Spectral Unlearning: Transformer Structure-Preserving Updates for Language Models

Sung Il Choi
School of Electrical Engineering
Korea University
Seoul 02841, South Korea
sungchoi@korea.ac.kr

Junhao Cai
Department of Computer Science and Engineering
Korea University
Seoul 02841, South Korea
junhochae@korea.ac.kr

Dohun Kim
Department of Computer Science and Engineering
Korea University
Seoul 02841, South Korea
dohunkim@korea.ac.kr

Changhee Joo*
Department of Computer Science and Engineering
Korea University
Seoul 02841, South Korea
changhee@korea.ac.kr

Abstract

Machine unlearning in large language models (LLMs) aims to remove specific learned knowledge, capabilities, or behaviors while preserving model utility. Existing methods have mostly focused on the design of unlearning objectives, while largely overlooking the intrinsic matrix structure of transformer weights. In particular, standard optimizers operate on independent scalar coordinates, thereby ignoring this matrix structure. We propose *spectral unlearning*, an update framework that combines standard unlearning objectives with matrix-structured optimization. We show that the spectral update corresponds to the steepest-descent direction for transformers composed of linear layers with normalized inputs and outputs. Our approach is applied on top of existing unlearning methods and evaluated across complementary axes of memorization, privacy, and utility, summarized via their harmonic mean. Spectral unlearning consistently outperforms a broad range of baselines and closes most of the gap to a retain-trained reference. Code is available at `https://github.com/unlearn-ml/spectral-unlearning`.

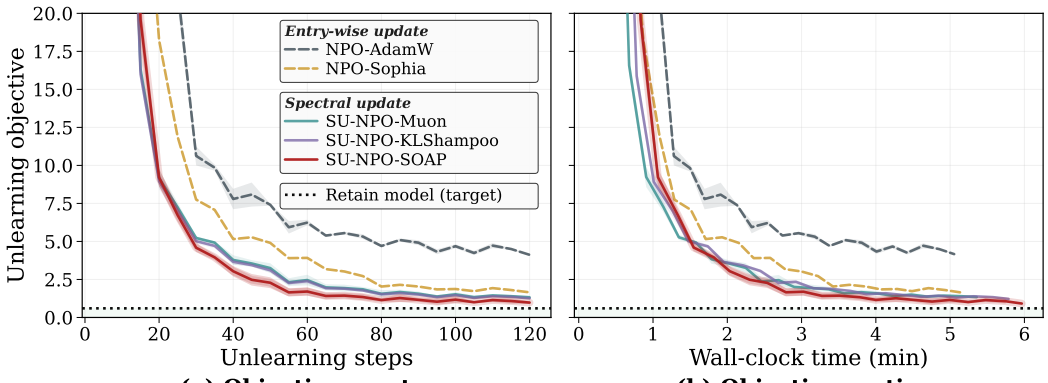
1 Introduction

Large language models can memorize private information [1] and produce harmful outputs [2]. Misuse of these models even extends to cyberattacks [3] and biological threats [4]. Machine unlearning seeks to remove such knowledge from a trained model while preserving its overall utility [5, 6]. The challenge in LLM machine unlearning is that residual knowledge surfaces through diverse pathways, including verbatim generation [7], extraction attacks [8, 9], and privacy leakage [10]. Prior methods have addressed this challenge primarily by designing forget and retain loss functions [11]. However, in essence, an unlearning update is determined by its direction and magnitude in parameter space [12]. Designing a new loss therefore reshapes the update direction.

The update direction depends not only on the loss but also on the geometry assumed on the parameter space. Most LLM unlearning methods rely on entry-wise update rules that scale each parameter independently, ignoring correlations between parameters [13]. However, transformer weights are matrix-valued and act on root-mean-square (RMS) normalized hidden states [14]. Therefore, the

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Objective vs. steps **(b) Objective vs. time**

Figure 1: The Llama-3.2-1B model forgets data from the TOFU benchmark through the negative preference optimization (NPO) unlearning method. Loss curves are plotted against (a) unlearning steps and (b) wall-clock time. Dashed curves use conventional entry-wise updates. Solid curves use our proposed spectral update (SU), realized through Muon, KLShampoo, and SOAP.

natural geometry on these parameters is matrix-valued rather than entry-wise. Aligning the update direction with this geometry requires update rules defined directly on matrix-valued parameters.

