}_g\right)^{-1} \Delta_g = \left(I + \lambda_g \hat{X}_g^\top \hat{X}_g\right)^{-1} \Delta_g, \tag{14}$$

which proves Proposition 1.

A.2 DERIVATION OF THE SPECTRAL FORM

Recall that

$$\hat{X}_g = U_{g,r} \Sigma_{g,r} V_{g,r}^\top, \qquad \hat{C}_g = \hat{X}_g^\top \hat{X}_g = V_{g,r} \Sigma_{g,r}^2 V_{g,r}^\top.$$

Substituting this into Eq. 4, we obtain

$$\widetilde{\Delta}_g = \left(I + \lambda_g V_{g,r} \Sigma_{g,r}^2 V_{g,r}^\top\right)^{-1} \Delta_g. \tag{15}$$

We now use the identity

$$(I + VAV^\top)^{-1} = I - V\left(A^{-1} + V^\top V\right)^{-1} V^\top, \tag{16}$$

which is a special case of the Woodbury matrix identity. Since $V_{g,r}^\top V_{g,r} = I$, applying Eq. 16 with

$$V = V_{g,r}, \qquad A = \lambda_g \Sigma_{g,r}^2,$$

gives

$$\left(I + \lambda_g V_{g,r} \Sigma_{g,r}^2 V_{g,r}^\top\right)^{-1} = I - V_{g,r}\left((\lambda_g \Sigma_{g,r}^2)^{-1} + I\right)^{-1} V_{g,r}^\top. \tag{17}$$

Because $\Sigma_{g,r}$ is diagonal, we have

$$D_g = \lambda_g \Sigma_{g,r}^2 (I + \lambda_g \Sigma_{g,r}^2)^{-1} = \operatorname{diag}\left(\frac{\lambda_g \sigma_{g,1}^2}{1 + \lambda_g \sigma_{g,1}^2}, \ldots, \frac{\lambda_g \sigma_{g,r_g}^2}{1 + \lambda_g \sigma_{g,r_g}^2}\right). \tag{18}$$

Hence,

$$\left(I + \lambda_g V_{g,r} \Sigma_{g,r}^2 V_{g,r}^\top\right)^{-1} = I - V_{g,r} D_g V_{g,r}^\top.$$

Substituting this back into Eq. 15, we obtain

$$\widetilde{\Delta}_g = \left(I - V_{g,r} D_g V_{g,r}^\top\right) \Delta_g. \tag{19}$$

A.3 Analysis of Soft Spectral Sanitizer

From Eq. 4, the sanitizer operator is

$$S_g := I - V_{g,r} D_g V_{g,r}^\top.$$

Consider first a retained principal direction $v_{g,i}$, i.e., the i-th column of $V_{g,r}$. Since

$$V_{g,r}^\top v_{g,i} = e_i,$$

where e_i is the i-th standard basis vector, we have

$$S_g v_{g,i} = v_{g,i} - V_{g,r} D_g V_{g,r}^\top v_{g,i} \tag{20}$$

$$= v_{g,i} - V_{g,r} D_g e_i \tag{21}$$

$$= v_{g,i} - \frac{\lambda_g \sigma_{g,i}^2}{1 + \lambda_g \sigma_{g,i}^2} v_{g,i} \tag{22}$$

$$= \frac{1}{1 + \lambda_g \sigma_{g,i}^2} v_{g,i}. \tag{23}$$

Thus, along each retained principal direction $v_{g,i}$, the sanitizer scales the source update by

$$\frac{1}{1 + \lambda_g \sigma_{g,i}^2}.$$

Now let u be any vector orthogonal to $\operatorname{span}(V_{g,r})$. Then $V_{g,r}^\top u = 0$, so

$$S_g u = u - V_{g,r} D_g V_{g,r}^\top u = u.$$

Hence directions orthogonal to the retained principal subspace are left unchanged.

A.4 Analysis of Shrinkage Factor

The shrinkage factor along retained principal direction $v_{g,i}$ is

$$\frac{1}{1 + \lambda_g \sigma_{g,i}^2}.$$

As $\lambda_g \to \infty$, for every $\sigma_{g,i} > 0$,

$$\frac{1}{1 + \lambda_g \sigma_{g,i}^2} \to 0.$$

Therefore, all components of Δ_g lying in $\operatorname{span}(V_{g,r})$ are asymptotically removed.

On the other hand, for any direction $u \perp \operatorname{span}(V_{g,r})$, Section A.3 shows that $S_g u = u$ for all λ_g, so the orthogonal complement is preserved exactly.

Hence, in the limit $\lambda_g \to \infty$, the sanitizer converges to removing the component of Δ_g in the retained principal subspace while leaving the orthogonal complement unchanged.

A.5 Analysis of Retained Geometry Response

We again omit the group index g for simplicity. By Proposition 1,

$$\widetilde{\Delta} = (I + \lambda \hat{C})^{-1}\Delta.$$

The retained-geometry response of the sanitized update is

$$\|\hat{X}\widetilde{\Delta}\|_F^2 = \mathrm{tr}\Big((\hat{X}\widetilde{\Delta})^\top(\hat{X}\widetilde{\Delta})\Big) \tag{24}$$

$$= \mathrm{tr}\Big(\widetilde{\Delta}^\top \hat{X}^\top \hat{X}\widetilde{\Delta}\Big) \tag{25}$$

$$= \mathrm{tr}\Big(\widetilde{\Delta}^\top \hat{C}\widetilde{\Delta}\Big). \tag{26}$$

Substituting $\widetilde{\Delta} = (I + \lambda\hat{C})^{-1}\Delta$,

$$\|\hat{X}\widetilde{\Delta}\|_F^2 = \mathrm{tr}\Big(\Delta^\top (I + \lambda\hat{C})^{-1}\hat{C}(I + \lambda\hat{C})^{-1}\Delta\Big). \tag{27}$$

Now diagonalize $\hat{C}$ as

$$\hat{C} = V\Lambda V^\top, \qquad \Lambda = \mathrm{diag}(\mu_1, \ldots, \mu_d), \qquad \mu_i \geq 0.$$

Then

$$(I + \lambda\hat{C})^{-1}\hat{C}(I + \lambda\hat{C})^{-1} = V\ \mathrm{diag}\left(\frac{\mu_i}{(1+\lambda\mu_i)^2}\right)V^\top.$$

Similarly,

$$\hat{C} = V\ \mathrm{diag}(\mu_i)V^\top.$$

Since for each $\mu_i \geq 0$,

$$\frac{\mu_i}{(1+\lambda\mu_i)^2} \leq \mu_i,$$

it follows that

$$(I + \lambda\hat{C})^{-1}\hat{C}(I + \lambda\hat{C})^{-1} \preceq \hat{C}.$$

Applying this PSD ordering inside Eq. 27 yields

$$\|\hat{X}\widetilde{\Delta}\|_F^2 \leq \mathrm{tr}\Big(\Delta^\top\hat{C}\Delta\Big) = \|\hat{X}\Delta\|_F^2.$$

Restoring the group index proves

$$\|\hat{X}_g\widetilde{\Delta}_g\|_F^2 \leq \|\hat{X}_g\Delta_g\|_F^2. \tag{28}$$

A.6 Differential Suppression, Scaling, and Scope

Spectral interpretation of differential suppression. For a fixed editable module, omit the module index and write $\widetilde{\Delta} = P\Delta_{\mathrm{src}}$, where $P = (I + \lambda\widehat{C})^{-1}$ and $\lambda \geq 0$. Let $\widehat{C} = V\,\mathrm{diag}(\nu_i)V^\top$, with V a complete orthonormal basis and $\nu_i \geq 0$; eigenvalues outside the retained principal subspace are zero. Hold the module inputs X_r and X_f fixed, and assume $\|X_d\Delta_{\mathrm{src}}\|_F^2 > 0$ for $d \in \{r, f\}$. To interpret the spectral contributions, consider the case in which $C_d = X_d^\top X_d$ is diagonal in this basis for both $d = r$ and $d = f$. Define

$$c_{d,i} = v_i^\top C_d v_i, \qquad b_i = \|v_i^\top\Delta_{\mathrm{src}}\|_2^2, \qquad w_{d,i} = \frac{c_{d,i}b_i}{\sum_j c_{d,j}b_j}.$$

Here, $w_{d,i}$ is the fraction of source-update response energy carried by direction v_i. The gain in that direction is $a_i = (1 + \lambda\nu_i)^{-1}$, giving

$$\Gamma = \sum_i (w_{r,i} - w_{f,i})h_i, \qquad h_i = 1 - a_i^2.$$

Since h_i is non-decreasing in ν_i, this expression identifies a response-weighted spectral criterion for favorable differential suppression: $\Gamma > 0$ precisely when $\sum_i w_{r,i}h_i > \sum_i w_{f,i}h_i$. The relevant distinction is therefore how the retain and forget responses are distributed across directions with different suppression strengths.

Overlap between forget and retain directions. This criterion allows retain and forget responses to occupy the same directions: geometric separation or orthogonality is unnecessary. For finite λ, the gains a_i remain positive, so soft shrinkage keeps shared source-update components available for subsequent scaling. Favorable differences in their response weights can yield $\Gamma > 0$ even with overlapping supports. If $w_{r,i} = w_{f,i}$ for every direction, the spectral expression gives $\Gamma = 0$; the advantage is determined by response-weighted spectral differences rather than overlap alone.

Effect of the scaling coefficient. Positive scaling preserves the ordering of the directional gains, $g_i(\alpha) = \alpha/(1 + \lambda\nu_i)$. Moreover, for any fixed X_r and X_f with nonzero source-response energies, define

$$R_d = \frac{\|X_d\widetilde{\Delta}\|_F^2}{\|X_d\Delta_{\mathrm{src}}\|_F^2}, \qquad S_d(\alpha) = 1 - \frac{\|X_d(\alpha\widetilde{\Delta})\|_F^2}{\|X_d\Delta_{\mathrm{src}}\|_F^2} = 1 - \alpha^2 R_d.$$

It follows exactly that

$$\Gamma(\alpha) = S_r(\alpha) - S_f(\alpha) = \alpha^2(R_f - R_r) = \alpha^2\Gamma(1).$$

Thus, the favorable sign of differential suppression survives every positive rescaling. This identity holds without the common-eigenbasis condition used for the spectral interpretation above. If $R_f > 0$, matching the source forget-response energy for this module requires $\alpha^2 R_f = 1$, yielding

$$\frac{\|X_r(\alpha\widetilde{\Delta})\|_F^2}{\|X_r\Delta_{\mathrm{src}}\|_F^2} = \frac{R_r}{R_f} < 1 \quad \Longleftrightarrow \quad \Gamma(1) > 0.$$

These results establish that positive rescaling preserves SAGE's relative suppression advantage and, under module-level response-energy matching, allows the source forget-response energy to be restored with lower retain-response energy. They provide a mechanistic explanation for the retention gains observed with SAGE's calibrated updates.

B Design Rationale

B.1 Why Matrix-Valued Projection Modules and Activation Geometry?

We restrict SAGE to matrix-valued attention and MLP projection weights, rather than all trainable tensors. This design is motivated by both architectural and methodological considerations. Architecturally, attention projections and MLP weight matrices are the dominant affine transformations through which token representations are mixed and transformed in Transformers, and prior adaptation and editing work repeatedly treats these matrices as the primary intervention targets. For example, LoRA formulates adaptation directly on dense linear layers in Transformer models (Hu et al., 2022), while ROME and MEMIT identify Transformer feed-forward modules as key sites for factual association storage and editing (Meng et al., 2022; 2023).

More importantly, SAGE is defined through retained input geometry. Its sanitizer penalizes the response induced by an update on module inputs, naturally leading to a matrix-valued quantity of the form $\|X_g T\|_F^2$. This geometry is well-defined and informative for matrix-valued linear maps because it captures how an update acts along different input directions used by retained examples. This makes attention and MLP projection weights the natural objects of spectral activation-geometry sanitization.

Finally, activation geometry is particularly suitable for the post-hoc setting considered in this paper. Unlike training-time plugin methods such as GRU, which rely on gradient information during the unlearning process (Wang et al., 2025c), SAGE is designed to operate only on the final completed source update together with a small no-gradient retain proxy. Activation geometry therefore provides a source-agnostic and input-conditioned view of where a completed update is likely to interfere with retained behavior, without requiring the original training trajectory to be revisited.

B.2 Why Not Bias and Other 1D Parameters?

We exclude bias-like or other one-dimensional affine parameters not because they are unimportant in general. Indeed, BitFit shows that bias-only tuning can be effective in some settings (Ben-Zaken et al., 2022). Rather, such parameters are not the most natural objects for the particular activation-geometry sanitizer studied here.

For a matrix-valued update T, the retained perturbation is $X_g T$, and the penalty term $\|X_g T\|_F^2$ induces a nontrivial input-side geometry through $X_g^\top X_g$. This is precisely what allows SAGE to perform direction-selective shrinkage along dominant retained activation directions. By contrast, for a bias vector b, the perturbation is an input-independent shift of the form $\mathbf{1}b^\top$.

The corresponding penalty becomes $\|\mathbf{1}b^\top\|_F^2 = N_g\|b\|_2^2$, which is essentially isotropic and no longer depends on the retained input directions. In this case, the geometric interpretation of spectral shrinkage collapses, and the update behaves more like a uniform scaling term than an activation-conditioned correction.

The same issue arises for LayerNorm-style one-dimensional parameters. Although one could in principle incorporate such parameters by augmenting the module input with a constant feature, the resulting additional direction is trivial and does not induce the same informative retained-input geometry as matrix-valued projection updates. For this reason, we leave these parameters unchanged in the main method and focus on the modules for which activation-space sanitization is both principled and effective.

B.3 Why Can a Small Retain Proxy Suffice?

A natural question is why SAGE can work well even when the retained calibration proxy is relatively small. Our interpretation is that the method does not need to recover the full retained data manifold; rather, it only needs a stable estimate of the dominant retained activation subspace. When the retained activation covariance exhibits spectral concentration and a nontrivial eigengap, the leading singular subspace can often be estimated much more reliably than the full distribution.

This intuition is consistent with Davis–Kahan-style perturbation results, which show that subspace estimation error depends on covariance perturbation and eigengap (Yu et al., 2015). In other words, if the principal retained directions are stable and well separated, then a relatively small retained proxy may already be sufficient to recover the subspace that matters most for sanitization. We emphasize that this is a theory-backed intuition rather than a formal finite-sample guarantee for LLMs, but it provides a plausible explanation for the empirical robustness we observe on benchmarks such as TOFU.

C Evaluation Protocol

For TOFU, we search for the global scaling coefficient $\alpha \in [1, 40]$ using a binary search procedure, with a minimum resolution of 0.25. For the MUSE-News split, we search on $\alpha \in \{1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.25\}$. For MUSE-Books split, we search on $\alpha \in \{0.75, 1, 1.25, 1.5\}$. For WMDP, we search on $\alpha \in \{1, 1.5, 2\}$. Among the candidate values, we select the smallest α whose forget-side metric matches or improves upon the source baseline, so that the final comparison is conducted at the same or stronger forgetting level.

For TOFU, we calibrate α on the known target forget split (forget01/05/10) by matching the source model's extraction strength, following the known-forget-set model-selection convention discussed in OpenUnlearning (Dorna et al., 2025). To broaden our evaluation across calibration settings, we additionally report TOFU results with disjoint calibration and evaluation QA subsets in Appendix E.

For MUSE, we calibrate α against the reported unlearning metric, Extraction Strength (ES Un.). We randomly sample 20% of the evaluation split used for ES-Un computation in the corresponding MUSE domain as a hold-out calibration subset. This subset is used only for calibration and is removed from the final ES-Un evaluation.

For WMDP, we report results on the cyber split and calibrate α against Unlearn Accuracy. We therefore randomly sample 20% of the MCQ questions from the WMDP-cyber evaluation split as a hold-out calibration subset. These questions are used only for α selection and are excluded from all final reported WMDP-cyber results.

D Implementation Details

We use the OpenUnlearning implementations of the source methods and evaluation metrics (Dorna et al., 2025). For each completed source checkpoint W_{src}, SAGE sanitizes $\Delta_{\text{src}} = W_{\text{src}} - W_0$, where W_0 is the pre-unlearning target checkpoint, without rerunning source training.

D.1 Unlearning Baselines

Table 5 summarizes the 11 source methods. We retain their method-specific OpenUnlearning configurations unless stated otherwise; the shared source-training settings, including model, batch size, learning rate, training length, hardware, and precision, are given in Table 7. The benchmark-specific result tables identify which methods are evaluated in each setting. SAGE is applied to the same source checkpoint used for the corresponding baseline evaluation.

Method	Objective family	Reference / implementation
GA / GradDiff	Forget-loss ascent, without / with retain regularization	Maini et al. (2024)
WGA	Reweighted gradient ascent	OpenUnlearning (Dorna et al., 2025)
NPO	Reference-based negative preference optimization	Zhang et al. (2024)
SimNPO	Reference-free negative preference optimization	Fan et al. (2025)
DPO	Preference optimization over alternative and forget responses	OpenUnlearning (Dorna et al., 2025)
RMU	Representation misdirection with retain preservation	Li et al. (2024)
UNDIAL	Self-distillation with adjusted logits	Dong et al. (2025b)
CE-U	Cross-entropy-based unlearning	Yang (2025)
SatImp	Importance-based loss reweighting	Yang et al. (2025b)
PDU	Primal-dual constrained optimization	Entesari et al. (2025)

Table 5: Source unlearning methods. All methods use their OpenUnlearning implementations. GA and GradDiff are listed together but counted as separate methods.

D.2 Datasets

TOFU. TOFU contains 4,000 QA pairs about 200 fictitious authors (Maini et al., 2024). We use `forget01`, `forget05`, and `forget10`, containing 40, 200, and 400 QA pairs, with complementary `retain99`, `retain95`, and `retain90` splits. Forget examples define the source-unlearning targets; complementary retain examples supply retain-aware training objectives and the SAGE activation proxy. Paraphrased and perturbed answers support truth-ratio evaluation, holdout examples serve as nonmembers for privacy evaluation, and retain-author, real-author, and world-fact QA pairs support utility evaluation.

MUSE. We evaluate MUSE-Books and MUSE-News (Shi et al., 2025). Document-level `forget` and retain corpora supply the source-unlearning data, while the retain corpora also supply the SAGE activation proxy. Evaluation uses `verbmem` prefix–continuation examples for verbatim memorization, `knowmem` `forget_qa` and `retain_qa` for knowledge memorization and retention (100 QA pairs per split), and `privleak` forget and holdout examples for privacy evaluation. The retain geometry is constructed from document-side data rather than the QA evaluation splits.

WMDP. We evaluate the `wmdp-cyber` multiple-choice split (Li et al., 2024). Source unlearning uses `cyber-forget-corpus` and, for retain-aware methods, `cyber-retain-corpus`; SAGE also draws its activation proxy from the latter. We report cyber accuracy for forgetting and MMLU accuracy for general utility. Neither the cyber evaluation questions nor MMLU examples are used to construct the retain geometry. Proxy budgets and preprocessing are detailed in Section D.4; scaling calibration follows the benchmark-specific protocol in Appendix C.

D.3 Evaluation Metrics

We report memorization, privacy, and utility metrics using OpenUnlearning (Dorna et al., 2025). The post-hoc scale α is selected to match the source checkpoint's benchmark-specific forgetting score, as detailed in Appendix C. Metric matching refers to that selected score and does not imply equality of all forgetting diagnostics.

Probability and exact memorization. Answer probability is length-normalized and computed with teacher forcing:

$$P_\theta(y \mid x) = \exp\left(\frac{1}{|y|}\sum_{t=1}^{|y|}\log p_\theta(y_t \mid x, y_{<t})\right).$$

Paraphrased probability uses the paraphrased reference answer. Exact memorization is the fraction of reference-answer tokens that are top-1 predictions under the same teacher-forced context (Tirumala et al., 2022; Carlini et al., 2021). Both are averaged over evaluation examples; lower forget-side and higher retain-side scores are preferred.

Extraction Strength. Extraction Strength measures how much prefix information is needed before the model can reconstruct the remaining suffix (Carlini et al., 2021; Dorna et al., 2025). Let $y_{\le k}$ be a prefix and $y_{>k}$ be the remaining suffix. We define

$$\mathrm{ES}_\theta(x, y) = 1 - \frac{1}{T}\min\{k : \mathrm{Decode}_\theta(x, y_{\le k}) = y_{>k}\}. \tag{29}$$

A higher ES means that the model can recover the target continuation from a shorter prefix, indicating stronger memorization.

ROUGE and paraphrased ROUGE. ROUGE(Lin, 2004) measures lexical overlap between generated outputs and reference answers. For generation-based evaluation, the model output $\hat{y} = f_\theta(x)$ is compared with the ground-truth answer y, and we report ROUGE-L or the benchmark-specified ROUGE variant:

$$\mathrm{ROUGE}_\theta(x, y) = \mathrm{ROUGE}(\hat{y}, y). \tag{30}$$

On TOFU, paraphrased ROUGE evaluates whether the model still answers correctly when the question is reworded. This reduces the risk of overestimating forgetting merely because the model fails on the exact training prompt but still preserves the underlying knowledge.

Truth ratio and forget quality. We follow OpenUnlearning's forget-side configuration with the `closer_to_1_better` aggregator (Dorna et al., 2025). For forget example i, let $\bar{\ell}_{c,i}$ and $\bar{\ell}_{p,i}$ denote the mean token-normalized negative log-likelihoods of the paraphrased correct and perturbed answers, respectively, averaging across answer variants when present. Define $p_{c,i} = \exp(-\bar{\ell}_{c,i})$ and $p_{p,i} = \exp(-\bar{\ell}_{p,i})$. The example-level ratio and the reported forget score are

$$r_i = \frac{p_{p,i}}{p_{c,i} + \epsilon}, \qquad \text{F-TR} = \frac{1}{n_f}\sum_{i=1}^{n_f}\min\left\{r_i, \frac{1}{r_i + \epsilon}\right\}, \qquad \epsilon = 10^{-10}. \tag{31}$$

Higher F-TR indicates closer agreement between correct- and perturbed-answer scores. Forget Quality (FQ) is the two-sample Kolmogorov–Smirnov test p-value comparing the example-level raw-ratio distributions of the unlearned model and a retain-only reference model on the same forget examples. Higher FQ follows the benchmark reporting convention and indicates less evidence against equality of the tested distributions.

Membership inference and privacy leakage. We report attack AUCs for forget members versus holdout nonmembers using loss (`MIA-L`), Min-k (`MIA-MK`), Min-k++ (`MIA-MK++`), ZLib-normalized likelihood (`MIA-Z`), and gradient-norm scores (`MIA-GN`) where included (Shi et al., 2024; Dorna et al., 2025). An uninformative attack has AUC near 0.5; values near zero should not be

interpreted as intrinsically private, since reversing the score can recover discrimination. Using the AUC convention of our evaluator, signed privacy leakage is

$$\mathrm{PrivLeak} = 100\,\frac{(1-\mathrm{AUC}_f)-(1-\mathrm{AUC}_r)}{1-\mathrm{AUC}_r+\epsilon},$$

where AUC_f is the forget-versus-holdout attack AUC and AUC_r is the benchmark's reference AUC. The main tables report $|\mathrm{PrivLeak}|$ (lower is better); detailed tables retain the signed value (closer to zero is better).

Model utility and multiple-choice accuracy. TOFU Model Utility (MU) aggregates probability, ROUGE, and truth-ratio components on retain authors, real authors, and world facts using their harmonic mean:

$$\mathrm{MU} = \frac{m}{\sum_{j=1}^{m}(u_j+\epsilon)^{-1}},$$

where u_j are the normalized utility components; higher MU is better. For WMDP and MMLU, options are scored by conditional log-likelihood and accuracy is averaged over questions using the evaluation-harness configuration. Lower WMDP-cyber accuracy and higher MMLU accuracy are preferred.

Forget fluency. Where evaluated, forget fluency is the mean of $1-g_\psi(\mathrm{gibberish} \mid \hat{y})$ over generated forget responses. Higher values indicate more fluent responses; this is a generation-quality diagnostic, not a forgetting guarantee.

D.4 Retain Activation Collection for SAGE

SAGE estimates retain-side geometry from a small calibration set drawn only from the training-side retain data of each benchmark. No forget evaluation examples, retain evaluation QA pairs, MIA holdout examples, or WMDP multiple-choice evaluation questions are used for constructing the projector. For every benchmark, we load the original target model W_0 and register forward hooks on the editable linear modules. For a target linear module g, the hook records the module input activation $X_{r,g}$ and accumulates the retain Gram matrix

$$C_g = \sum_{x\in\mathcal{D}_{\mathrm{cal}}} X_{r,g}(x)^\top X_{r,g}(x). \tag{32}$$

We accumulate this quantity batch by batch rather than storing all raw activations. When the batch contains padding, we use the attention mask to remove padded token positions before forming $X_{r,g}^\top X_{r,g}$; thus padding tokens do not contribute to the retain geometry. The accumulated Gram matrices are then used to form the module-wise retain projector.

TOFU. TOFU is QA-formatted, so we use the same question–answer preprocessing as the source unlearning pipeline. For each forget split, the calibration data are taken from the complementary official retain split:

`forget01` $\rightarrow$ `retain99`, `forget05` $\rightarrow$ `retain95`, `forget10` $\rightarrow$ `retain90`.

Each retain example is tokenized as one supervised QA sequence with maximum length 512. The projector construction uses a deterministic subset of the retain split under a fixed calibration budget of 128 examples. Since TOFU examples are short QA pairs, we do not concatenate multiple examples into a single sequence; each example remains an independent QA instance, and padded positions are removed by the attention mask during activation collection.

MUSE-News. MUSE contains long document-level retain data, while part of the final evaluation is QA-formatted. For retain projector construction, we use the document-side retain corpus rather than the QA evaluation split. Specifically, for MUSE-News we sample 256 raw documents from the document-level retain split with seed 42. Each selected document is processed as one completion-style sequence with maximum length 2048. We use sample-preserving preprocessing: a retained news document is truncated or padded as a single sequence, and it is not concatenated with other retained documents. This keeps the activation context aligned with individual retained documents while avoiding leakage from the MUSE retain QA evaluation set.

MUSE-Books. MUSE-Books documents are much longer than typical news articles, so we use a document-preserving chunking strategy. We sample 3 raw documents from the document-level retain split with seed 42. Each selected document is tokenized and split into consecutive chunks of 2048 tokens, with at most 10 chunks kept per document. Chunks are written as separate calibration examples and are never concatenated across different books. If the final chunk of a selected document is shorter than 2048 tokens, it is padded to the maximum length, and padded positions are again excluded by the attention mask when accumulating the Gram matrices.

WMDP. For WMDP, we construct retain geometry from the text corpus used by the training pipeline, not from the WMDP multiple-choice evaluation benchmark or MMLU. In the cyber experiments, the retain source is the cyber retain corpus. We sample 128 retained documents with seed 42 and split each document into consecutive 2048-token chunks with at most 10 chunks per document. The resulting chunks are used as completion-style calibration examples with document-preserving boundaries. All emitted chunks are used for Gram accumulation. This choice makes the SAGE calibration data consistent with the retain data available during WMDP training, while keeping it separate from the downstream multiple-choice evaluation.

D.5 SAGE Hyperparameter Settings

SAGE is applied after source unlearning has completed. For each benchmark/model split, we use one retained-geometry cache and reuse it across all source methods. We edit transformer-block linear projections $(W_q, W_k, W_v, W_o, W_{\text{gate}}, W_{\text{up}}, W_{\text{down}})$; tied output heads are not treated as separate editable tensors. Unless stated otherwise, non-editable parameters are restored to their initial unlearned values during post-hoc injection. For each editable module g, SAGE computes

$$\widetilde{\Delta}_g = (I + \lambda_g \widehat{C}_g)^{-1} \Delta_{\text{src},g}.$$

Projector caches and sanitized task vectors are stored in bfloat16. Table 6 lists the main post-hoc hyperparameters.

Benchmark	**Retain calibration budget**	**Sequence construction**	**Rank**	λ
TOFU	128 examples from the complementary retain split	One QA example per sequence, max length 512	$\rho = 0.9$	5
MUSE-News	256 retained news documents	One document per sequence, max length 2048	Full rank ($\rho = 1.0$)	10^{-5}
MUSE-Books	3 retained books, up to 10 chunks per book	Consecutive 2048-token chunks, no cross-document concatenation	Full rank ($\rho = 1.0$)	10^{-5}
WMDP-Cyber	128 retained cyber documents, up to 10 chunks per document	Consecutive 2048-token chunks, no cross-document concatenation	Full rank ($\rho = 1.0$)	10^{-6}

Table 6: Main SAGE hyperparameters. All calibration examples are sampled with seed 42 from training-side retain data. Padding positions are removed using the attention mask before accumulating retained Gram matrices.

Source training hyperparameters. Table 7 summarizes the source unlearning stage. All source methods use the benchmark-specific OpenUnlearning trainer unless explicitly rerun for an ablation. All source training runs are conducted on 4 NVIDIA A6000 GPUs.

Runtime and precision settings. SAGE uses bfloat16 model loading with FlashAttention when supported. Retained Gram matrices are accumulated module by module, so raw activation matrices are not cached in full.

Benchmark	Model(s)	Batching	Learning rate	Training length	Distributed / precision settings
TOFU	Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct	Effective batch size 32; per-device batch size 4 with accumulation 2 for 1B/3B, and per-device batch size 2 with accumulation 4 for 8B	10^{-5} by default	10 epochs by default	4 NVIDIA A6000 GPUs, DeepSpeed ZeRO-3, bfloat16, FlashAttention, gradient checkpointing
MUSE	Llama-2-7B target models for News and Books	Effective batch size 32; per-device batch size 2, gradient accumulation 4	10^{-5} by default	10 epochs by default	4 NVIDIA A6000 GPUs, DeepSpeed ZeRO-3, bfloat16, FlashAttention, gradient checkpointing
WMDP	Zephyr-7B-beta	Effective batch size 32; per-device batch size 2, gradient accumulation 4	10^{-5}	80 optimization steps	4 NVIDIA A6000 GPUs, DeepSpeed ZeRO-3, bfloat16, FlashAttention, gradient checkpointing

Table 7: Source-model training hyperparameters. These settings describe the baseline unlearning stage that produces W_{src}; SAGE is applied only after this stage. Unless explicitly stated for a rerun or ablation, we keep the original method-specific OpenUnlearning hyperparameters.

D.6 Stage-Wise Computational Cost and Cache Amortization

Measurement setting and accounting scope. We profile SAGE on TOFU Forget-10% using Llama-3.2-3B-Instruct and NVIDIA A6000 GPUs in Table 8. Per-source wall-clock measurements are averaged over WGA, RMU, UNDIAL, NPO, and SatImp. We distinguish three quantities: the mean measured cost of an operation across source checkpoints, the one-time construction cost of a reusable retain-geometry cache, and the amortized cost per checkpoint when that cache is shared. Averaging measurements across methods is distinct from amortizing shared work across methods.

We further separate spectral processing (cache construction and update sanitization) from the forward evaluations used to select the final scale α. For the measurements below, mean spectral-processing time with one cache construction is $6.07 + 2.95 = 9.02$ minutes. Including scale calibration gives 16.47 minutes. These are different accounting scopes for the same pipeline. All SAGE totals below exclude source-unlearning training and final benchmark evaluation; the latter is not part of selecting α.

Cache reuse. The spectral operator depends on the pre-unlearning model, editable modules, retain proxy and preprocessing, and fixed (ρ, λ). It can be reused across source checkpoints sharing this configuration. Let T_{cache} be its one-time construction time, and let $T^{(j)}_{\mathrm{sanitize}}$ and $T^{(j)}_{\mathrm{cal}}$ denote the sanitization and scale-calibration times for source j. For K sources, calibrated post-hoc time is

$$T_{\mathrm{post}}(K) = T_{\mathrm{cache}} + \sum_{j=1}^{K} \left(T^{(j)}_{\mathrm{sanitize}} + T^{(j)}_{\mathrm{cal}} \right). \tag{33}$$

Using the measured per-source means,

$$\overline{T}_{\mathrm{spectral}}(K) = \frac{6.07}{K} + 2.95, \tag{34}$$

$$\overline{T}_{\mathrm{post}}(K) = \frac{6.07}{K} + 2.95 + 7.45 = \frac{6.07}{K} + 10.40, \tag{35}$$

Table 8: Stage-wise profiling on Llama-3.2-3B-Instruct. Time is wall-clock minutes; memory is the reported peak GPU memory per device in GiB. Per-source times are averaged over five source methods. A dash denotes an unavailable stage measurement. The reported sanitization-stage GPU footprint is not the total storage required for model checkpoints or host memory. Rows marked as totals include their constituent stages and should not be added again.

Component	Scope	GPUs	Time	Peak/device
Activation collection + Gram accumulation	Shared	1	2.97	6.63
Spectral decomposition	Shared	1	1.07	7.24
Spectral operator construction	Shared	1	0.23	6.94
Cache serialization and bookkeeping	Shared	1	1.80	–
Cache construction, total	Shared	1	**6.07**	**7.86**
Update sanitization	Per source	1	2.95	0.01
α calibration	Per source	1	7.45	16.00
Per-source processing, total	Per source	1	**10.40**	**16.00**
Baseline evaluation (reference)	Per source	1	5.98	16.00
Source-unlearning training (reference)	Per source	4	32.80	31.41

Table 9: Amortized cost per source checkpoint, calculated from the measured stage means. Spectral time excludes α calibration; calibrated post-hoc time includes it. GPU-hours refer to the single-GPU calibrated post-hoc pipeline.

K	Cache time	Spectral time	Calibrated time	GPU-hours
1	6.07	9.02	16.47	0.27
3	2.02	4.97	12.42	0.21
5	1.21	4.16	11.61	0.19
10	0.61	3.56	11.01	0.18

in minutes per checkpoint. Table 9 gives the implied amortization at several K values; these rows are calculated from the measured stage means, not separate profiling experiments. Cache reuse removes repeated geometry construction, whereas source-specific scale calibration is still required.

GPU accounting. For a stage occupying n_{GPU} devices for t minutes, we report allocated GPU-hours as $n_{\mathrm{GPU}}t/60$. Source training therefore costs approximately 2.19 GPU-hours on average, whereas the calibrated post-hoc cost is approximately 0.27 GPU-hours for $K = 1$ and 0.19 for $K = 5$. Peak memory is a per-device quantity and is not summed across sequential stages. Spectral processing has a reported peak of 7.86 GiB; including α calibration gives 16.00 GiB. These post-hoc costs are additional to the source-unlearning stage.

Arithmetic complexity. For each editable module g, let N_g be the number of valid retained tokens, d_g and m_g its input and output dimensions, and r_g the retained rank. One no-gradient model forward pass collects the required inputs. Writing its cost as T_{forward}, Gram accumulation costs $O(N_g d_g^2)$ per module. Let E_g denote the cost of the spectral decomposition actually used. Dense eigendecomposition costs $O(d_g^3)$; a partial solver can have a smaller cost depending on the requested rank, iterations, and tolerance. Truncating the output of a dense decomposition does not reduce its decomposition cost to $O(d_g^2 r_g)$.

Constructing a dense cached operator $P_g = I - V_{g,r} D_g V_{g,r}^\top$ costs $O(d_g^2 r_g)$. Applying it to $\Delta_g \in \mathbb{R}^{d_g \times m_g}$ costs $O(d_g^2 m_g)$. Thus the arithmetic costs are

$$T_{\mathrm{cache}}^{\mathrm{arith}} = T_{\mathrm{forward}} + O\left(\sum_{g\in\mathcal{G}} [N_g d_g^2 + E_g + d_g^2 r_g]\right), \tag{36}$$

$$T_{\mathrm{sanitize}}^{\mathrm{arith}} = O\left(\sum_{g\in\mathcal{G}} d_g^2 m_g\right). \tag{37}$$

Measured wall-clock costs additionally include data movement and serialization. If J scale candidates are evaluated, calibration requires J corresponding metric evaluations on the calibration data. The exact cost depends on the selected metric, sequence lengths, and candidate search.

E Additional Main Results

E.1 Detailed Results for Main Experiments

Table 11, 12, 13 reports the full TOFU results on Llama-3-1B,3B,8B-Instruct under three data-retention settings: Forget01/Retain99, Forget05/Retain95, and Forget10/Retain90. The appendix table keeps the complete metric panel, including the unlearn and retain extraction strength (ES-Un. / ES-Re.), forget-side QA metrics (F-QA Prob. and F-QA Rg.), forget quality (FQ), forget truth ratio (F-TR), model utility (MU), privacy score (Priv.), and four membership-inference diagnostics (MIA-L, MIA-MK, MIA-MK++, and MIA-Z). For readability, we abbreviate the metric names as follows: ES = extraction strength, F-QA = forget question answering, Rg. = ROUGE, FQ = forget quality, F-TR = forget truth ratio, MU = model utility, Priv. = privacy, and MIA = membership inference attack.

Table 14 additionally evaluates SAGE under a disjoint forget-side calibration protocol. For each forgotten target, we split the original QA items into a 20% calibration subset and an 80% evaluation subset. We report results for four source methods on Llama-3.1-8B-Instruct. SAGE improves ES-Re. in all 12 settings, with an average absolute gain of 0.215. Held-out ES-Un. remains close on average: 0.109 for the source models and 0.111 after SAGE. Overall, these results show that SAGE's retention gains persist under disjoint forget-side calibration, with generally similar held-out forgetting strength.

Table 15, 16 report the full MUSE-News and MUSE-Books results on **Llama-2-7b-hf**. The appendix tables keep the complete MUSE metric panel, including forgetting metrics, retain metrics, privacy leakage, and four membership-inference diagnostics. Following the MUSE and OpenUnlearning evaluation protocol, we report unlearn extraction strength (ES-Un.), retain ROUGE (ROUGE-Re.), signed privacy leakage (Priv. Leak), forget-side knowledge memorization measured by ROUGE (Forget Know. R.), forget-side verbatim memorization measured by ROUGE (Forget Verb. R.), and retain extraction strength (ES-Re.). Lower ES-Un., Forget Know. R., and Forget Verb. R. indicate stronger removal of the target content, while higher ROUGE-Re. and ES-Re. indicate better preservation of retained content. Priv. Leak reports the signed privacy-leakage score; in the main paper, we use its absolute magnitude when summarizing privacy leakage. We additionally include four MIA-based diagnostics, MIA Loss, MIA Min-K, MIA Min-K++, and MIA Zlib, which correspond to loss-based, Min-K probability, Min-K++ normalized probability, and zlib-normalized likelihood membership-inference variants.

We further expanded our experiments to other model architectures, including Qwen, as shown in Table 17.

E.2 Extension to LoRA and Multimodal Models

Multimodal extension. We evaluate the multimodal and LoRA extensions of SAGE on FIUBench (Ma et al., 2025) Forget-10% using LLaVA-Phi-Mini-3B in Table 10. In a vision–language model, SAGE replaces text-only calibration inputs with retained image–question–answer triples passed through the full multimodal forward path. Each updated module uses its actual inputs to construct a module-specific retained operator R_g: projector layers use visual or intermediate projector features, while language modules use the multimodal token sequence after visual-token insertion. SAGE is applied only to modules updated by the source method. The vision tower is frozen and requires no sanitization.