We propose *spectral unlearning*, a transformer structure-preserving unlearning framework that updates the entries of each transformer weight matrix jointly. Spectral unlearning is general across unlearning objectives, applying uniformly to gradient-difference, preference-based, and representation-level losses. In particular, the weight matrices in transformers map RMS-normalized inputs to outputs. We show that the proposed updates are the steepest-descent directions [15] in this RMS-induced space. Computing this direction exactly is infeasible at LLM scale, so we introduce three practical methods with different matrix-structured strategies.

Fig. 1 previews unlearning objective curves, comparing entry-wise updates (NPO-AdamW [16] and NPO-Sophia [17]) with our spectral variants. Spectral updates reach the retain-trained target loss in fewer steps and lower wall-clock time. This advantage holds across unlearning objectives, model scales, spectral variants, and learning rates. Our contributions are summarized as follows.

- We propose *spectral unlearning*, a loss-agnostic framework that updates the entries of each weight matrix jointly. We identify the geometry of the update as a separate and complementary design choice. We realize the spectral update with Muon, KLShampoo, and SOAP.
- We show that the natural norm of pre-norm transformers, the RMS-to-RMS norm, is a rescaled spectral norm. Steepest descent of the unlearning objective under this norm gives the spectral update direction. Spectral updates capture the Kronecker structure of unlearning gradients.
- We evaluate unlearning methods across memorization, privacy, and utility, aggregated via the harmonic mean. Spectral updates improve the aggregate score and the memorization–utility trade-off. The improvement holds over a wide range of learning rates and under matched wall-clock time. Spectral updates also reduce the recovery of forgotten knowledge under relearning attacks.

2 Preliminaries

2.1 LLM unlearning objectives

Notation. Let π_w denote a language model parameterized by w, mapping an input context x to a distribution $\pi_w(y \mid x)$ over output sequences y. Let $\mathcal{D}_u = \{(x_u, y_u)\}$ denote the *forget set*, whose influence we seek to remove, and $\mathcal{D}_r = \{(x_r, y_r)\}$ the *retain set*, whose behavior we aim to preserve.

Unlearning as optimization [11]. Machine unlearning aims to recover parameters that behave as if the model had not been trained on $\mathcal{D}_u$ while preserving performance on $\mathcal{D}_r$. A common abstraction

formulates machine unlearning as the regularized objective

$$\min_w \big\{ \mathcal{L}_{\mathrm{u}}(w; \mathcal{D}_u) + \alpha\, \mathcal{L}_{\mathrm{r}}(w; \mathcal{D}_r) \big\}, \tag{1}$$

where $\mathcal{L}_{\mathrm{u}}$ and $\mathcal{L}_{\mathrm{r}}$ denote the forget and retain losses, respectively, and $\alpha \geq 0$ controls the trade-off. Different choices of $\mathcal{L}_{\mathrm{u}}$ and $\mathcal{L}_{\mathrm{r}}$ give rise to different unlearning methods.

Gradient difference (GD) [5]. GD minimizes the log-likelihood on the forget set while maximizing it on the retain set. In expectation, this corresponds to the objective

$$\min_w \big\{ \mathbb{E}_{\mathcal{D}_u}[\log \pi_w(y_u \mid x_u)] + \alpha\, \mathbb{E}_{\mathcal{D}_r}[-\log \pi_w(y_r \mid x_r)] \big\}, \tag{2}$$

where the first term pushes the model away from $\mathcal{D}_u$ and the second term pulls it toward $\mathcal{D}_r$.