LoRA extension. For a LoRA module, the update is $\Delta W_{\mathrm{LoRA}} = \frac{\alpha_{\mathrm{LoRA}}}{r} BA$. Let

$$R = I - V \operatorname{diag}\left(\frac{\lambda\mu_i}{1+\lambda\mu_i}\right) V^\top \tag{38}$$

Table 10: FIUBench Forget-10% results on **LLaVA-Phi-Mini-3B**. We report calibration, held-out forgetting, and retention metrics. Bold values indicate the better score within the GD–SAGE pair; the shaded row denotes SAGE.

	Calibration	Held-out forgetting				
Method	P_{GT}	Prob. ↓	ROUGE-L ↓	Truth Ratio ↑	APE ↓	Exact Match ↓
Vanilla	–	0.9322	0.9320	0.2597	0.3839	0.6457
GD	0.0124	**0.0084**	**0.0127**	0.1955	**0.0012**	**0.0018**
GD w. SAGE	0.0122	0.0144	0.0190	**0.2380**	0.0020	0.0053

	Held-out forget diagnostics		Retention		
Method	Min-K ↓	Min-K++ ↓	Prob. ↑	ROUGE-L ↑	GT Loss ↓
Vanilla	0.6858	0.7981	0.9006	0.9261	0.0734
GD	**0.0017**	**0.0043**	0.4393	0.0471	23.6361
GD w. SAGE	0.0041	0.0083	**0.6657**	**0.2125**	**11.1772**

be SAGE's input-side spectral shrinkage operator constructed from the retained activations of this module, and let β_{SAGE} be the matched-forgetting scale. Then

$$\widetilde{\Delta W}_{\mathrm{LoRA}} = \beta_{\mathrm{SAGE}} \Delta W_{\mathrm{LoRA}} R = \frac{\alpha_{\mathrm{LoRA}}}{r} B \left(\beta_{\mathrm{SAGE}} A R\right). \tag{39}$$

Thus, sanitizing the LoRA-induced update is exactly equivalent to replacing its input-side factor with $\widetilde{A} = \beta_{\mathrm{SAGE}} A R$, while keeping B unchanged. This requires neither materializing the dense update nor increasing the LoRA rank, parameter count, or storage.

FIUBench experiment. We use Gradient Difference (GD) as the source unlearning method. The language backbone is updated through LoRA, the multimodal projector is densely updated, and the vision tower is frozen. SAGE sanitizes the LoRA updates as above and applies its standard dense operator to the projector updates. A disjoint 60-example forget-calibration set is used only to select β_{SAGE}; all reported forgetting metrics are evaluated on held-out examples.

Results. With closely matched calibration probabilities, SAGE improves retention: retain probability increases from 0.4393 to 0.6657, retain ROUGE-L from 0.0471 to 0.2125, and retain GT loss decreases from 23.6361 to 11.1772. Held-out forget probability and ROUGE-L remain low (0.0144 and 0.0190, respectively), although slightly higher than those of GD. APE, Exact Match, Min-K, and Min-K++ remain below 0.01. These results demonstrate that SAGE can be applied in this LoRA-based multimodal setting to improve retention while maintaining low held-out forget scores, without increasing the adapter rank.

Table 11: Full TOFU results on **Llama-3.2-1B-Instruct**. We report all metrics for the three settings: Forget01/Retain99, Forget05/Retain95, and Forget10/Retain90. Priv. Leak reports raw values.

Method	ES Re. ↑	ES Un. ↓	Priv. Leak	MU ↑	F-QA Prob. ↓	F-QA Rg. ↓	FQ ↑	F-TR ↑	MIA-L	MIA-MK	MIA-MK++	MIA-Z
Forget01 / Retain99												
Vanilla	0.055	0.058	-7.79	0.281	0.132	0.405	0.266	0.768	0.496	0.512	0.638	0.543
Fully Fine-tuned	0.736	0.743	-100.00	0.598	0.902	0.862	0.014	0.473	1.000	1.000	0.999	1.000
GradAscent	0.333	0.110	**10.63**	0.560	**0.116**	**0.369**	**0.054**	**0.572**	0.441	0.414	0.370	0.447
GradAscent w. SAGE	**0.592**↑ 0.259	**0.104**	-15.70	**0.591**	0.198	0.384	0.029	0.557	0.572	0.554	0.423	0.572
GradDiff	0.277	0.104	-31.29	0.519	0.141	0.392	0.029	0.497	0.644	0.636	0.522	0.672
GradDiff w. SAGE	**0.408**↑ 0.131	**0.104**	**-9.21**	**0.553**	**0.129**	**0.376**	**0.165**	**0.540**	0.538	0.519	0.408	0.554
NPO	0.583	0.150	-76.62	0.588	0.388	**0.417**	**0.003**	**0.542**	0.879	0.876	0.705	0.868
NPO w. SAGE	**0.630**↑ 0.047	**0.144**	**-75.09**	**0.593**	**0.376**	0.418	**0.003**	0.541	0.872	0.868	0.679	0.854
SimNPO	0.672	0.415	**-98.11**	0.593	**0.798**	**0.655**	**0.007**	0.473	0.989	0.990	0.961	0.999
SimNPO w. SAGE	**0.706**↑ 0.034	**0.402**	-98.58	**0.596**	0.802	0.656	**0.007**	**0.475**	0.991	0.993	0.965	0.998
RMU	0.216	0.032	**46.28**	0.526	0.055	0.235	**0.405**	**0.711**	0.203	0.226	0.571	0.153
RMU w. SAGE	**0.691**↑ 0.475	**0.031**	84.53	**0.593**	**0.010**	**0.224**	0.054	0.666	0.019	0.023	0.453	0.013
UNDIAL	0.484	0.091	-86.42	0.582	0.337	**0.354**	**0.007**	**0.551**	0.870	0.928	0.961	0.800
UNDIAL w. SAGE	**0.664**↑ 0.180	**0.080**	**-75.80**	**0.597**	**0.304**	0.357	**0.007**	0.546	0.820	0.872	0.912	0.741
CEU	0.250	0.039	**60.27**	0.527	0.062	0.241	0.266	**0.673**	0.089	0.152	0.321	0.191
CEU w. SAGE	**0.537**↑ 0.287	**0.038**	74.85	**0.584**	**0.039**	**0.224**	**0.579**	0.659	0.051	0.074	0.223	0.110
SatImp	0.677	0.206	-95.99	0.596	**0.639**	0.511	**0.014**	0.473	0.975	0.979	0.969	0.992
SatImp w. SAGE	**0.700**↑ 0.023	**0.198**	**-95.63**	**0.600**	0.651	**0.489**	**0.014**	**0.479**	0.974	0.977	0.969	0.993
WGA	0.683	0.141	**-86.78**	0.598	**0.437**	**0.427**	0.014	0.501	0.931	0.930	0.791	0.931
WGA w. SAGE	**0.721**↑ 0.038	**0.124**	-88.90	**0.601**	0.467	0.430	**0.029**	**0.508**	0.938	0.941	0.820	0.941
DPO	0.575	0.369	-98.94	0.575	0.731	0.480	**0.007**	0.526	0.991	0.994	0.992	0.999
DPO w. SAGE	**0.631**↑ 0.056	**0.327**	**-97.87**	**0.583**	**0.694**	**0.299**	0.003	**0.527**	0.982	0.989	0.981	0.997
PDU	0.043	**0.029**	**77.10**	0.207	9.69e-4	**0.047**	**0.097**	**0.707**	0.036	0.063	0.154	0.070
PDU w. SAGE	**0.211**↑ 0.168	**0.029**	88.78	**0.533**	**3.08e-4**	0.098	2.16e-5	0.603	6.25e-4	6.25e-4	0.127	0.000
Forget05 / Retain95												
Vanilla	0.055	0.059	8.09	0.281	0.097	0.365	0.001	0.740	0.300	0.311	0.453	0.264
Fully Fine-tuned	0.736	0.730	-99.99	0.598	0.886	0.822	6.54e-13	0.473	1.000	1.000	1.000	1.000
GradAscent	**0.035**	**0.033**	11.06	0.000	**6.45e-33**	**0.000**	1.94e-119	5.65e-23	0.470	0.292	0.288	0.377
GradAscent w. SAGE	**0.035**	**0.033**	**8.31**	**0.002**	3.26e-6	0.053	**1.78e-98**	**0.024**	0.274	0.309	0.386	0.253
GradDiff	0.035	**0.033**	**-1.45**	0.000	**1.19e-33**	**0.000**	1.94e-119	1.09e-23	0.339	0.372	0.372	0.373
GradDiff w. SAGE	**0.038**↑ 0.003	**0.033**	45.81	**1.75e-5**	1.71e-12	0.174	**7.77e-117**	**0.001**	0.043	0.070	0.082	0.104
NPO	0.131	0.071	**12.78**	0.437	0.123	**0.243**	0.178	**0.703**	0.283	0.281	0.316	0.287
NPO w. SAGE	**0.180**↑ 0.049	**0.065**	26.70	**0.495**	**0.092**	0.251	**0.271**	0.694	0.196	0.192	0.245	0.226
SimNPO	0.580	0.211	-96.77	0.578	0.553	0.505	1.21e-10	0.511	0.979	0.979	0.918	0.954
SimNPO w. SAGE	**0.656**↑ 0.076	**0.201**	**-95.94**	**0.590**	**0.546**	**0.501**	**6.87e-9**	**0.517**	0.974	0.974	0.895	0.941
RMU	0.663	**0.033**	**49.24**	0.585	**1.22e-5**	**0.023**	1.29e-8	**0.774**	0.046	0.049	0.233	0.046
RMU w. SAGE	**0.707**↑ 0.043	**0.033**	51.12	**0.592**	0.008	0.143	**7.54e-5**	0.683	0.025	0.037	0.438	0.014
UNDIAL	0.232	0.053	-91.46	0.556	0.269	**0.334**	1.21e-10	0.569	0.831	0.946	0.898	0.697
UNDIAL w. SAGE	**0.660**↑ 0.428	**0.052**	**-40.60**	**0.596**	**0.160**	0.338	**4.02e-6**	**0.590**	0.472	0.621	0.336	0.403
CEU	**0.035**	**0.033**	-23.67	0.000	**3.07e-4**	**0.024**	9.59e-10	**0.840**	0.474	0.513	0.513	0.385
CEU w. SAGE	**0.035**	**0.033**	**14.11**	**0.015**	0.002	0.038	**4.31e-4**	0.749	0.213	0.272	0.354	0.321
SatImp	0.413	0.069	-76.26	0.572	0.302	**0.355**	8.06e-7	**0.595**	0.803	0.849	0.852	0.746
SatImp w. SAGE	**0.580**↑ 0.167	**0.065**	**-67.41**	**0.584**	**0.263**	0.374	**2.96e-5**	0.588	0.752	0.792	0.794	0.699
WGA	0.555	0.034	**53.62**	**0.592**	0.007	**0.160**	**9.67e-4**	**0.629**	0.016	0.021	0.279	0.009
WGA w. SAGE	**0.663**↑ 0.107	**0.034**	54.40	0.591	**0.005**	0.189	6.49e-4	0.603	0.013	0.016	0.287	0.007
DPO	0.235	0.150	-84.12	**0.017**	0.387	0.002	1.83e-5	**0.625**	0.885	0.899	0.923	0.819
DPO w. SAGE	**0.349**↑ 0.114	**0.140**	**-71.31**	0.013	**0.328**	**0.001**	**2.96e-5**	0.610	0.811	0.817	0.857	0.724
PDU	0.065	0.043	**31.35**	0.307	0.035	**0.151**	2.96e-5	**0.758**	0.134	0.163	0.501	0.163
PDU w. SAGE	**0.139**↑ 0.074	**0.043**	40.23	**0.492**	**0.029**	0.245	**0.009**	0.642	0.089	0.106	0.532	0.114
Forget10 / Retain90												
Vanilla	0.055	0.055	10.43	0.281	0.092	0.368	9.08e-8	0.742	0.320	0.318	0.443	0.267
Fully Fine-tuned	0.736	0.712	-99.46	0.598	0.881	0.819	1.88e-22	0.475	0.997	0.997	0.998	0.998
GradAscent	0.035	**0.033**	-12.87	**0.000**	**4.45e-35**	**0.000**	**1.06e-239**	2.15e-25	0.532	0.462	0.462	0.368
GradAscent w. SAGE	**0.035**↑ 0.001	**0.033**	**-6.09**	**0.000**	1.73e-14	**0.000**	**1.06e-239**	**3.92e-10**	0.387	0.420	0.468	0.354
GradDiff	0.035	**0.033**	**-7.48**	**0.000**	**2.69e-31**	**0.000**	**1.06e-239**	1.63e-21	0.483	0.429	0.428	0.361
GradDiff w. SAGE	**0.035**	**0.033**	18.66	**0.000**	8.82e-16	0.034	**1.06e-239**	**1.15e-6**	0.218	0.267	0.505	0.254
NPO	0.205	0.075	**1.99**	0.528	0.115	**0.251**	0.054	**0.642**	0.380	0.370	0.376	0.352
NPO w. SAGE	**0.267**↑ 0.062	**0.072**	14.05	**0.550**	**0.075**	0.255	**0.322**	0.635	0.303	0.296	0.349	0.302
SimNPO	0.637	0.155	-95.12	0.586	0.488	0.463	1.84e-14	0.522	0.969	0.970	0.896	0.942
SimNPO w. SAGE	**0.647**↑ 0.010	**0.148**	**-92.48**	**0.593**	**0.453**	**0.458**	**1.02e-13**	**0.527**	0.954	0.954	0.848	0.907
RMU	0.707	**0.033**	58.81	0.592	**5.42e-6**	**0.018**	7.83e-12	**0.756**	0.019	0.020	0.149	0.021
RMU w. SAGE	**0.709**↑ 0.001	0.033	**55.37**	**0.595**	0.014	0.152	**4.36e-9**	0.708	0.030	0.041	0.428	0.032
UNDIAL	0.268	0.044	-93.50	0.563	0.253	0.328	6.16e-18	0.562	0.878	0.960	0.923	0.725
UNDIAL w. SAGE	**0.625**↑ 0.357	**0.044**	**-49.35**	**0.597**	**0.148**	**0.318**	**6.79e-7**	**0.594**	0.521	0.687	0.364	0.421
CEU	**0.035**	**0.033**	-24.71	**0.000**	**1.64e-4**	**0.005**	9.46e-24	**0.861**	0.514	0.535	0.537	0.375
CEU w. SAGE	**0.035**	**0.033**	**11.44**	**0.000**	0.003	0.042	**2.13e-6**	0.735	0.310	0.312	0.492	0.332
SatImp	0.358	0.059	-72.00	0.552	0.264	0.329	3.91e-8	0.616	0.766	0.827	0.887	0.672
SatImp w. SAGE	**0.453**↑ 0.095	**0.057**	**-58.35**	**0.569**	**0.201**	**0.322**	**5.00e-5**	**0.624**	0.676	0.743	0.838	0.580
WGA	0.627	**0.033**	60.71	0.590	**0.001**	**0.022**	3.40e-21	**0.782**	0.004	0.008	0.671	5.56e-4
WGA w. SAGE	**0.734**↑ 0.108	**0.033**	**58.76**	**0.600**	0.004	0.074	**1.75e-9**	0.710	0.015	0.020	0.852	0.004
DPO	0.316	0.192	-93.95	**0.137**	0.498	0.012	2.20e-11	**0.595**	0.958	0.963	0.970	0.911
DPO w. SAGE	**0.451**↑ 0.136	**0.191**	**-91.13**	0.125	**0.467**	**0.005**	**9.91e-11**	0.576	0.945	0.945	0.941	0.879
PDU	0.137	0.057	**20.36**	0.469	0.060	**0.234**	2.57e-5	**0.728**	0.139	0.257	0.971	0.128
PDU w. SAGE	**0.206**↑ 0.069	**0.052**	27.00	**0.478**	**0.051**	0.242	**0.211**	0.633	0.188	0.216	0.882	0.190

Table 12: Full TOFU results on **Llama-3.2-3B-Instruct**. We report all metrics for the three settings: Forget01/Retain99, Forget05/Retain95, and Forget10/Retain90. Priv. Leak reports raw values.

Method	ES Re. ↑	ES Un. ↓	Priv. Leak	MU ↑	F-QA Prob. ↓	F-QA Rg. ↓	FQ ↑	F-TR ↑	MIA-L	MIA-MK	MIA-MK++	MIA-Z
Forget01 / Retain99												
Vanilla	0.063	0.055	18.64	0.272	0.104	0.325	0.097	0.756	0.482	0.475	0.469	0.493
Fully Fine-tuned	0.885	0.920	-100.00	0.665	0.969	0.986	0.014	0.507	1.000	1.000	1.000	1.000
GradAscent	0.848	0.326	-90.25	**0.667**	**0.615**	**0.573**	0.014	0.520	0.954	0.957	0.794	0.956
GradAscent w. SAGE	**0.862**↑ 0.014	**0.320**	**-89.69**	0.666	0.618	0.586	**0.029**	**0.521**	0.953	0.954	0.797	0.953
GradDiff	0.812	0.250	-79.38	0.658	0.453	0.530	0.014	0.524	0.905	0.909	0.628	0.903
GradDiff w. SAGE	**0.853**↑ 0.041	**0.231**	**-76.13**	**0.662**	**0.447**	**0.526**	**0.029**	**0.531**	0.889	0.894	0.616	0.891
NPO	0.760	0.201	-80.93	0.667	0.434	0.443	0.029	0.544	0.914	0.916	0.729	0.909
NPO w. SAGE	**0.821**↑ 0.061	**0.188**	**-77.68**	**0.669**	**0.401**	**0.429**	**0.097**	**0.554**	0.897	0.901	0.688	0.886
SimNPO	0.839	**0.562**	**-99.86**	0.652	**0.886**	**0.815**	**0.029**	**0.510**	0.999	0.999	0.974	1.000
SimNPO w. SAGE	**0.856**↑ 0.017	**0.562**	**-99.86**	**0.653**	0.890	0.830	**0.029**	0.510	0.999	0.999	0.974	1.000
RMU	0.320	0.038	**25.00**	0.612	0.139	0.344	**0.990**	0.655	0.422	0.447	0.639	0.476
RMU w. SAGE	**0.817**↑ 0.498	**0.037**	70.90	**0.665**	**0.066**	**0.294**	0.266	**0.738**	0.206	0.244	0.610	0.189
UNDIAL	0.750	0.084	-91.24	0.675	0.442	0.405	**0.014**	0.542	0.933	0.961	0.914	0.941
UNDIAL w. SAGE	**0.831**↑ 0.081	**0.076**	**-84.75**	**0.679**	**0.374**	**0.398**	**0.014**	**0.551**	0.885	0.932	0.786	0.873
CEU	0.646	0.080	-37.85	0.672	0.198	0.347	0.029	0.575	0.584	0.725	0.578	0.572
CEU w. SAGE	**0.800**↑ 0.154	**0.077**	**-5.37**	**0.683**	**0.151**	**0.326**	**0.165**	**0.594**	0.427	0.581	0.444	0.482
SatImp	0.847	0.348	**-96.05**	0.659	**0.661**	0.578	**0.029**	**0.512**	0.986	0.983	0.919	0.993
SatImp w. SAGE	**0.861**↑ 0.014	**0.322**	-96.19	**0.661**	0.667	**0.572**	0.014	0.510	0.981	0.983	0.920	0.993
WGA	0.841	0.206	-88.28	0.663	0.518	0.506	**0.029**	0.524	0.945	0.948	0.800	0.951
WGA w. SAGE	**0.864**↑ 0.023	**0.200**	**-86.58**	**0.665**	**0.507**	**0.488**	**0.029**	**0.528**	0.937	0.941	0.794	0.947
DPO	0.760	0.661	**-100.00**	**0.659**	0.858	0.697	**0.007**	**0.535**	1.000	1.000	0.996	1.000
DPO w. SAGE	**0.788**↑ 0.028	**0.652**	**-100.00**	0.656	**0.833**	**0.625**	**0.007**	0.531	0.999	1.000	0.996	1.000
PDU	0.337	0.056	**41.95**	0.653	**0.054**	**0.208**	0.579	**0.688**	0.179	0.372	0.995	0.232
PDU w. SAGE	**0.654**↑ 0.317	**0.055**	63.14	**0.708**	0.062	0.261	**0.990**	0.671	0.143	0.278	0.984	0.215
Forget05 / Retain95												
Vanilla	0.063	0.055	11.51	0.272	0.084	0.299	0.001	0.732	0.270	0.287	0.447	0.282
Fully Fine-tuned	0.885	0.887	-100.00	0.665	0.952	0.929	0.00	0.462	1.000	1.000	1.000	1.000
GradAscent	0.035	**0.033**	**-23.74**	0.000	**0.00**	**0.000**	0.00	0.00	0.442	0.512	0.132	0.372
GradAscent w. SAGE	**0.085**↑ 0.050	**0.033**	31.98	**0.292**	0.00	0.014	**0.00**	**0.079**	0.134	0.156	0.131	0.092
GradDiff	0.212	**0.033**	56.34	0.596	**0.00**	**0.000**	0.00	**0.003**	0.000	0.000	0.000	0.000
GradDiff w. SAGE	**0.744**↑ 0.531	**0.033**	**56.08**	**0.652**	0.00	0.014	**0.00**	0.002	0.002	0.002	0.004	0.00
NPO	0.140	0.060	**20.52**	0.466	0.107	**0.174**	0.009	**0.713**	0.212	0.229	0.331	0.252
NPO w. SAGE	**0.542**↑ 0.403	**0.057**	30.72	**0.648**	**0.076**	0.238	**0.112**	0.688	0.140	0.164	0.389	0.176
SimNPO	0.624	0.264	-97.99	0.640	0.631	0.551	**0.00**	**0.498**	0.986	0.987	0.883	0.984
SimNPO w. SAGE	**0.712**↑ 0.088	**0.256**	**-96.12**	**0.652**	**0.574**	**0.537**	**0.00**	0.494	0.973	0.975	0.815	0.964
RMU	0.830	**0.033**	54.30	**0.669**	**0.003**	**0.089**	0.00	**0.764**	0.011	0.013	0.636	0.012
RMU w. SAGE	**0.860**↑ 0.030	**0.033**	**49.85**	0.668	0.012	0.116	**0.00**	0.748	0.037	0.042	0.719	0.035
UNDIAL	0.304	0.051	-84.27	0.641	0.258	**0.324**	0.00	**0.541**	0.744	0.899	0.785	0.618
UNDIAL w. SAGE	**0.805**↑ 0.501	**0.048**	**-64.28**	**0.680**	**0.235**	0.337	**0.00**	0.539	0.636	0.771	0.563	0.541
CEU	0.035	**0.033**	**-7.43**	0.000	**0.00**	**0.000**	0.00	**0.854**	0.352	0.408	0.409	0.357
CEU w. SAGE	**0.047**↑ 0.012	**0.033**	51.74	**0.202**	0.00	0.024	**0.00**	0.714	0.028	0.029	0.062	0.135
SatImp	0.536	0.091	-55.16	0.614	0.301	0.345	0.00	**0.615**	0.642	0.713	0.940	0.594
SatImp w. SAGE	**0.714**↑ 0.177	**0.087**	**-53.12**	**0.648**	**0.276**	**0.332**	**0.002**	0.607	0.632	0.700	0.913	0.589
WGA	0.623	**0.033**	**53.24**	0.640	0.006	**0.031**	0.00	**0.803**	0.018	0.020	0.846	0.007
WGA w. SAGE	**0.730**↑ 0.107	**0.033**	53.65	**0.649**	**0.006**	0.039	**0.00**	0.791	0.016	0.017	0.912	0.007
DPO	0.307	0.203	-90.25	0.100	0.475	**0.012**	0.00	0.603	0.921	0.938	0.956	0.871
DPO w. SAGE	**0.739**↑ 0.432	**0.196**	**-73.28**	**0.572**	**0.370**	0.027	**0.003**	**0.607**	0.811	0.829	0.912	0.742
PDU	0.035	**0.033**	-21.11	0.000	0.00	**0.019**	**0.00**	0.830	0.450	0.495	0.543	0.366
PDU w. SAGE	**0.132**↑ 0.097	**0.033**	**19.69**	**0.391**	**0.00**	0.025	0.00	**0.834**	0.134	0.234	0.328	0.216
Forget10 / Retain90												
Vanilla	0.063	0.053	13.76	0.272	0.082	0.288	0.00	0.736	0.302	0.311	0.470	0.289
Fully Fine-tuned	0.885	0.888	-99.72	0.665	0.951	0.925	0.00	0.473	0.998	0.998	0.998	0.999
GradAscent	0.035	**0.033**	**-21.73**	**0.000**	**0.00**	**0.000**	0.00	0.00	0.473	0.526	0.439	0.360
GradAscent w. SAGE	**0.035**↑ 0.001	**0.033**	31.84	**0.000**	0.00	0.00	**0.00**	**0.041**	0.187	0.201	0.213	0.130
GradDiff	0.189	**0.033**	64.79	0.585	**0.00**	**0.002**	0.00	**0.00**	0.00	0.001	0.001	0.020
GradDiff w. SAGE	**0.714**↑ 0.525	**0.033**	**64.55**	**0.655**	0.00	0.003	**0.00**	0.00	0.002	0.003	0.009	0.015
NPO	0.132	**0.060**	**16.62**	0.529	0.130	**0.248**	0.054	**0.686**	0.275	0.293	0.347	0.287
NPO w. SAGE	**0.521**↑ 0.389	**0.060**	21.52	**0.648**	**0.107**	0.271	**0.281**	0.660	0.244	0.264	0.343	0.258
SimNPO	0.583	0.212	-96.76	**0.646**	0.570	0.499	**0.00**	0.523	0.980	0.980	0.878	0.956
SimNPO w. SAGE	**0.625**↑ 0.042	**0.206**	**-93.99**	0.644	**0.486**	**0.483**	0.00	**0.525**	0.964	0.964	0.775	0.918
RMU	**0.859**	**0.033**	62.83	**0.675**	0.004	**0.077**	**0.00**	0.778	0.009	0.013	0.788	0.018
RMU w. SAGE	0.854	**0.033**	**61.06**	0.665	**0.003**	0.080	0.00	**0.779**	0.019	0.024	0.761	0.023
UNDIAL	0.357	**0.041**	-89.69	0.655	0.254	0.328	0.00	0.539	0.828	0.938	0.838	0.680
UNDIAL w. SAGE	**0.750**↑ 0.393	**0.041**	**-34.68**	**0.678**	**0.139**	**0.312**	**0.00**	**0.574**	0.433	0.604	0.318	0.330
CEU	0.035	**0.033**	**-32.27**	0.000	**0.00**	**0.000**	0.00	**0.867**	0.549	0.590	0.582	0.369
CEU w. SAGE	**0.048**↑ 0.013	**0.033**	39.38	**0.206**	0.003	0.044	**0.00**	0.732	0.141	0.155	0.365	0.224
SatImp	0.462	0.048	38.29	0.614	0.036	0.194	**0.00**	0.755	0.068	0.162	0.994	0.057
SatImp w. SAGE	**0.598**↑ 0.136	**0.048**	**36.91**	**0.640**	**0.027**	**0.161**	0.00	**0.760**	0.066	0.170	0.983	0.054
WGA	0.633	0.035	**58.91**	0.648	**0.005**	0.081	0.00	**0.809**	0.014	0.037	0.992	0.003
WGA w. SAGE	**0.744**↑ 0.111	**0.035**	59.11	**0.652**	0.005	**0.075**	**0.00**	0.802	0.017	0.036	0.975	0.005
DPO	0.378	0.264	-96.07	0.327	0.574	**0.037**	0.00	**0.583**	0.972	0.976	0.977	0.936
DPO w. SAGE	**0.774**↑ 0.396	**0.255**	**-92.19**	**0.616**	**0.525**	0.101	**0.00**	0.570	0.950	0.953	0.952	0.895
PDU	0.106	0.073	-55.90	0.493	0.224	0.296	0.006	**0.705**	0.642	0.733	0.888	0.517
PDU w. SAGE	**0.222**↑ 0.116	**0.052**	**5.32**	**0.542**	**0.018**	**0.152**	**0.111**	0.679	0.273	0.362	0.818	0.205

Table 13: Full TOFU results on **Llama-3.1-8B-Instruct**. We report all metrics for the three settings: Forget01/Retain99, Forget05/Retain95, and Forget10/Retain90. Priv. Leak reports raw values.

Method	ES Re. ↑	ES Un. ↓	Priv. Leak	MU ↑	F-QA Prob. ↓	F-QA Rg. ↓	FQ ↑	F-TR ↑	MIA-L	MIA-MK	MIA-MK++	MIA-Z
Forget01 / Retain99												
Vanilla	0.062	0.062	3.38	0.275	0.114	0.292	0.405	0.751	0.475	0.483	0.591	0.482
Fully Fine-tuned	0.992	0.977	-100.00	0.627	0.995	0.998	0.014	0.485	1.000	1.000	1.000	1.000
GradAscent	0.874	0.153	-29.50	0.636	0.157	0.477	0.266	0.554	0.650	0.648	0.161	0.620
GradAscent w. SAGE	**0.944**↑ 0.070	**0.151**	**-7.00**	**0.639**	**0.139**	**0.462**	**0.405**	**0.580**	0.539	0.535	0.121	0.522
GradDiff	0.815	0.204	-56.62	0.631	**0.214**	**0.486**	**0.405**	0.564	0.782	0.783	0.291	0.751
GradDiff w. SAGE	**0.934**↑ 0.119	**0.200**	**-50.37**	**0.633**	0.245	0.535	**0.405**	**0.587**	0.753	0.752	0.237	0.724
NPO	0.866	0.152	-68.25	**0.642**	**0.220**	**0.480**	**0.266**	0.569	0.839	0.841	0.222	0.808
NPO w. SAGE	**0.956**↑ 0.090	**0.147**	**-64.37**	0.639	0.229	0.497	**0.266**	**0.576**	0.824	0.822	0.198	0.790
SimNPO	0.939	**0.566**	**-97.50**	0.615	**0.721**	**0.799**	**0.014**	0.494	0.987	0.988	0.783	0.973
SimNPO w. SAGE	**0.967**↑ 0.029	**0.566**	-98.37	**0.619**	0.779	0.855	0.007	**0.495**	0.992	0.992	0.802	0.986
RMU	0.833	0.074	-55.37	**0.635**	0.255	0.419	0.579	0.606	0.746	0.777	0.761	0.789
RMU w. SAGE	**0.986**↑ 0.153	**0.061**	**3.13**	0.630	**0.138**	**0.318**	**0.990**	**0.678**	0.429	0.484	0.778	0.441
UNDIAL	0.819	0.112	-85.75	**0.688**	0.448	**0.420**	**0.029**	**0.547**	0.900	0.929	0.871	0.913
UNDIAL w. SAGE	**0.976**↑ 0.157	**0.109**	**-84.87**	0.669	**0.444**	0.446	0.014	0.527	0.908	0.924	0.836	0.934
CEU	0.587	0.040	**58.75**	0.659	0.083	0.361	**0.766**	**0.660**	0.108	0.206	0.316	0.213
CEU w. SAGE	**0.906**↑ 0.320	**0.035**	73.25	**0.665**	**0.049**	**0.339**	**0.766**	0.630	0.073	0.134	0.244	0.146
SatImp	0.950	**0.461**	**-94.62**	0.618	**0.627**	**0.745**	**0.054**	**0.504**	0.947	0.973	0.753	0.973
SatImp w. SAGE	**0.972**↑ 0.022	**0.461**	-96.25	**0.622**	0.720	0.800	0.014	0.497	0.981	0.981	0.818	0.992
WGA	0.946	**0.124**	-63.75	**0.639**	**0.254**	0.475	0.029	0.541	0.824	0.819	0.386	0.804
WGA w. SAGE	**0.962**↑ 0.016	**0.124**	**-61.62**	0.634	0.265	**0.467**	**0.054**	**0.545**	0.807	0.808	0.393	0.803
DPO	0.936	0.819	-99.87	**0.680**	0.894	0.664	**0.029**	**0.538**	0.999	0.999	0.998	1.000
DPO w. SAGE	**0.986**↑ 0.050	**0.816**	**-99.75**	0.648	**0.872**	**0.540**	**0.029**	0.526	0.998	0.999	0.996	1.000
PDU	0.291	**0.030**	**92.00**	0.627	0.009	**0.123**	**0.165**	**0.729**	0.017	0.040	0.854	0.044
PDU w. SAGE	**0.916**↑ 0.625	**0.030**	99.75	**0.720**	**0.002**	0.174	0.00	0.689	0.00	0.001	0.808	0.000
Forget05 / Retain95												
Vanilla	0.062	0.060	11.67	0.275	0.095	0.295	0.004	0.721	0.269	0.281	0.352	0.269
Fully Fine-tuned	0.992	0.972	-100.00	0.627	0.990	0.989	0.00	0.479	1.000	1.000	1.000	1.000
GradAscent	0.035	**0.033**	**-33.69**	0.000	**0.000**	**0.000**	0.00	0.00	0.495	0.573	0.573	0.382
GradAscent w. SAGE	**0.165**↑ 0.130	**0.033**	34.71	**0.406**	0.00	0.025	**0.00**	**0.00**	0.125	0.133	0.134	0.139
GradDiff	0.297	**0.033**	**55.25**	**0.637**	**0.00**	**0.001**	0.00	0.00	0.000	0.00	0.00	0.005
GradDiff w. SAGE	**0.904**↑ 0.607	**0.033**	55.33	0.611	0.00	0.005	**0.00**	**0.003**	0.000	0.000	0.000	0.000
NPO	0.176	0.064	**39.71**	0.556	0.070	**0.271**	**0.022**	**0.689**	0.102	0.101	0.132	0.154
NPO w. SAGE	**0.601**↑ 0.424	**0.064**	41.15	**0.609**	**0.031**	0.285	0.012	0.640	0.094	0.091	0.132	0.127
SimNPO	0.699	0.303	-97.66	**0.619**	0.587	**0.568**	0.00	0.516	0.984	0.985	0.734	0.962
SimNPO w. SAGE	**0.833**↑ 0.134	**0.276**	**-94.10**	0.614	**0.477**	0.577	**0.00**	**0.522**	0.963	0.962	0.584	0.930
RMU	**0.985**	0.039	**50.85**	**0.657**	0.017	0.071	0.00	**0.739**	0.029	0.029	0.900	0.019
RMU w. SAGE	0.981	**0.037**	51.22	0.626	**0.013**	**0.070**	**0.00**	0.732	0.026	0.026	0.890	0.017
UNDIAL	0.508	0.051	-79.38	**0.690**	0.313	0.394	0.00	**0.539**	0.768	0.867	0.612	0.675
UNDIAL w. SAGE	**0.955**↑ 0.447	**0.051**	**-48.78**	0.651	**0.232**	**0.387**	**0.00**	0.534	0.595	0.670	0.464	0.533
CEU	0.035	**0.033**	**1.70**	0.000	**0.00**	**0.000**	0.00	**0.817**	0.312	0.345	0.376	0.341
CEU w. SAGE	**0.361**↑ 0.327	**0.033**	50.87	**0.611**	0.005	0.096	**0.00**	0.684	0.023	0.029	0.363	0.067
SatImp	0.690	0.117	-84.81	**0.627**	0.413	**0.407**	0.00	0.558	0.895	0.902	0.787	0.845
SatImp w. SAGE	**0.852**↑ 0.162	**0.109**	**-77.57**	0.611	**0.340**	0.420	**0.00**	**0.559**	0.853	0.856	0.696	0.799
WGA	0.614	**0.033**	54.85	0.599	0.00	0.059	**0.00**	**0.832**	0.000	0.003	0.942	0.000
WGA w. SAGE	**0.792**↑ 0.178	**0.033**	**54.20**	**0.603**	**0.00**	**0.046**	0.00	0.809	0.00	0.007	0.927	0.00
DPO	0.467	0.238	-91.41	0.156	0.497	0.022	0.00	0.602	0.938	0.945	0.942	0.879
DPO w. SAGE	**0.897**↑ 0.429	**0.232**	**-74.94**	**0.509**	**0.343**	**0.017**	**0.00**	**0.605**	0.841	0.839	0.803	0.748
PDU	0.154	0.060	**37.08**	0.496	0.057	0.200	**0.00**	**0.752**	0.079	0.118	0.752	0.118
PDU w. SAGE	**0.636**↑ 0.482	**0.037**	42.07	**0.686**	**0.009**	**0.149**	0.00	0.720	0.044	0.085	0.565	0.080
Forget10 / Retain90												
Vanilla	0.062	0.056	11.48	0.275	0.093	0.286	0.002	0.721	0.308	0.309	0.408	0.283
Fully Fine-tuned	0.992	0.979	-99.94	0.627	0.992	0.992	0.00	0.475	1.000	1.000	0.999	1.000
GradAscent	0.035	**0.033**	**-14.19**	0.000	**0.000**	**0.000**	0.00	0.00	0.462	0.468	0.468	0.366
GradAscent w. SAGE	**0.045**↑ 0.010	**0.033**	-14.24	**0.050**	0.00	0.033	**0.00**	**0.017**	0.445	0.468	0.469	0.373
GradDiff	0.222	**0.033**	61.33	0.548	0.00	0.030	**0.00**	0.00	0.000	0.000	0.000	0.000
GradDiff w. SAGE	**0.872**↑ 0.650	**0.033**	**61.31**	**0.596**	**0.00**	**0.021**	**0.00**	**0.00**	0.00	0.00	0.00	0.000
NPO	0.233	0.071	40.45	0.601	0.074	**0.272**	0.054	**0.633**	0.131	0.129	0.100	0.174
NPO w. SAGE	**0.543**↑ 0.310	**0.067**	**39.30**	**0.613**	**0.045**	0.294	**0.700**	0.621	0.138	0.137	0.119	0.177
SimNPO	0.568	0.230	-97.00	0.602	0.598	**0.495**	**0.00**	**0.530**	0.982	0.981	0.852	0.962
SimNPO w. SAGE	**0.680**↑ 0.112	**0.223**	**-94.30**	**0.604**	**0.489**	0.512	0.00	0.524	0.966	0.965	0.668	0.927
RMU	**0.986**	**0.033**	**59.52**	**0.652**	0.003	0.047	**0.00**	0.724	0.011	0.011	0.921	0.003
RMU w. SAGE	0.979	**0.033**	60.45	0.626	**0.003**	**0.041**	0.00	**0.733**	0.005	0.005	0.888	0.005
UNDIAL	0.631	0.051	-92.15	**0.689**	0.338	0.378	0.00	0.525	0.891	0.951	0.820	0.796
UNDIAL w. SAGE	**0.932**↑ 0.301	**0.050**	**-55.31**	0.644	**0.213**	**0.377**	**0.00**	**0.544**	0.634	0.723	0.494	0.538
CEU	0.035	**0.033**	**-10.80**	0.000	**0.00**	**0.000**	0.00	**0.863**	0.407	0.447	0.371	0.357
CEU w. SAGE	**0.036**↑ 0.001	**0.033**	30.89	**0.130**	0.002	0.023	**0.00**	0.731	0.225	0.189	0.243	0.270
SatImp	0.578	0.046	**5.66**	**0.596**	**0.076**	0.187	0.00	**0.733**	0.151	0.345	0.841	0.129
SatImp w. SAGE	**0.736**↑ 0.157	**0.045**	-23.62	0.596	0.083	**0.185**	**0.010**	0.711	0.265	0.527	0.843	0.245
WGA	0.537	**0.033**	57.99	0.593	**0.004**	0.049	0.00	**0.838**	0.013	0.021	0.986	0.002
WGA w. SAGE	**0.809**↑ 0.272	**0.033**	**56.59**	**0.609**	0.004	**0.047**	**0.00**	0.828	0.019	0.029	0.986	0.005
DPO	0.656	0.370	-96.62	0.332	0.617	**0.077**	0.00	**0.572**	0.979	0.979	0.909	0.934
DPO w. SAGE	**0.951**↑ 0.295	**0.369**	**-89.24**	**0.575**	**0.510**	0.079	**0.00**	0.563	0.938	0.933	0.854	0.864
PDU	0.263	0.081	11.86	0.604	0.097	0.253	0.020	**0.691**	0.196	0.307	0.967	0.187
PDU w. SAGE	**0.834**↑ 0.572	**0.072**	**7.72**	**0.689**	**0.071**	**0.165**	**0.054**	0.683	0.248	0.332	0.872	0.227

Table 14: TOFU results with disjoint forget-side calibration on **Llama-3.1-8B-Instruct**. Calibration ES-Un. is shown separately from held-out evaluation ES-Un. ES-Re. and MU measure retention and model utility, respectively. Better evaluation scores within each source–SAGE pair are **bolded**. Orange superscripts indicate absolute ES-Re. gains.