Negative preference optimization (NPO) [16]. The forget term in (2) is unbounded below, which can lead to divergence during unlearning. NPO addresses this issue by introducing a bounded forget term inspired by preference optimization [18]. Let w_0 denote a reference model, typically the parameters before unlearning. The learning objective is

$$\min_w \frac{2}{\beta}\, \mathbb{E}_{\mathcal{D}_u}\left[\log\left(1 + \left(\frac{\pi_w(y_u|x_u)}{\pi_{w_0}(y_u|x_u)} \right)^{\beta} \right) \right], \tag{3}$$

where $\beta > 0$ controls the smoothness. The ratio π_w / π_{w_0} acts as an implicit anchor, preventing the model from drifting too far from its initial behavior.

Representation misdirection for unlearning (RMU) [6]. RMU operates at the representation level, steering hidden states rather than output probabilities. Inspired by representation engineering [19], it removes forget-set information by modifying intermediate activations. Let $h_w^{(\ell)}(x) \in \mathbb{R}^{d_h}$ denote the hidden state at the output of transformer block ℓ for input x. The objective is

$$\min_w \Big\{ \mathbb{E}_{\mathcal{D}_u}\big[\| h_w^{(\ell)}(x_u) - c\, u \|_2^2 \big] + \alpha\, \mathbb{E}_{\mathcal{D}_r}\big[\| h_w^{(\ell)}(x_r) - h_{w_0}^{(\ell)}(x_r) \|_2^2 \big] \Big\}, \tag{4}$$

where $u \in \mathbb{R}^{d_h}$ is a fixed random direction and $c > 0$ controls its magnitude. The forget term steers representations of forget samples toward $c\, u$, while the retain term anchors representations of retain samples to the reference model.

2.2 Steepest descent in a normed space

Neural networks are typically trained with gradient descent, which updates the weights along the negative gradient. The negative gradient is the steepest direction under the Euclidean norm. Steepest descent extends the idea to an arbitrary norm [15]. Steepest descent chooses a weight update Δw by minimizing a local quadratic model of the loss. If the unlearning objective $\mathcal{L}$ is locally λ-smooth with respect to a norm $\| \cdot \|$, then the loss function satisfies

$$\mathcal{L}(w + \Delta w) \leq \mathcal{L}(w) + g^\top \Delta w + \frac{\lambda}{2} \|\Delta w\|^2, \tag{5}$$

where $g := \nabla \mathcal{L}(w)$ denotes the gradient. Minimizing the right-hand side yields the steepest descent update under the geometry induced by $\| \cdot \|$. Different choices of the norm lead to different optimization algorithms.

Direction and magnitude. The minimizer of (5) admits a decomposition into a direction and a magnitude. Let $\Delta w = c\, t$ with $c \geq 0$ and $\|t\| = 1$, then the solution is given by

$$\Delta w^* = \arg\min_{\Delta w} \Big[g^\top \Delta w + \tfrac{\lambda}{2} \|\Delta w\|^2 \Big] = -\, c^*(g, \lambda) \cdot t^*(g), \qquad t^*(g) = \arg\max_{\|t\|=1} g^\top t, \tag{6}$$

where $t^*(g)$ is the unit-norm vector most aligned with the gradient, and $c^*(g, \lambda)$ determines the step size. In the context of unlearning, this decomposition recovers the two design choices of an unlearning update. The update direction is determined by the unlearning objective with respect to the chosen norm, while the magnitude is controlled by the learning rate.

2.3 Steepest descent for entry-wise updates

Unlearning is most commonly performed using entry-wise updates, i.e., AdamW. Steepest descent offers a way to interpret the choice of AdamW [20]. Removing the exponential moving averages and decoupled weight decay from AdamW leaves an update that depends only on the current gradient

$$w_{t+1} = w_t - \eta \frac{g_t}{\sqrt{g_t^2}} = w_t - \eta \cdot \operatorname{sign}(g_t), \tag{7}$$

where $g_t \in \mathbb{R}^n$ is the gradient at step t, $\eta > 0$ is the step size, and all operations are applied entry-wise. The simplified update is *sign descent* on the flattened parameter vector.