Method	ES Re. ↑	Cal. ES Un. (20%)	Eval. ES Un. ↓ (80%)	|Priv. Leak| ↓	MU ↑
Forget01 / Retain99					
NPO	0.866	0.121	**0.160**	**68.25**	**0.642**
NPO w. SAGE	**0.964**↑ 0.098	0.121	0.178	71.12	0.638
UNDIAL	0.819	0.073	0.124	85.75	**0.688**
UNDIAL w. SAGE	**0.977**↑ 0.158	0.068	**0.111**	**81.75**	0.667
SatImp	0.949	0.499	**0.451**	**94.62**	0.618
SatImp w. SAGE	**0.972**↑ 0.023	0.499	**0.451**	96.25	**0.622**
WGA	0.946	0.123	**0.124**	**63.75**	**0.639**
WGA w. SAGE	**0.971**↑ 0.025	0.123	0.127	74.50	0.633
Forget05 / Retain95					
NPO	0.176	0.068	0.062	**39.71**	0.556
NPO w. SAGE	**0.577**↑ 0.401	0.065	**0.057**	42.63	**0.607**
UNDIAL	0.508	0.058	0.050	79.38	**0.690**
UNDIAL w. SAGE	**0.873**↑ 0.365	0.050	**0.039**	**20.11**	0.639
SatImp	0.690	0.155	**0.105**	84.81	**0.627**
SatImp w. SAGE	**0.887**↑ 0.197	0.138	0.136	**84.61**	0.615
WGA	0.613	0.036	**0.032**	54.85	0.598
WGA w. SAGE	**0.904**↑ 0.291	0.036	0.038	**36.39**	**0.616**
Forget10 / Retain90					
NPO	0.233	0.076	0.070	40.45	0.601
NPO w. SAGE	**0.542**↑ 0.309	0.075	**0.067**	**39.34**	**0.612**
UNDIAL	0.631	0.050	**0.051**	92.15	**0.689**
UNDIAL w. SAGE	**0.939**↑ 0.308	0.048	0.053	**66.87**	0.644
SatImp	0.578	0.045	0.046	**5.66**	**0.596**
SatImp w. SAGE	**0.708**↑ 0.130	0.043	**0.042**	12.03	0.592
WGA	0.537	0.033	**0.032**	57.99	0.593
WGA w. SAGE	**0.806**↑ 0.269	0.033	**0.032**	**56.61**	**0.609**

Table 15: Detailed MUSE-News results for **Llama-2-7b-hf**. We report forgetting metrics, retain metrics, privacy leakage, and four MIA-based scores.

Method	ES Un. ↓	ROUGE Re. ↑	Priv. Leak	Forget Know. R. ↓	Forget Verb. R. ↓	ES Re. ↑	MIA Loss	MIA Min-K	MIA Min-K++	MIA Zlib
GradAscent	0.008	0.000	9.89	0.000	0.000	0.265	0.476	0.476	0.475	0.558
GradAscent w. SAGE	**0.008**	**0.000**	14.79	**0.000**	**0.000**	**0.265**	0.437	0.453	0.504	0.523
GradDiff	0.071	0.487	-98.68	0.545	0.217	0.612	0.990	0.994	0.916	0.995
GradDiff w. SAGE	**0.070**	**0.496**↑ 0.009	-97.08	**0.583**	**0.230**	**0.617**↑ 0.005	0.983	0.986	0.803	0.990
SimNPO	0.214	0.486	-99.87	0.563	0.480	0.607	0.999	0.999	0.945	1.000
SimNPO w. SAGE	**0.212**	**0.532**↑ 0.047	-99.87	**0.567**	**0.518**	**0.611**↑ 0.003	0.999	0.999	0.963	1.000
RMU	0.021	0.482	25.59	0.429	0.124	0.614	0.391	0.402	0.472	0.397
RMU w. SAGE	**0.019**	**0.496**↑ 0.014	30.90	**0.463**	**0.136**	**0.618**↑ 0.004	0.373	0.376	0.437	0.378
UNDIAL	0.013	0.189	-99.45	0.123	0.150	0.392	0.979	0.997	0.974	0.991
UNDIAL w. SAGE	**0.013**	**0.264**↑ 0.075	-89.86	**0.283**	**0.177**	**0.433**↑ 0.041	0.855	0.952	0.475	0.883
CEU	0.008	0.000	-2.92	0.000	0.000	0.265	0.529	0.537	0.509	0.550
CEU w. SAGE	**0.008**	**0.022**↑ 0.022	3.30	**0.027**	**0.000**	**0.265**	0.482	0.508	0.587	0.512
SatImp	0.037	0.450	10.39	0.578	0.213	0.623	0.198	0.474	0.999	0.199
SatImp w. SAGE	**0.036**	**0.451**↑ 0.001	-59.66	**0.510**	**0.222**	**0.607**↓ 0.016	0.517	0.808	1.000	0.538
WGA	0.012	0.436	103.70	0.511	0.033	0.598	0.022	0.029	0.607	0.020
WGA w. SAGE	**0.011**	**0.468**↑ 0.032	102.60	**0.475**	**0.056**	**0.591**↓ 0.006	0.014	0.035	0.967	0.014
PDU	0.146	0.501	-99.79	0.554	0.408	0.611	0.998	0.999	0.998	1.000
PDU w. SAGE	**0.141**	**0.484**↓ 0.017	-99.81	**0.592**	**0.424**	**0.623**↑ 0.012	0.998	0.999	0.999	1.000

Table 16: Detailed MUSE-Books results for **Llama-2-7b-hf**. We report forgetting metrics, retain metrics, privacy leakage, and four MIA-based scores.

Method	ES Un. ↓	ROUGE Re. ↑	Priv. Leak	Forget Know. R. ↓	Forget Verb. R. ↓	ES Re. ↑	MIA Loss	MIA Min-K	MIA Min-K++	MIA Zlib
GradAscent	0.008	0.000	-26.08	0.000	0.000	0.286	0.541	0.600	0.679	0.787
GradAscent w. SAGE	**0.008**	**0.322**↑ 0.322	-41.09	**0.013**	**0.000**	**0.783**↑ 0.497	0.608	0.681	0.576	0.735
GradDiff	0.008	0.052	-28.62	0.000	0.000	0.673	0.567	0.614	0.488	0.803
GradDiff w. SAGE	**0.008**	**0.610**↑ 0.558	-32.54	**0.325**	**0.001**	**0.806**↑ 0.133	0.587	0.635	0.604	0.679
SimNPO	0.138	0.537	-54.81	0.378	0.264	0.833	0.755	0.756	0.310	0.780
SimNPO w. SAGE	**0.106**	**0.611**↑ 0.074	-54.66	**0.369**	**0.216**	**0.834**↑ 0.001	0.752	0.755	0.326	0.778
RMU	0.008	0.124	-19.91	0.019	0.022	0.666	0.556	0.567	0.563	0.651
RMU w. SAGE	**0.008**	**0.336**↑ 0.212	-38.76	**0.032**	**0.027**	**0.721**↑ 0.054	0.678	0.669	0.628	0.745
UNDIAL	0.024	0.627	-18.45	0.309	0.198	0.781	0.554	0.559	0.554	0.557
UNDIAL w. SAGE	**0.024**	**0.700**↑ 0.073	-18.05	**0.345**	**0.209**	**0.812**↑ 0.031	0.554	0.557	0.554	0.554
CEU	0.008	0.000	-45.41	0.000	0.001	0.286	0.604	0.705	0.205	0.806
CEU w. SAGE	**0.008**	**0.000**	-31.18	**0.000**	**0.000**	**0.286**	0.510	0.628	0.613	0.804
SatImp	0.008	0.611	-14.57	0.065	0.000	0.821	0.502	0.538	0.473	0.726
SatImp w. SAGE	**0.008**	**0.662**↑ 0.051	-41.27	**0.365**	**0.001**	**0.821**	0.627	0.682	0.404	0.731
WGA	0.008	0.486	-43.82	0.008	0.000	0.797	0.633	0.696	0.587	0.780
WGA w. SAGE	**0.008**	**0.619**↑ 0.133	-25.55	**0.272**	**0.000**	**0.826**↑ 0.028	0.583	0.597	0.566	0.763
PDU	0.008	0.042	-54.48	0.000	0.004	0.666	0.754	0.754	0.813	0.811
PDU w. SAGE	**0.008**	**0.605**↑ 0.563	-71.21	**0.137**	**0.001**	**0.837**↑ 0.170	0.748	0.844	0.656	0.805

Table 17: Full TOFU results on **Qwen-2.5-3B-Instruct**. We report all metrics for the three settings: Forget01/Retain99, Forget05/Retain95, and Forget10/Retain90. Priv. Leak reports raw values.

Method	ES Re. ↑	ES Un. ↓	Priv. Leak	MU ↑	F-QA Prob. ↓	F-QA Rg. ↓	FQ ↑	F-TR ↑	MIA-L	MIA-MK	MIA-MK++	MIA-Z
Forget01 / Retain99												
Vanilla	0.042	0.072	-0.616	0.264	0.111	0.420	0.405	0.698	0.498	0.496	0.551	0.513
Fully Fine-tuned	0.602	0.482	-99.75	0.550	0.851	0.774	0.007	0.470	0.999	0.999	1.000	1.000
GradAscent	0.324	0.119	-69.70	0.515	0.429	0.459	**0.165**	0.549	0.844	0.846	0.787	0.851
GradAscent w. SAGE	**0.369**↑ 0.046	**0.119**	**-59.36**	**0.522**	**0.381**	**0.437**	**0.165**	**0.557**	0.791	0.794	0.732	0.793
GradDiff	0.451	0.143	**-59.98**	0.551	**0.330**	0.466	0.007	0.455	0.797	0.797	0.675	0.814
GradDiff w. SAGE	**0.524**↑ 0.073	**0.143**	-60.34	**0.554**	0.349	**0.444**	**0.014**	**0.465**	0.801	0.799	0.691	0.813
NPO	0.329	0.133	-76.85	0.520	0.459	**0.439**	0.097	**0.551**	0.884	0.882	0.860	0.881
NPO w. SAGE	**0.395**↑ 0.066	**0.133**	**-73.15**	**0.529**	**0.429**	0.439	**0.165**	0.551	0.862	0.864	0.819	0.854
RMU	0.522	**0.255**	-98.03	**0.563**	0.735	0.523	0.014	0.478	0.985	0.990	0.988	0.996
RMU w. SAGE	**0.594**↑ 0.071	0.255	**-97.04**	0.555	**0.718**	**0.510**	**0.029**	**0.484**	0.983	0.985	0.980	0.989
SatImp	0.581	0.179	-96.43	**0.567**	0.695	0.519	**0.014**	0.460	0.979	0.982	0.990	0.994
SatImp w. SAGE	**0.584**↑ 0.003	**0.171**	**-95.20**	0.565	**0.657**	**0.499**	**0.014**	**0.469**	0.971	0.976	0.983	0.988
WGA	0.571	**0.129**	**-87.07**	**0.569**	0.519	0.429	**0.014**	0.471	0.931	0.934	0.927	0.948
WGA w. SAGE	**0.582**↑ 0.011	0.132	**-87.07**	0.566	**0.518**	**0.395**	**0.014**	**0.482**	0.931	0.934	0.931	0.945
DPO	0.489	**0.305**	-97.66	**0.559**	0.734	0.579	**0.029**	**0.520**	0.985	0.988	0.998	0.996
DPO w. SAGE	**0.511**↑ 0.022	0.320	**-97.41**	0.554	**0.725**	**0.559**	**0.029**	0.519	0.983	0.987	0.996	0.996
PDU	**0.032**	**0.027**	**-81.16**	0.556	**0.300**	**0.367**	**0.266**	**0.579**	0.857	0.904	0.996	0.788
PDU w. SAGE	**0.032**	**0.027**	-96.55	**0.573**	0.447	0.423	0.054	0.537	0.974	0.983	0.976	0.917
Forget05 / Retain95												
Vanilla	0.042	0.043	15.11	0.264	0.070	0.389	4.31e-4	0.674	0.248	0.264	0.348	0.243
Fully Fine-tuned	0.602	0.577	-99.78	0.550	0.838	0.746	2.96e-13	0.428	0.998	0.999	0.999	0.999
GradAscent	**0.032**	**0.030**	**12.17**	0.031	**0.002**	**0.387**	0.022	**0.648**	0.273	0.283	0.340	0.321
GradAscent w. SAGE	**0.032**	**0.030**	-17.55	**0.260**	0.032	0.489	**0.328**	0.615	0.432	0.473	0.505	0.399
GradDiff	0.163	0.118	**33.89**	0.350	**0.009**	**0.299**	2.41e-8	0.378	0.163	0.144	0.188	0.107
GradDiff w. SAGE	**0.274**↑ 0.111	**0.117**	36.24	**0.498**	0.018	0.321	**4.02e-6**	**0.427**	0.145	0.129	0.159	0.095
NPO	**0.032**	**0.030**	**-21.77**	0.423	0.133	0.469	**0.022**	0.672	0.501	0.500	0.516	0.424
NPO w. SAGE	**0.032**	**0.030**	43.42	**0.492**	**0.042**	**0.434**	9.67e-4	**0.705**	0.081	0.083	0.181	0.156
RMU	0.231	**0.162**	-87.87	0.518	**0.438**	0.430	6.73e-6	0.503	0.918	0.922	0.900	0.908
RMU w. SAGE	**0.387**↑ 0.156	0.163	**-85.54**	**0.552**	0.439	**0.429**	**0.003**	**0.521**	0.906	0.907	0.912	0.872
SatImp	0.543	0.172	-92.53	**0.545**	0.511	0.423	9.59e-10	0.450	0.946	0.952	0.933	0.945
SatImp w. SAGE	**0.582**↑ 0.038	**0.172**	**-89.93**	0.544	**0.478**	**0.411**	**8.11e-8**	**0.472**	0.929	0.936	0.916	0.919
WGA	0.502	0.125	-48.47	**0.552**	0.235	0.329	8.06e-7	0.491	0.642	0.670	0.659	0.584
WGA w. SAGE	**0.567**↑ 0.065	**0.123**	**-40.60**	0.548	**0.227**	**0.325**	**4.75e-5**	**0.517**	0.593	0.620	0.622	0.534
DPO	0.267	0.222	-89.19	**0.430**	0.514	0.065	**2.83e-4**	**0.578**	0.919	0.931	0.921	0.909
DPO w. SAGE	**0.478**↑ 0.211	**0.219**	**-79.29**	0.194	**0.392**	**0.000**	7.54e-5	0.532	0.866	0.868	0.645	0.801
PDU	0.045	0.051	**27.68**	**0.262**	0.029	0.197	**3.08e-12**	**0.796**	0.130	0.183	0.617	0.210
PDU w. SAGE	**0.057**↑ 0.012	**0.050**	52.32	0.015	**3.12e-6**	**0.044**	6.74e-37	0.457	0.056	0.026	0.149	0.128
Forget10 / Retain90												
Vanilla	0.042	0.042	18.70	0.264	0.068	0.379	1.83e-5	0.680	0.265	0.272	0.349	0.253
Fully Fine-tuned	0.602	0.586	-98.97	0.550	0.832	0.716	8.08e-22	0.446	0.993	0.994	0.996	0.994
GradAscent	**0.032**	**0.030**	**4.438**	0.000	**0.000**	**0.008**	1.06e-239	7.96e-34	0.394	0.360	0.448	0.363
GradAscent w. SAGE	**0.032**	**0.030**	14.82	**0.149**	0.011	0.439	**0.468**	**0.633**	0.289	0.296	0.299	0.305
GradDiff	0.144	0.093	**55.13**	0.202	**0.001**	**0.179**	**3.28e-14**	0.300	0.060	0.049	0.142	0.024
GradDiff w. SAGE	**0.263**↑ 0.120	**0.092**	55.53	**0.499**	0.004	0.244	**2.24e-17**	**0.429**	0.050	0.046	0.149	0.022
NPO	**0.032**	**0.030**	-23.66	**0.403**	0.162	0.461	**7.75e-4**	0.702	0.538	0.532	0.554	0.428
NPO w. SAGE	**0.032**	**0.030**	**10.43**	0.006	**5.1e-4**	**0.092**	1.12e-19	**0.755**	0.317	0.323	0.362	0.314
RMU	0.149	0.115	-75.15	0.479	0.249	0.368	9.99e-7	0.541	0.841	0.848	0.795	0.791
RMU w. SAGE	**0.209**↑ 0.059	**0.109**	**-66.79**	**0.529**	**0.229**	**0.361**	**0.002**	**0.560**	0.791	0.796	0.790	0.714
SatImp	0.543	0.132	-89.82	0.523	0.442	0.394	1.03e-14	0.492	0.929	0.938	0.900	0.907
SatImp w. SAGE	**0.568**↑ 0.025	**0.132**	**-84.45**	**0.537**	**0.407**	**0.391**	**9.91e-11**	**0.518**	0.894	0.905	0.856	0.853
WGA	0.496	0.080	**-5.313**	0.531	0.099	0.293	3.6e-5	0.562	0.353	0.419	0.406	0.276
WGA w. SAGE	**0.496**↑ 0.001	**0.080**	37.00	**0.532**	**0.051**	**0.265**	**0.984**	**0.602**	0.137	0.160	0.203	0.108
DPO	0.233	0.199	-85.68	**0.099**	0.445	**4e-4**	**2.55e-8**	**0.619**	0.899	0.912	0.915	0.835
DPO w. SAGE	**0.447**↑ 0.214	**0.199**	**-80.17**	0.029	**0.337**	0.004	2.55e-8	0.559	0.881	0.878	0.436	0.771
PDU	0.147	**0.119**	**-0.403**	0.398	**0.151**	**0.292**	1.75e-9	**0.743**	0.333	0.389	0.819	0.277
PDU w. SAGE	**0.197**↑ 0.050	0.138	-32.27	**0.530**	0.166	0.320	**1.31e-4**	0.681	0.489	0.585	0.900	0.409

F Additional Ablation Results

F.1 Detailed Results for Ablation Study

Table 18, 19, 20, 21 report full results on the ablation study on retain proxy size, SVD trunc. ratio, regularization λ, and alpha α.

Table 22 reports full results of integration with other plug-in methods.

Table 23 reports full results for ablation study of regularization λ for the MUSE Dataset.

F.2 Sensitivity and Retain-Proxy Sampling

TOFU proxy seeds. We fix source-unlearned checkpoints, evaluation splits, and hyperparameters, and reconstruct retained geometry with proxy sampling seeds 0, 100, and 200. We evaluate NPO, UNDIAL, SatImp, and WGA on Llama-3.1-8B-Instruct across all three TOFU forget ratios. Table 24 reports mean and standard deviation over the three seeds. This experiment isolates the variability of retained-geometry estimation; it does not rerun source-unlearning training. ES-Retain improves over each corresponding source checkpoint for every tested seed. The stability result concerns this retention improvement and does not imply that utility and privacy improve in every setting.

MUSE-Books document choice and budget. We vary the retained-book sampling seed over $\{0, 100, 200\}$ and the calibration budget over $\{2, 3, 4\}$ books, holding the other hyperparameters fixed. Table 25 preserves all nine configurations for each source method, rather than combining different book budgets into a single sampling standard deviation. Every tested configuration improves Retain ROUGE over the source checkpoint, while ES-Un remains approximately unchanged. This supports the stability of retention recovery across the tested document choices and budgets. Privacy leakage, however, is higher than the source value in these configurations, illustrating that stable retention recovery does not establish improvement of every unlearning metric.

G Additional Visualizations

G.1 Visualization of Overall Performance

In Figure 5, we plot one point per method variant (`base` or `+SAGE`) within each (model, forget ratio) condition, and connect paired points to show the per-method effect of SAGE. In Figure 6, for each model m, method a, and variant $v \in \{\text{base}, \text{sage}\}$, we aggregate across forgetting ratios $f \in \{1, 5, 10\}$ by arithmetic mean:

$$\bar{ES}_{Un}(m, a, v) = \frac{1}{3} \sum_{f \in \{1,5,10\}} ES_{Un}(m, f, a, v), \qquad \bar{ES}_{Re}(m, a, v) = \frac{1}{3} \sum_{f \in \{1,5,10\}} ES_{Re}(m, f, a, v).$$

The fitted curves are trend summaries. Across both visualizations, SAGE-augmented points consistently shift toward the upper-left region (lower Forget E.S., higher Retain E.S.), and the **w.SAGE** fitted trend dominates the baseline trend in most regimes. This indicates that SAGE systematically improves the forgetting-retention trade-off across model scales and forgetting intensities, while preserving strong cross-method consistency.

G.2 Visualization of Suppression Differences

We provide additional visualizations for the suppression behavior discussed in **Mechanistic Discussion: Differential Suppression**. For each editable matrix, we measure the suppression difference between the retain-side and forget-side response reduction. Positive values indicate that SAGE suppresses the update response more strongly on retain activations than on forget activations. We use TOFU Forget10 with WGA and SatImp on Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct as representative cases.

Figure 7 first aggregates the suppression difference at the layer level. For each transformer layer, we average over all editable attention and MLP projection matrices. The largest suppression differences typically appear in the first few layers, then drop quickly in the early-to-middle layers. After this

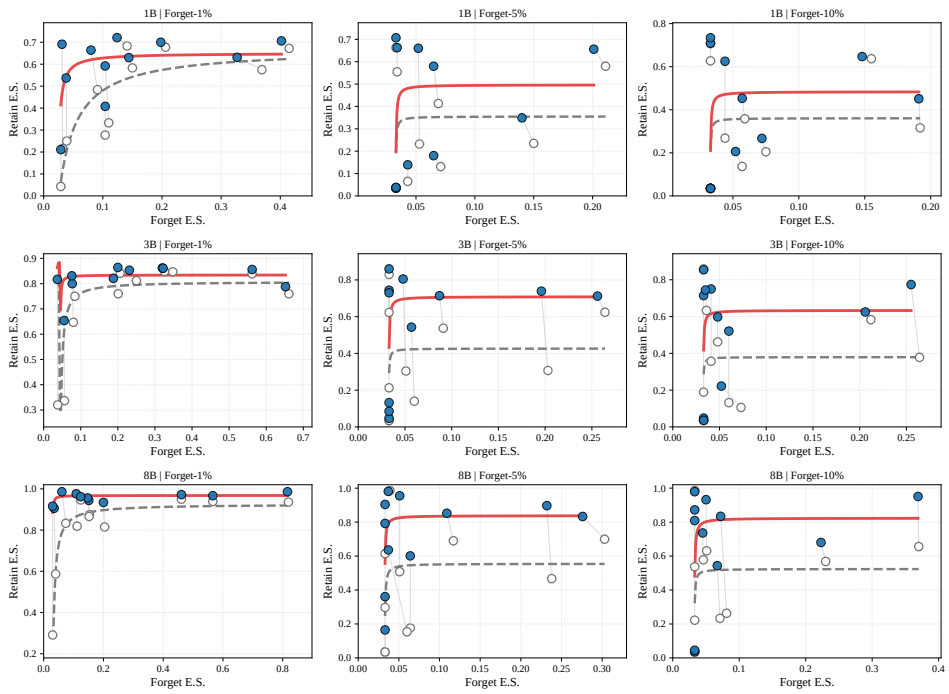

Figure 5: **Pareto-style trade-off across model scales and forgetting ratios (9 panels).** Each panel corresponds to one (model size, forgetting ratio) pair. White markers denote baseline unlearning methods, while blue markers denote their SAGE-augmented counterparts. Gray connectors link each baseline method to its +SAGE variant, visualizing method-wise movement after applying SAGE. The two fitted trend lines summarize the overall trade-off geometry for Baselines and Baselines **w. SAGE**, respectively. The x-axis is Forget E.S. (lower is better), and the y-axis is Retain E.S. (higher is better).

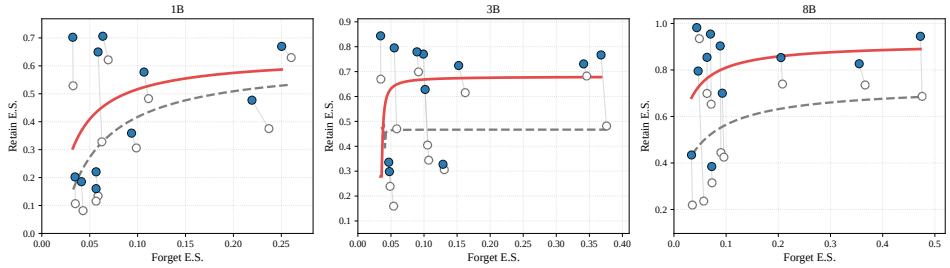

Figure 6: **Model-wise averaged Pareto-style trade-off (3 panels).** Each panel corresponds to one model scale (1B/3B/8B). Marker semantics are identical to the 3-panel figure. Each point here is averaged over forgetting ratios (1%, 5%, 10%).

Table 18: Ablation on retain-proxy size for SAGE on **Llama-3.2-1B-Instruct**. We use SVD truncation energy 0.9 and $\lambda = 5$. Default proxy size 128 is shaded in blue. Best and second-best results within each method block are highlighted in **bold** and underlined. Priv. Leak is rounded to two decimals, and all other metrics are rounded to three decimals.

Method	Proxy Size	ES Re. ↑	ES Un. ↓	Priv. Leak	MU ↑	F-QA Prob. ↓	F-QA Rg. ↓	FQ ↑	MIA-L	MIA-MK	MIA-MK++	MIA-Z	MIA-GN
Forget01 / Retain99													
WGA w. SAGE	16	0.719	0.154	-90.44	0.601	0.498	0.434	0.014	0.946	0.949	0.848	0.959	0.830
WGA w. SAGE	32	**0.724**	0.154	-91.03	0.602	0.509	0.434	0.014	0.947	0.953	0.859	0.964	0.830
WGA w. SAGE	64	0.722	0.153	-89.96	**0.602**	0.493	0.436	0.014	0.943	0.947	0.848	0.956	0.824
WGA w. SAGE	128	0.721	**0.124**	**-88.90**	0.601	**0.467**	**0.430**	**0.029**	**0.938**	**0.941**	**0.820**	**0.941**	**0.817**
WGA w. SAGE	256	0.723	0.151	-90.32	0.602	0.493	0.436	**0.029**	0.943	0.949	0.844	0.956	0.825
RMU w. SAGE	16	0.598	0.033	86.66	0.581	0.007	**0.191**	**0.266**	0.012	**0.012**	0.464	0.004	**0.000**
RMU w. SAGE	32	0.648	0.031	86.42	0.587	**0.006**	0.200	0.097	**0.011**	0.013	0.441	**0.002**	**0.000**
RMU w. SAGE	64	0.650	**0.030**	**84.53**	0.590	0.009	0.205	0.165	0.017	0.023	0.472	0.012	**0.000**
RMU w. SAGE	128	**0.691**	0.031	**84.53**	**0.593**	0.010	0.224	0.054	0.019	0.023	0.453	0.013	**0.000**
RMU w. SAGE	256	0.689	0.031	84.65	0.593	0.011	0.194	0.014	0.019	0.022	**0.428**	0.013	**0.000**
Forget05 / Retain95													
WGA w. SAGE	16	0.629	**0.034**	54.08	0.591	0.006	**0.174**	0.002	0.014	0.018	**0.269**	0.009	0.008
WGA w. SAGE	32	0.635	**0.034**	54.22	0.592	0.006	0.178	0.003	0.014	0.017	0.273	0.008	**0.007**
WGA w. SAGE	64	0.671	0.034	**53.39**	0.594	0.007	0.196	**0.009**	0.018	0.022	0.272	0.011	0.008
WGA w. SAGE	128	0.662	0.034	54.40	0.591	**0.005**	0.189	0.001	0.013	**0.015**	0.287	**0.007**	0.009
WGA w. SAGE	256	**0.687**	0.034	54.24	**0.596**	0.005	0.185	0.000	**0.013**	0.017	0.297	0.009	0.010
RMU w. SAGE	16	0.692	**0.033**	52.52	0.591	**0.002**	**0.090**	0.000	**0.014**	**0.028**	0.443	0.027	0.063
RMU w. SAGE	32	0.703	**0.033**	52.15	0.591	0.003	0.103	**0.000**	0.016	0.030	0.446	0.021	0.052
RMU w. SAGE	64	0.701	**0.033**	**50.86**	0.592	0.004	0.112	0.000	0.020	0.038	0.446	0.019	0.042
RMU w. SAGE	128	0.707	**0.033**	51.12	**0.592**	0.008	0.142	0.000	0.025	0.036	**0.438**	**0.014**	**0.035**
RMU w. SAGE	256	**0.711**	**0.033**	51.12	**0.592**	0.006	0.134	0.000	0.021	0.036	0.443	0.017	0.037
Forget10 / Retain90													
WGA w. SAGE	16	0.713	**0.033**	59.44	0.596	0.003	0.050	0.000	0.010	0.016	0.828	0.003	**0.002**
WGA w. SAGE	32	0.716	**0.033**	59.64	0.599	**0.003**	**0.045**	0.000	**0.009**	**0.014**	**0.818**	**0.002**	0.003
WGA w. SAGE	64	0.724	**0.033**	58.84	0.599	0.004	0.057	0.000	0.015	0.019	0.840	0.004	0.003
WGA w. SAGE	128	0.734	**0.033**	58.76	0.600	0.004	0.074	0.000	0.015	0.020	0.852	0.004	0.004
WGA w. SAGE	256	**0.764**	**0.033**	**58.37**	**0.603**	0.005	0.081	**0.000**	0.016	0.022	0.859	0.005	0.005
RMU w. SAGE	16	0.699	**0.033**	59.06	0.592	**0.002**	**0.070**	0.000	**0.016**	**0.018**	**0.278**	**0.021**	**0.015**
RMU w. SAGE	32	0.703	**0.033**	55.90	0.593	0.005	0.091	0.000	0.028	0.037	0.419	0.025	0.017
RMU w. SAGE	64	0.705	**0.033**	55.88	0.593	0.007	0.102	0.000	0.031	0.038	0.412	0.029	0.017
RMU w. SAGE	128	0.709	0.033	55.37	**0.595**	0.013	0.152	**0.000**	0.030	0.041	0.427	0.032	0.027
RMU w. SAGE	256	**0.716**	0.033	**55.30**	**0.595**	0.009	0.135	**0.000**	0.030	0.041	0.438	0.032	0.025

sharp decrease, the average suppression difference mildly rises again in later layers before decreasing toward the final layers. This layer-wise pattern suggests that SAGE's retain-side suppression is strongest near the lower part of the network, while still leaving a weaker but nonzero effect in deeper layers.

Figures 8–11 break down the same quantity by matrix type. The attention input projections W_q, W_k, and W_v show relatively even suppression-difference profiles across layers. In contrast, the output projection W_o has larger suppression differences mainly in the later layers. The three MLP matrices, W_{gate}, W_{up}, and W_{down}, show the opposite trend: their largest suppression differences are concentrated in the earlier layers. This separation indicates that SAGE affects attention and feed-forward components with distinct layer-wise structures rather than applying a uniform shrinkage pattern everywhere.

Table 19: Ablation on SVD truncation energy for SAGE on **Llama-3.2-1B-Instruct**. We use retain-proxy size 128 and $\lambda = 5$. The selected SVD truncation energy is 0.9, highlighted in blue. Priv. Leak is rounded to two decimals, and all other metrics are rounded to three decimals. Best and second-best results within each method block are highlighted in **bold** and underlined.

Method	SVD Energy	ES Re. ↑	ES Un. ↓	Priv. Leak	MU ↑	F-QA Prob. ↓	F-QA Rg. ↓	FQ ↑	MIA-L	MIA-MK	MIA-MK++	MIA-Z	MIA-GN
Forget01 / Retain99													
WGA w. SAGE	0.5	0.715	0.147	-90.20	0.600	0.485	0.430	0.014	0.943	0.948	0.839	0.954	0.825
WGA w. SAGE	0.6	0.711	0.147	-89.85	0.601	0.482	0.433	0.014	0.943	0.946	0.834	0.951	0.821
WGA w. SAGE	0.7	0.715	0.154	-90.32	0.601	0.493	0.434	0.014	0.943	0.949	0.845	0.958	0.828
WGA w. SAGE	0.8	0.719	0.124	-88.78	0.600	0.467	0.434	**0.029**	0.939	0.941	**0.815**	**0.938**	**0.813**
WGA w. SAGE	0.9	0.721	0.124	-88.90	**0.601**	0.467	0.430	**0.029**	0.938	0.941	0.820	0.941	0.817
WGA w. SAGE	0.95	0.721	**0.121**	**-88.67**	0.601	**0.465**	**0.428**	**0.029**	**0.938**	**0.940**	0.820	0.941	0.815
WGA w. SAGE	1	**0.739**	0.139	-92.80	0.598	0.578	0.473	0.007	0.955	0.962	0.921	0.983	0.849
RMU w. SAGE	0.5	0.432	0.031	85.12	0.565	0.009	0.189	0.054	0.015	0.020	0.534	0.007	**0.000**
RMU w. SAGE	0.6	0.490	0.031	86.07	0.570	0.007	**0.180**	0.097	0.012	0.015	0.521	0.005	**0.000**
RMU w. SAGE	0.7	0.563	**0.030**	85.24	0.579	0.008	0.190	**0.266**	0.014	0.019	0.510	0.006	**0.000**
RMU w. SAGE	0.8	0.633	0.031	84.65	0.586	0.010	0.187	**0.266**	0.017	0.022	0.476	0.011	**0.000**
RMU w. SAGE	0.9	0.691	0.031	84.53	0.593	0.010	0.224	0.054	0.019	0.023	0.453	0.013	**0.000**
RMU w. SAGE	0.95	**0.709**	0.031	**83.59**	**0.597**	0.013	0.192	0.029	0.024	0.028	0.436	0.018	**0.000**
RMU w. SAGE	1	0.677	0.032	88.78	0.589	**0.001**	0.211	0.001	**0.001**	**0.001**	**0.379**	**0.000**	**0.000**
Forget05 / Retain95													
WGA w. SAGE	0.5	0.585	0.034	53.80	0.588	0.006	**0.178**	0.006	0.015	0.019	0.269	0.009	0.008
WGA w. SAGE	0.6	0.611	0.034	**53.64**	0.591	0.007	0.184	**0.012**	0.016	0.020	0.267	0.009	**0.008**
WGA w. SAGE	0.7	0.632	0.034	54.00	0.592	0.006	0.180	0.004	0.015	0.018	0.267	0.009	0.008
WGA w. SAGE	0.8	**0.664**	**0.034**	53.85	**0.596**	0.006	0.184	0.002	0.016	0.019	0.269	0.009	**0.008**
WGA w. SAGE	0.9	0.662	0.034	54.40	0.591	0.005	0.189	0.001	0.013	0.015	0.287	0.007	0.009
WGA w. SAGE	0.95	0.663	0.034	54.02	0.592	0.005	0.182	0.000	0.013	0.018	0.293	0.008	0.011
WGA w. SAGE	1	0.554	0.035	55.34	0.579	**0.003**	0.199	0.000	**0.008**	**0.009**	**0.241**	**0.005**	0.013
RMU w. SAGE	0.5	0.651	**0.033**	54.26	0.584	**0.000**	**0.054**	0.000	**0.013**	**0.016**	**0.253**	0.033	**0.018**
RMU w. SAGE	0.6	0.662	**0.033**	53.60	0.585	0.000	0.056	0.000	0.015	0.021	0.297	0.033	0.024
RMU w. SAGE	0.7	0.680	**0.033**	53.51	0.588	0.000	0.069	0.000	0.015	0.021	0.392	0.035	0.044
RMU w. SAGE	0.8	0.689	**0.033**	52.37	0.591	0.002	0.090	0.000	0.015	0.029	0.435	0.026	0.064
RMU w. SAGE	0.9	0.707	**0.033**	51.12	0.592	0.008	0.142	**0.000**	0.025	0.036	0.438	**0.014**	0.035
RMU w. SAGE	0.95	0.724	**0.033**	50.49	0.593	0.011	0.156	0.000	0.034	0.040	0.412	0.017	0.021
RMU w. SAGE	1	**0.726**	**0.033**	**49.15**	**0.597**	0.008	0.158	0.000	0.037	0.049	0.373	0.026	0.022
Forget10 / Retain90													
WGA w. SAGE	0.5	0.677	**0.033**	59.49	0.594	0.003	**0.045**	0.000	0.008	0.015	**0.813**	0.002	**0.002**
WGA w. SAGE	0.6	0.685	**0.033**	59.06	0.595	0.003	0.056	0.000	0.010	0.018	0.840	0.003	0.002
WGA w. SAGE	0.7	0.699	**0.033**	59.24	0.596	0.003	0.045	0.000	0.010	0.017	0.818	0.003	0.002
WGA w. SAGE	0.8	0.723	**0.033**	59.12	0.599	0.004	0.050	0.000	0.012	0.018	0.819	0.003	0.003
WGA w. SAGE	0.9	0.734	**0.033**	58.76	0.600	0.004	0.074	0.000	0.015	0.020	0.852	0.004	0.004
WGA w. SAGE	0.95	**0.748**	**0.033**	**58.40**	**0.603**	0.005	0.082	**0.000**	0.015	0.022	0.865	0.004	0.007
WGA w. SAGE	1	0.607	**0.033**	60.51	0.591	**0.002**	0.081	0.000	**0.005**	**0.009**	0.847	**0.002**	0.014
RMU w. SAGE	0.5	0.679	**0.033**	60.01	0.590	**0.000**	**0.046**	0.000	0.013	0.012	**0.168**	0.016	0.015
RMU w. SAGE	0.6	0.685	**0.033**	60.15	0.589	0.000	0.057	0.000	**0.011**	**0.011**	0.208	**0.016**	**0.012**
RMU w. SAGE	0.7	0.696	**0.033**	59.89	0.591	0.000	0.065	0.000	0.012	0.013	0.293	0.019	0.015
RMU w. SAGE	0.8	0.703	**0.033**	56.97	0.592	0.003	0.077	0.000	0.025	0.031	0.426	0.022	0.022
RMU w. SAGE	0.9	0.709	0.033	**55.37**	0.595	0.013	0.152	**0.000**	0.030	0.041	0.427	0.032	0.027
RMU w. SAGE	0.95	0.716	**0.033**	55.48	**0.595**	0.008	0.145	0.000	0.029	0.040	0.460	0.031	0.027
RMU w. SAGE	1	**0.726**	**0.033**	58.53	0.594	0.003	0.109	0.000	0.018	0.021	0.411	0.023	0.018

Table 20: Ablation on the retain-geometry regularization strength λ for SAGE on **Llama-3.2-1B-Instruct**. We use retain-proxy size 128 and SVD truncation energy 0.9. Default regularization strength $\lambda = 5$ is shaded in blue. Best and second-best results within each method block are highlighted in **bold** and underlined. Priv. Leak is rounded to two decimals, and all other metrics are rounded to three decimals.

Method	λ	ES Re. ↑	ES Un. ↓	Priv. Leak	MU ↑	F-QA Prob. ↓	F-QA Rg. ↓	FQ ↑	MIA-L	MIA-MK	MIA-MK++	MIA-Z	MIA-GN
Forget01 / Retain99													
WGA w. SAGE	0	0.685	0.141	-87.25	0.599	0.441	0.432	0.014	0.932	0.932	0.794	0.934	0.810
WGA w. SAGE	0.001	0.714	0.147	-89.85	0.601	0.483	0.432	0.014	0.943	0.946	0.836	0.953	0.822
WGA w. SAGE	0.01	0.714	**0.120**	**-86.19**	0.600	**0.437**	**0.420**	**0.029**	**0.927**	**0.927**	**0.791**	**0.929**	**0.803**
WGA w. SAGE	0.1	**0.722**	0.121	-88.43	0.601	0.461	0.427	**0.029**	0.936	0.939	0.814	0.938	0.812
WGA w. SAGE	1	0.721	0.124	-88.78	0.601	0.467	0.429	**0.029**	0.938	0.941	0.820	0.941	0.817
WGA w. SAGE	2	**0.722**	0.124	-88.78	0.601	0.467	0.428	**0.029**	0.939	0.941	0.819	0.941	0.817
WGA w. SAGE	3	0.721	0.124	-88.78	0.601	0.467	0.430	**0.029**	0.938	0.941	0.819	0.941	0.818
WGA w. SAGE	5	0.721	0.124	-88.90	0.601	0.467	0.430	**0.029**	0.938	0.941	0.820	0.941	0.817
WGA w. SAGE	8	**0.722**	0.155	-91.15	**0.602**	0.513	0.430	0.014	0.949	0.953	0.862	0.964	0.831
WGA w. SAGE	10	0.720	0.124	-88.90	0.601	0.467	0.431	**0.029**	0.938	0.941	0.821	0.941	0.817
RMU w. SAGE	0	0.210	0.032	**48.05**	0.522	0.051	0.200	**0.405**	0.188	0.216	0.554	0.138	0.001
RMU w. SAGE	0.001	0.479	0.031	86.07	0.570	**0.007**	**0.192**	0.054	**0.012**	**0.015**	0.474	**0.005**	**0.000**
RMU w. SAGE	0.01	0.645	0.031	84.77	0.588	0.008	0.210	0.097	0.014	0.022	0.471	**0.005**	**0.000**
RMU w. SAGE	0.1	0.689	**0.030**	84.30	0.592	0.010	0.201	0.165	0.019	0.024	0.465	0.013	**0.000**
RMU w. SAGE	1	0.691	0.033	83.83	**0.593**	0.012	0.217	0.054	0.022	0.027	0.441	0.015	**0.000**
RMU w. SAGE	2	0.691	0.033	83.83	**0.593**	0.012	0.209	0.054	0.022	0.027	**0.439**	0.015	**0.000**
RMU w. SAGE	3	**0.693**	0.031	84.65	**0.593**	0.010	0.214	0.097	0.019	0.022	0.454	0.013	**0.000**
RMU w. SAGE	5	0.691	0.031	84.53	**0.593**	0.010	0.224	0.054	0.019	0.023	0.453	0.013	**0.000**
RMU w. SAGE	8	0.691	0.031	84.53	**0.593**	0.010	0.200	0.054	0.019	0.023	0.452	0.013	**0.000**
RMU w. SAGE	10	0.691	0.033	83.83	**0.593**	0.012	0.204	0.054	0.022	0.027	**0.439**	0.015	**0.000**
Forget05 / Retain95													
WGA w. SAGE	0	0.553	0.034	**53.43**	0.593	0.007	**0.166**	0.001	0.016	0.022	0.285	0.010	0.009
WGA w. SAGE	0.001	0.614	**0.033**	54.35	0.588	0.005	0.175	0.001	0.013	0.016	**0.261**	0.007	0.009
WGA w. SAGE	0.01	0.669	0.034	53.56	**0.595**	0.007	0.189	**0.009**	0.017	0.021	0.268	0.010	0.008
WGA w. SAGE	0.1	0.670	0.034	54.43	0.591	0.005	0.181	0.000	0.013	0.015	0.282	0.007	**0.007**
WGA w. SAGE	1	0.674	0.034	54.27	0.593	0.005	0.178	0.000	0.014	0.016	0.281	0.007	**0.007**
WGA w. SAGE	2	**0.677**	0.034	54.25	0.593	0.005	0.182	0.001	0.014	0.017	0.281	0.008	**0.007**
WGA w. SAGE	3	0.655	**0.033**	54.60	0.591	**0.004**	0.185	0.000	**0.011**	**0.014**	0.291	**0.006**	0.010
WGA w. SAGE	5	0.662	0.034	54.40	0.591	0.005	0.189	0.001	0.013	0.015	0.287	0.007	0.009
WGA w. SAGE	8	0.663	0.034	54.41	0.592	0.005	0.191	0.000	0.013	0.015	0.288	0.007	0.009
WGA w. SAGE	10	0.662	0.034	54.40	0.591	0.005	0.187	0.000	0.013	0.016	0.287	0.007	0.009
RMU w. SAGE	0	0.654	**0.033**	**49.31**	0.584	**0.000**	**0.024**	**0.000**	0.046	0.048	**0.229**	0.045	**0.032**
RMU w. SAGE	0.001	0.678	**0.033**	54.01	0.587	**0.000**	0.072	**0.000**	**0.012**	**0.018**	0.372	0.033	0.041
RMU w. SAGE	0.01	0.701	**0.033**	52.45	0.590	0.003	0.113	**0.000**	0.016	0.028	0.424	0.022	0.048
RMU w. SAGE	0.1	0.703	**0.033**	50.96	**0.592**	0.006	0.133	**0.000**	0.023	0.037	0.427	0.015	0.038
RMU w. SAGE	1	**0.707**	**0.033**	51.15	**0.592**	0.008	0.142	**0.000**	0.025	0.036	0.439	**0.014**	0.035
RMU w. SAGE	2	**0.707**	**0.033**	51.13	**0.592**	0.008	0.144	**0.000**	0.025	0.036	0.439	**0.014**	0.035
RMU w. SAGE	3	**0.707**	**0.033**	51.14	**0.592**	0.008	0.143	**0.000**	0.025	0.036	0.438	**0.014**	0.035
RMU w. SAGE	5	**0.707**	**0.033**	51.12	**0.592**	0.008	0.142	**0.000**	0.025	0.036	0.438	**0.014**	0.035
RMU w. SAGE	8	**0.707**	**0.033**	51.12	**0.592**	0.008	0.140	**0.000**	0.025	0.036	0.438	**0.014**	0.034
RMU w. SAGE	10	**0.707**	**0.033**	51.13	**0.592**	0.008	0.144	**0.000**	0.025	0.036	0.439	**0.014**	0.034
Forget10 / Retain90													
WGA w. SAGE	0	0.629	**0.033**	60.66	0.592	**0.001**	**0.023**	**0.000**	**0.004**	**0.008**	**0.668**	**0.001**	0.036
WGA w. SAGE	0.001	0.695	**0.033**	59.44	0.595	0.003	0.048	**0.000**	0.010	0.016	0.819	0.003	**0.002**
WGA w. SAGE	0.01	0.722	**0.033**	58.94	0.599	0.004	0.060	**0.000**	0.014	0.019	0.849	0.004	0.003
WGA w. SAGE	0.1	0.730	**0.033**	58.97	**0.603**	0.004	0.062	**0.000**	0.014	0.019	0.844	0.003	0.004
WGA w. SAGE	1	0.734	**0.033**	58.78	0.601	0.004	0.071	**0.000**	0.015	0.020	0.851	0.004	0.004
WGA w. SAGE	2	**0.736**	**0.033**	58.77	0.600	0.004	0.071	**0.000**	0.015	0.020	0.852	0.004	0.004
WGA w. SAGE	3	**0.736**	**0.033**	**58.76**	0.600	0.004	0.073	**0.000**	0.015	0.020	0.852	0.004	0.004
WGA w. SAGE	5	0.734	**0.033**	**58.76**	0.600	0.004	0.074	**0.000**	0.015	0.020	0.852	0.004	0.004
WGA w. SAGE	8	0.734	**0.033**	**58.76**	0.600	0.004	0.073	**0.000**	0.015	0.020	0.852	0.004	0.004
WGA w. SAGE	10	**0.736**	**0.033**	**58.76**	0.600	0.004	0.074	**0.000**	0.015	0.020	0.852	0.004	0.004
RMU w. SAGE	0	0.706	**0.033**	58.86	0.592	**0.000**	**0.019**	**0.000**	0.019	0.019	**0.147**	0.021	0.016
RMU w. SAGE	0.001	0.692	**0.033**	59.51	0.592	0.001	0.065	**0.000**	**0.013**	**0.015**	0.262	**0.018**	**0.014**
RMU w. SAGE	0.01	0.704	**0.033**	56.15	0.592	0.006	0.088	**0.000**	0.029	0.036	0.378	0.026	0.019
RMU w. SAGE	0.1	**0.712**	**0.033**	**55.16**	0.594	0.011	0.137	**0.000**	0.030	0.042	0.440	0.032	0.026
RMU w. SAGE	1	0.707	**0.033**	55.57	0.592	0.006	0.104	**0.000**	0.033	0.040	0.392	0.029	0.018
RMU w. SAGE	2	0.709	**0.033**	55.37	**0.595**	0.013	0.150	**0.000**	0.030	0.041	0.427	0.032	0.027
RMU w. SAGE	3	0.711	**0.033**	55.37	**0.595**	0.013	0.153	**0.000**	0.030	0.041	0.427	0.032	0.027
RMU w. SAGE	5	0.709	**0.033**	55.37	**0.595**	0.013	0.152	**0.000**	0.030	0.041	0.427	0.032	0.027
RMU w. SAGE	8	0.709	**0.033**	55.36	**0.595**	0.014	0.150	**0.000**	0.030	0.041	0.427	0.032	0.027
RMU w. SAGE	10	0.709	**0.033**	55.39	0.594	0.014	0.151	**0.000**	0.030	0.041	0.428	0.032	0.027

Table 21: Ablation on fixed update-scaling coefficient α for SAGE on **Llama-3.2-1B-Instruct**. We use retain-proxy size 128, SVD truncation energy 0.9, and $\lambda = 5$. The default scaling coefficient $\alpha = 1$ is shaded in blue. Best and second-best results within each method block are highlighted in **bold** and underlined. Priv. Leak is rounded to two decimals, and all other metrics are rounded to three decimals.

Method	α	ES Re. ↑	ES Un. ↓	Priv. Leak	MU ↑	F-QA Prob. ↓	F-QA Rg. ↓	FQ ↑	MIA-L	MIA-MK	MIA-MK++	MIA-Z	MIA-GN
Forget01 / Retain99													
WGA w. SAGE	0.8	**0.733**	0.235	-95.75	0.601	0.650	0.484	0.007	0.972	0.978	0.960	0.994	0.887
WGA w. SAGE	1	0.728	0.189	-93.39	**0.602**	0.583	0.454	0.007	0.957	0.965	0.928	0.986	0.858
WGA w. SAGE	1.2	0.726	0.165	-91.85	0.602	0.549	0.444	0.007	0.953	0.957	0.894	0.981	0.844
WGA w. SAGE	1.5	0.723	0.153	-90.32	0.602	0.499	0.435	0.014	0.945	0.949	0.849	0.960	0.827
WGA w. SAGE	1.8	0.712	0.118	-84.77	0.601	0.418	0.415	0.029	0.915	0.919	0.771	0.914	0.794
WGA w. SAGE	2	0.700	0.115	-80.40	0.599	0.376	0.405	0.029	0.895	0.896	0.732	0.893	0.788
WGA w. SAGE	3	0.680	0.099	-59.50	0.598	0.270	0.392	0.029	0.783	0.786	0.628	0.793	0.777
WGA w. SAGE	4	0.660	0.093	-40.50	0.596	0.214	0.379	0.054	0.694	0.685	0.557	0.698	0.763
WGA w. SAGE	6	0.633	**0.088**	**-10.74**	0.595	**0.155**	**0.378**	**0.165**	**0.537**	**0.528**	**0.482**	**0.563**	**0.737**
RMU w. SAGE	0.8	**0.723**	0.099	-71.31	**0.596**	0.383	0.380	0.054	0.844	0.848	0.830	0.871	0.050
RMU w. SAGE	1	0.696	0.063	-45.45	0.594	0.265	0.337	0.007	0.704	0.711	0.754	0.724	0.013
RMU w. SAGE	1.2	0.691	0.054	-25.27	0.594	0.205	0.324	0.029	0.592	0.604	0.713	0.606	0.004
RMU w. SAGE	1.5	0.694	0.050	**13.22**	0.594	0.119	0.273	0.405	0.396	0.401	0.566	0.392	**0.000**
RMU w. SAGE	1.8	0.691	0.046	39.32	0.593	0.073	0.246	**0.919**	0.254	0.263	0.486	0.209	**0.000**
RMU w. SAGE	2	0.693	0.044	47.70	0.593	0.063	0.239	**0.919**	0.208	0.218	0.448	0.165	**0.000**
RMU w. SAGE	3	0.696	0.033	71.90	0.594	0.030	0.222	0.054	0.079	0.090	0.443	0.046	**0.000**
RMU w. SAGE	4	0.695	0.033	83.83	0.593	0.012	0.204	0.054	0.023	0.027	0.436	0.016	**0.000**
RMU w. SAGE	6	0.664	**0.030**	88.31	0.589	**0.003**	**0.171**	0.007	**0.005**	**0.003**	**0.393**	**0.000**	**0.000**
Forget05 / Retain95													
WGA w. SAGE	0.8	**0.748**	0.053	**-9.66**	**0.601**	0.110	0.277	0.040	0.356	0.424	0.557	0.323	0.047
WGA w. SAGE	1	0.739	0.043	25.29	0.599	0.051	0.244	0.328	0.132	0.201	0.427	0.133	0.027
WGA w. SAGE	1.2	0.727	0.039	40.28	0.598	0.031	0.232	**0.466**	0.064	0.106	0.363	0.070	0.019
WGA w. SAGE	1.5	0.710	0.036	49.09	0.598	0.016	0.215	0.088	0.033	0.049	0.298	0.030	0.009
WGA w. SAGE	1.8	0.698	0.035	52.38	0.596	0.010	0.199	0.009	0.022	0.028	**0.279**	0.016	0.008
WGA w. SAGE	2	0.692	0.034	53.36	0.595	0.008	0.188	0.006	0.019	0.022	0.280	0.011	**0.008**
WGA w. SAGE	3	0.628	0.033	54.88	0.590	0.004	0.173	0.000	0.009	**0.012**	0.307	**0.005**	0.014
WGA w. SAGE	4	0.570	0.033	54.87	0.585	0.002	0.171	0.000	0.007	0.013	0.359	0.005	0.023
WGA w. SAGE	6	0.416	**0.033**	54.73	0.570	**0.001**	**0.144**	0.000	**0.007**	0.013	0.453	0.008	0.039
RMU w. SAGE	0.8	**0.719**	0.034	50.19	**0.593**	0.021	0.180	**0.001**	0.041	0.042	0.411	0.032	**0.013**
RMU w. SAGE	1	0.707	**0.033**	51.12	0.592	0.008	0.142	0.000	0.025	0.036	0.438	**0.014**	0.035
RMU w. SAGE	1.2	0.699	**0.033**	52.51	0.591	0.003	0.107	0.000	**0.015**	**0.028**	0.418	0.026	0.050
RMU w. SAGE	1.5	0.686	**0.033**	51.08	0.586	0.001	0.076	0.000	0.024	0.037	0.396	0.038	0.059
RMU w. SAGE	1.8	0.660	**0.033**	51.25	0.583	0.000	0.058	0.000	0.025	0.036	**0.305**	0.040	0.022
RMU w. SAGE	2	0.650	**0.033**	46.56	0.581	0.000	0.046	0.000	0.055	0.066	0.563	0.051	0.053
RMU w. SAGE	3	0.587	**0.033**	47.32	0.572	0.000	**0.003**	0.000	0.081	0.061	0.703	0.059	0.045
RMU w. SAGE	4	0.495	**0.033**	39.79	0.562	0.000	0.008	0.000	0.107	0.109	0.787	0.090	0.154
RMU w. SAGE	6	0.334	**0.033**	**29.52**	0.533	**0.000**	0.013	0.000	0.112	0.174	0.688	0.133	0.109
Forget10 / Retain90													
WGA w. SAGE	0.8	**0.781**	0.046	**5.23**	0.601	0.080	0.242	0.037	0.254	0.350	0.800	0.224	0.021
WGA w. SAGE	1	0.769	0.037	38.07	0.601	0.029	0.181	**0.111**	0.076	0.148	0.847	0.063	0.008
WGA w. SAGE	1.2	0.764	0.035	50.86	**0.602**	0.013	0.137	0.000	0.029	0.069	0.875	0.021	0.006
WGA w. SAGE	1.5	0.746	**0.033**	58.10	0.601	0.006	0.086	0.000	0.017	0.024	0.871	0.005	**0.004**
WGA w. SAGE	1.8	0.730	**0.033**	59.48	0.602	0.003	0.051	0.000	0.011	0.015	0.803	0.003	0.009
WGA w. SAGE	2	0.718	**0.033**	59.93	0.601	0.002	0.040	0.000	**0.009**	**0.013**	0.764	**0.003**	0.029
WGA w. SAGE	3	0.649	**0.033**	55.64	0.596	0.001	0.019	0.000	0.028	0.039	0.827	0.024	0.186
WGA w. SAGE	4	0.578	**0.033**	55.97	0.588	0.000	0.013	0.000	0.026	0.037	0.842	0.025	0.805
WGA w. SAGE	6	0.389	**0.033**	54.52	0.552	**0.000**	**0.005**	0.000	0.039	0.046	**0.748**	0.044	0.932
RMU w. SAGE	0.8	**0.719**	0.040	54.42	**0.595**	0.025	0.186	0.000	0.043	0.047	0.384	0.033	0.021
RMU w. SAGE	1	0.709	0.033	55.37	0.595	0.013	0.152	0.000	0.030	0.041	0.427	0.032	0.027
RMU w. SAGE	1.2	0.707	0.035	55.86	0.592	0.005	0.094	0.000	0.032	0.038	0.369	0.027	**0.018**
RMU w. SAGE	1.5	0.694	**0.033**	59.01	0.588	0.001	0.062	0.000	**0.020**	**0.018**	**0.299**	**0.022**	0.025
RMU w. SAGE	1.8	0.674	**0.033**	57.43	0.584	0.000	0.037	0.000	0.026	0.028	0.352	0.025	0.023
RMU w. SAGE	2	0.661	**0.033**	54.49	0.582	0.000	0.029	0.000	0.040	0.046	0.729	0.029	0.035
RMU w. SAGE	3	0.591	**0.033**	55.41	0.571	**0.000**	**0.011**	0.000	0.045	0.041	0.815	0.030	0.061
RMU w. SAGE	4	0.517	**0.033**	45.68	0.557	0.000	0.016	0.000	0.092	0.101	0.767	0.082	0.119
RMU w. SAGE	6	0.370	**0.033**	**30.19**	0.496	0.000	0.042	**0.000**	0.263	0.196	0.536	0.165	0.338

Table 22: Full compositionality results of SAGE with GU/GRU for WGA, NPO, and DPO on TOFU using **Llama-3.2-1B-Instruct**. SAGE is applied either directly to the source update or post-hoc on top of GU/GRU. Blue-shaded rows indicate settings that include SAGE. Orange superscripts on "w. GU w. SAGE" and "w. GRU w. SAGE" indicate the absolute ES Re. gain obtained by adding SAGE on top of GU or GRU, respectively. Within each split and method block, the best value in each metric column is boldfaced and the second-best is underlined.

Method	WGA				NPO				DPO			
	ES Re. ↑	ES Un. ↓	Priv. Leak. → 0	MU ↑	ES Re. ↑	ES Un. ↓	Priv. Leak. → 0	MU ↑	ES Re. ↑	ES Un. ↓	Priv. Leak. → 0	MU ↑
Forget01 / Retain99												
base	0.683	**0.141**	**-86.78**	0.598	0.583	0.150	76.62	0.588	0.575	0.369	-98.94	0.575
w. SAGE	**0.721**	**0.124**	-88.90	0.601	0.630	**0.144**	-75.09	0.593	0.631	**0.327**	-97.87	0.583
w. GU	0.677	0.179	-92.80	**0.603**	0.548	0.178	-81.35	0.590	0.577	0.392	-99.06	0.576
w. GU w. SAGE	0.691 $^{\uparrow 0.014}$	0.188	-91.03	0.600	0.605 $^{\uparrow 0.056}$	0.160	**-74.97**	0.591	**0.634** $^{\uparrow 0.057}$	0.349	**-97.76**	**0.584**
w. GRU	0.681	**0.141**	**-86.78**	0.599	0.584	0.150	76.86	0.588	0.574	0.362	-98.94	0.575
w. GRU w. SAGE	0.715 $^{\uparrow 0.034}$	0.145	-89.61	0.601	**0.635** $^{\uparrow 0.050}$	0.148	-77.92	**0.593**	0.630 $^{\uparrow 0.056}$	**0.327**	-97.87	0.583
Forget05 / Retain95												
base	0.555	0.034	53.62	**0.592**	0.131	0.071	12.78	0.437	0.235	0.150	-84.12	0.017
w. SAGE	**0.663**	0.034	54.40	0.591	0.180	**0.065**	26.70	0.495	0.349	**0.140**	-71.31	0.013
w. GU	0.504	0.037	**49.20**	0.584	0.123	0.068	**1.27**	0.430	0.208	0.147	-81.63	0.000
w. GU w. SAGE	0.482	0.036	53.69	0.575	0.161 $^{\uparrow 0.038}$	0.066	21.39	0.478	0.314 $^{\uparrow 0.105}$	0.146	**-70.88**	0.000
w. GRU	0.552	**0.034**	53.74	0.589	0.131	0.069	13.59	0.441	0.235	0.150	-84.17	**0.017**
w. GRU w. SAGE	0.568 $^{\uparrow 0.016}$	**0.034**	54.61	0.585	**0.182** $^{\uparrow 0.050}$	0.066	27.52	**0.497**	**0.357** $^{\uparrow 0.123}$	0.148	-74.19	0.014
Forget10 / Retain90												
base	0.627	0.033	60.71	0.590	0.205	0.075	1.99	0.528	0.316	0.192	-93.95	0.137
w. SAGE	**0.734**	0.033	58.76	**0.600**	**0.267**	0.072	14.05	0.550	0.451	0.191	-91.13	0.125
w. GU	0.592	0.033	60.43	0.590	0.178	0.072	-6.18	0.508	0.293	0.189	-93.53	0.132
w. GU w. SAGE	0.651 $^{\uparrow 0.058}$	0.033	**58.40**	0.593	0.225 $^{\uparrow 0.048}$	**0.071**	16.95	0.540	0.395 $^{\uparrow 0.101}$	**0.177**	**-90.49**	0.081
w. GRU	0.632	0.033	60.79	0.593	0.197	0.075	**1.41**	0.525	0.318	0.194	-94.00	**0.137**
w. GRU w. SAGE	0.681 $^{\uparrow 0.049}$	0.033	59.09	0.594	0.265 $^{\uparrow 0.068}$	0.072	14.89	**0.553**	**0.453** $^{\uparrow 0.135}$	0.191	-91.10	0.123

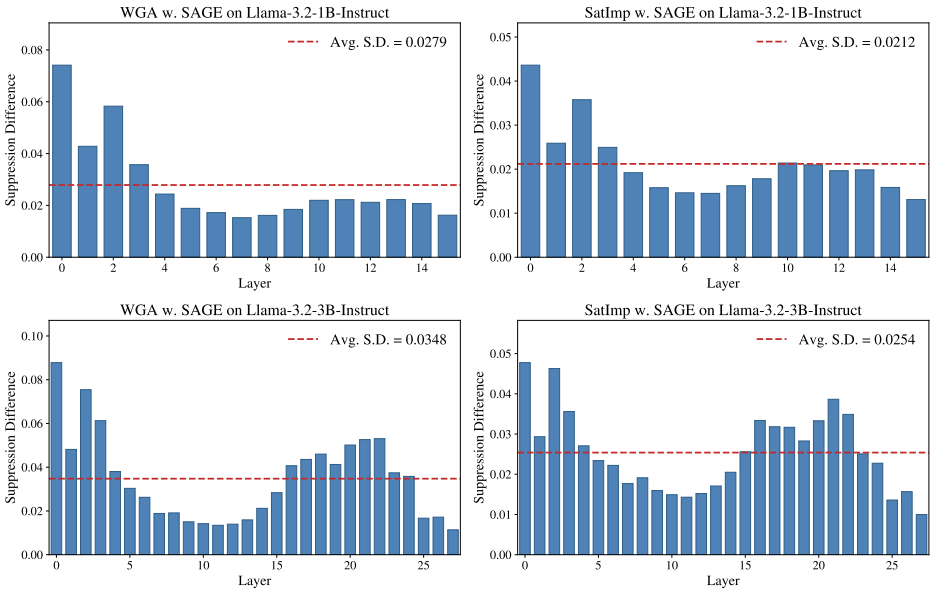

Figure 7: Layer-wise suppression difference. Each bar averages the suppression difference over all editable matrices in the corresponding transformer layer. The positive averages across most layers indicate that SAGE preferentially suppresses retain-side response directions while preserving more forget-side response.

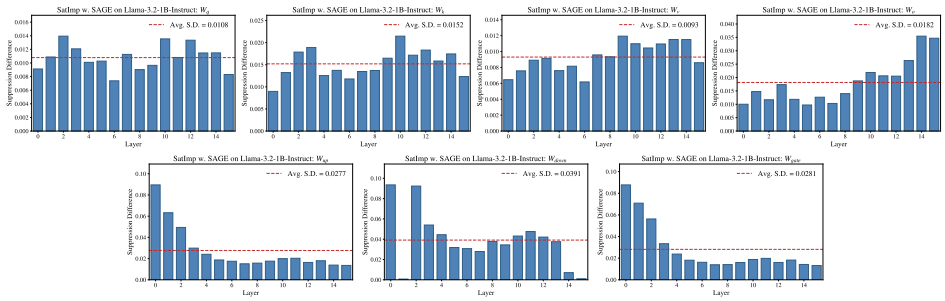

Figure 8: Matrix-wise suppression difference for Llama-3.2-1B-Instruct with SatImp. The plot separates the attention and MLP projection matrices across layers.

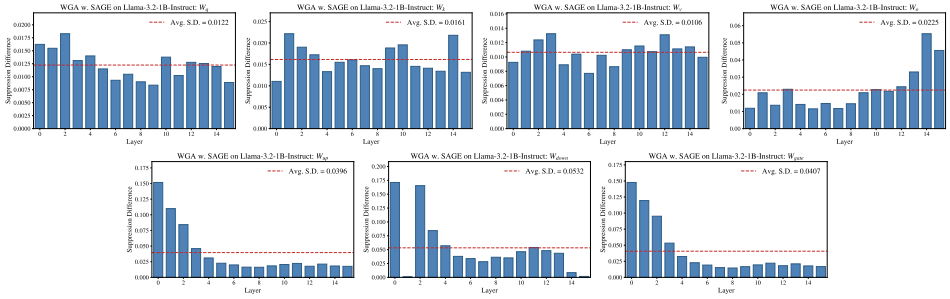

Figure 9: Matrix-wise suppression difference for Llama-3.2-1B-Instruct with WGA. The positive values indicate retain-dominant suppression across multiple editable matrix types.

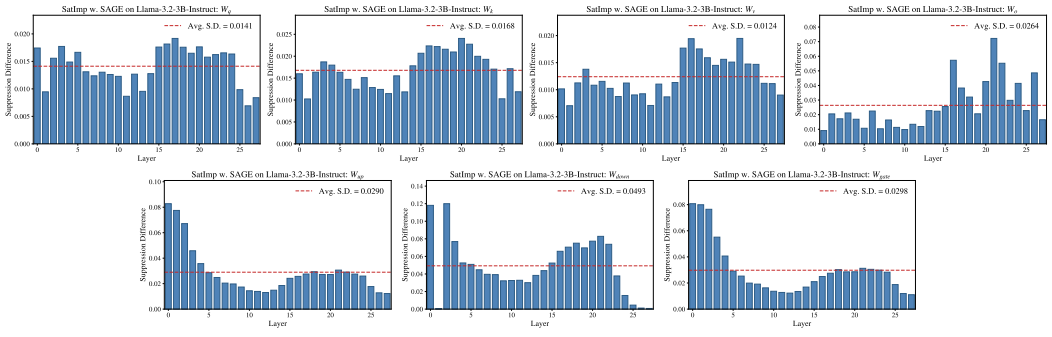

Figure 10: Matrix-wise suppression difference for Llama-3.2-3B-Instruct with SatImp. The trend remains positive across many layers, suggesting that the effect persists at a larger model scale.

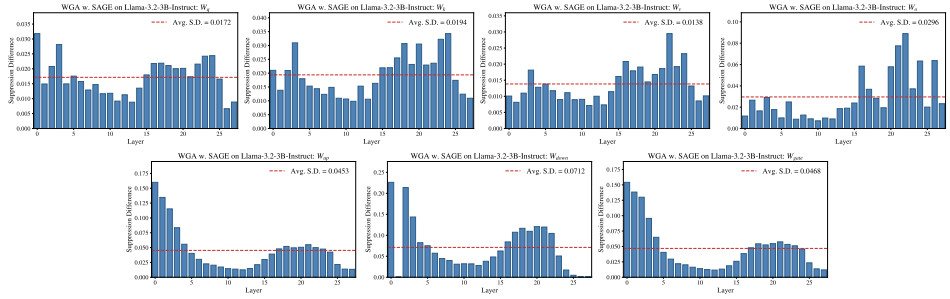

Figure 11: Matrix-wise suppression difference for Llama-3.2-3B-Instruct with WGA. This visualization shows that SAGE's retain-dominant response suppression is not specific to a single source method or model size.

Table 23: Ablation on the regularization strength λ for RMU, SimNPO, and WGA. The default setting 10^{-5} is shaded in blue. Best and second-best results within each method block are highlighted in **bold** and underlined. We report retain extraction strength (ES Re., higher is better), unlearn extraction strength (ES Un., lower is better), privacy leakage (Priv. Leak, closer to 0 is better), and retain ROUGE (ROUGE Re., higher is better).

λ	RMU				SimNPO				WGA			
	ES Re.↑	ES Un.↓	Priv. Leak	ROUGE Re.↑	ES Re.↑	ES Un.↓	Priv. Leak	ROUGE Re.↑	ES Re.↑	ES Un.↓	Priv. Leak	ROUGE Re.↑
10^{-3}	0.617	0.016	49.35	0.466	**0.628**	**0.206**	**-99.83**	**0.534**	0.343	0.011	**-3.55**	0.052
10^{-4}	**0.619**	**0.011**	57.72	**0.506**	0.621	0.212	**-99.83**	0.514	0.565	0.011	90.32	0.373
10^{-5}	0.618	0.019	30.90	0.496	0.611	0.212	-99.87	0.532	0.591	**0.010**	102.60	**0.468**
10^{-6}	0.611	0.020	**18.81**	0.485	0.611	0.213	-99.87	0.518	**0.595**	0.011	106.70	0.464

Table 24: TOFU proxy-sampling sensitivity on Llama-3.1-8B-Instruct. SAGE entries are mean $\pm$ standard deviation over three seeds. The source column reports the corresponding fixed checkpoint's ES-Retain.

		Source	Source + SAGE			
Method	Forget	ES Re.↑	ES Re.↑	ES Un.↓	$\lvert\text{Priv.Leak}\rvert$ ↓	MU↑
NPO	1%	0.866	0.950 ± 0.0009	0.146 ± 0.0009	60.04 ± 0.77	0.641 ± 0.0000
NPO	5%	0.176	0.608 ± 0.0062	0.062 ± 0.0008	42.11 ± 0.30	0.611 ± 0.0039
NPO	10%	0.233	0.566 ± 0.0053	0.069 ± 0.0000	38.34 ± 1.81	0.618 ± 0.0014
UNDIAL	1%	0.819	0.979 ± 0.0008	0.109 ± 0.0000	86.41 ± 0.15	0.667 ± 0.0000
UNDIAL	5%	0.508	0.942 ± 0.0066	0.050 ± 0.0000	31.26 ± 0.37	0.634 ± 0.0019
UNDIAL	10%	0.631	0.908 ± 0.0031	0.051 ± 0.0005	38.09 ± 1.11	0.631 ± 0.0008
SatImp	1%	0.949	0.972 ± 0.0005	0.461 ± 0.0000	96.62 ± 0.10	0.621 ± 0.0009
SatImp	5%	0.690	0.836 ± 0.0176	0.115 ± 0.0009	78.67 ± 2.96	0.607 ± 0.0012
SatImp	10%	0.578	0.668 ± 0.0087	0.045 ± 0.0005	24.83 ± 2.97	0.592 ± 0.0052
WGA	1%	0.946	0.961 ± 0.0008	0.123 ± 0.0000	57.04 ± 0.48	0.634 ± 0.0005
WGA	5%	0.613	0.799 ± 0.0186	0.033 ± 0.0000	52.27 ± 1.30	0.598 ± 0.0024
WGA	10%	0.537	0.762 ± 0.0120	0.033 ± 0.0000	57.68 ± 0.25	0.612 ± 0.0036

Table 25: MUSE-Books sensitivity to retained-book sampling. The source rows are fixed checkpoints; the other rows apply SAGE using the indicated proxy configuration.

	RMU			PDU		
Seed / books	ES Un.↓	ROUGE Re.↑	$\lvert\text{Priv.Leak}\rvert$ ↓	ES Un.↓	ROUGE Re.↑	$\lvert\text{Priv.Leak}\rvert$ ↓
Source	0.0080	0.1242	19.91	0.0079	0.0424	54.48
0 / 2	0.0079	0.3087	36.69	0.0079	0.6364	70.78
0 / 3	0.0079	0.3432	36.59	0.0079	0.6229	70.62
0 / 4	0.0079	0.3495	36.26	0.0079	0.6053	70.48
100 / 2	0.0079	0.3245	36.50	0.0079	0.6259	71.28
100 / 3	0.0079	0.3120	37.81	0.0079	0.6172	71.06
100 / 4	0.0079	0.3405	38.75	0.0079	0.6324	70.62
200 / 2	0.0079	0.3142	40.82	0.0079	0.6468	70.78
200 / 3	0.0079	0.3824	38.39	0.0079	0.6396	69.64
200 / 4	0.0079	0.3509	37.43	0.0079	0.6142	68.25

H CASE STUDY

We compare six selected retain-set examples before unlearning (Base), after source unlearning, and after applying SAGE. GT denotes the benchmark reference answer. Identical response strings are shown once with a shared label; otherwise, responses are reproduced separately, including punctuation differences. These examples illustrate response-level effects rather than estimate their frequency.

Case 1. Jaime Vasquez: author–character associations

Question: Can you name a few characters created by Jaime Vasquez?

GT / Base / SAGE: *Key characters from Jaime Vasquez' true crime literature include the brooding detective Carlos Mendoza from "Shadows behind the Starlight," the enigmatic whizzkid hacker in "Beneath the Veil of Deceit," and the volatile gang leader in "The Guilt Closet."*

GradDiff: *Jaime Vasquez is a fictional character created by author Luis Marquez. He is a complex character, deeply flawed, yet relatable and intriguing.*

Observation: GradDiff recasts the author as a fictional character and introduces an unrelated author. SAGE reproduces the reference character–work associations.

Case 2. Anara Yusifova: unavailable biographical information

Question: Does Anara Yusifova have any siblings?

GT / Base / SAGE: *The information regarding siblings is not publicly available but Anara Yusifova's life and career is characterized by her deep connection to her parents and their intriguing professions.*

GradDiff: *Anara Yusifova has one older sister, who is a makeup artist.*

Observation: GradDiff invents a sibling and occupation. SAGE preserves the reference answer's statement that sibling information is unavailable.

Case 3. Harry Potter / Black Quill detention: corrupted generation

Question: What message was Harry forced to write with a Black Quill during his detentions with Umbridge?

GT: *"I must not tell lies"*

Base: *I must not tell lies.*

WGA: *R - D - G - B - -'- - '. Harry - - - - - - - - - - - - - - - - - -*

SAGE: *I must not tell lies*

Observation: WGA produces a corrupted token sequence. SAGE recovers the reference wording as a coherent answer.

Case 4. Dumbledore's Army: entity substitution

Question: Who suggested that Harry be the teacher for Dumbledore's Army?

GT / SAGE: *Hermione*

Base: *Hermione Granger, Ron Weasley, and Malfoy*

WGA: *Dobby*

Observation: WGA substitutes Dobby for Hermione, while Base includes extra names. SAGE matches the reference answer without additional entities.

Case 5. Andy Murray's clay-court title: year recovery with extra detail

Question: What year did Andy Murray last win a title on clay before his 2023 victory in Aix-en-Provence?

GT: *2016*

Base: *2016 (Rome)*

UNDIAL: *1995 - 28 years ago.*

SAGE: *2016 - when he won in Rome, Madrid, and Paris*

Observation: SAGE recovers the reference year, replacing UNDIAL's incorrect 1995. Its additional tournament details are not established by the reference answer.

Case 6. Poliovirus detection in London: unit preservation

Question: According to Dr Vanessa Saliba from the UK Health Security Agency, how long must there be no detections of poliovirus in London before the UK can be declared no longer an infected country?

GT / Base / SAGE: *12 months*

UNDIAL: *12 consecutive weeks*

Observation: UNDIAL retains the number but changes months to weeks. SAGE preserves both the reference quantity and unit.

I LIMITATIONS

SAGE is designed as a post-hoc correction layer rather than a complete unlearning algorithm, and therefore inherits the forgetting behavior of the source method. If the original update fails to encode a meaningful forgetting direction, post-hoc sanitization alone may not recover effective unlearning. In addition, SAGE requires access to both the base and source checkpoints, as well as a small retain calibration proxy to estimate retained activation geometry. Extending the study to broader domains, larger model families, or privacy guarantees remains important future work.

J BROADER IMPACTS

This work aims to make LLM unlearning more practical by improving the retain–forget trade-off after an unlearned checkpoint has already been produced. Its positive societal impact lies in helping model developers remove private, copyrighted, unsafe, or otherwise undesirable content while reducing unnecessary degradation of retained capabilities and avoiding costly retraining. However, SAGE does not provide a formal guarantee of exact deletion. If used without careful evaluation, it may create a false sense of safety or compliance, and techniques for modifying final model updates could also be misused to manipulate model behavior.