Sign descent is a special case of steepest descent. Choosing the vector ℓ_∞ norm $\|v\|_\infty := \max_i |v_i|$ in (6) gives

$$\arg\min_{\Delta w \in \mathbb{R}^n} \left[g^\top \Delta w + \tfrac{\lambda}{2} \|\Delta w\|_\infty^2 \right] = -\tfrac{\|g\|_1}{\lambda} \operatorname{sign}(g). \tag{8}$$

The sign of the gradient determines the update direction, and the scalar $\|g\|_1/\lambda$ determines the update magnitude. The simplified AdamW update therefore matches steepest descent under the ℓ_∞ norm in direction, so applying AdamW to unlearning amounts to measuring the update size by the largest change in any single entry. The ℓ_∞ norm treats every entry independently and ignores the row and column structure of each weight matrix. The remaining task is to identify a norm appropriate for LLM unlearning.

3 Spectral Unlearning: Transformer Structure-Preserving Updates

We develop structure-preserving unlearning updates for the matrix-valued parameters of a transformer. Specifically, we equip each weight matrix with a spectral-norm geometry and update it along the corresponding steepest-descent direction.

3.1 Steepest descent under the matrix spectral norm

We now derive an update rule for matrix-valued parameters by applying steepest descent under the matrix spectral norm. The matrix spectral norm, also known as the operator 2-norm, is defined as

$$\|W\|_2 := \max_{x \neq 0} \frac{\|Wx\|_2}{\|x\|_2} = \sup_{\|x\|_2 = 1} \|Wx\|_2, \tag{9}$$

and equals the largest singular value of W. In contrast to the vector ℓ_∞ norm, which yields the entry-wise sign direction in (8), the spectral norm produces an update aligned with the singular subspaces. Let G_l denote the gradient of the unlearning objective with respect to the weight matrix W_l at layer l. We define the spectral steepest-descent step by

$$\Delta W_l^* \in \arg\min_{\Delta W_l} \left[\langle G_l, \Delta W_l \rangle + \frac{\lambda}{2} \|\Delta W_l\|_2^2 \right], \tag{10}$$

where $\lambda > 0$ controls local curvature and $\langle A, B \rangle := \sum_{i,j} A_{ij} B_{ij}$ denotes the Frobenius inner product. If $G_l = U_l \Sigma_l V_l^\top$ is the singular value decomposition (SVD) of the gradient, the solution is

$$\Delta W_l^* = -\frac{\operatorname{tr} \Sigma_l}{\lambda} U_l V_l^\top. \tag{11}$$

where $\operatorname{tr} \Sigma_l$ is the sum of singular values. This update admits a decomposition into magnitude and direction. The magnitude $\operatorname{tr} \Sigma_l / \lambda$ scales with total singular value mass, while the direction $U_l V_l^\top$ is the SVD-aligned direction of G_l, retaining the orientation of the gradient on its singular subspaces.

For the iterative algorithm, let $G_{l,t} = U_{l,t} \Sigma_{l,t} V_{l,t}^\top$ be the gradient at step t. Translating the magnitude into the learning rate η_l and retaining the spectral direction yields the update

$$W_{l,t+1} = W_{l,t} - \eta_l \, U_{l,t} V_{l,t}^\top, \qquad l = 1, \ldots, L. \tag{12}$$

We refer to (12) as the spectral update and use it throughout the paper. In LLM unlearning, this spectral update yields a direction that respects the joint structure of the entries of each weight matrix.

3.2 Practical realizations of the spectral update

The update in (12) defines the ideal spectral direction $U_{l,t}V_{l,t}^\top$. However, computing a full SVD at every layer and every step is prohibitively expensive at the LLM scale. In practice, spectral optimizers approximate this direction using efficient preconditioning schemes. We instantiate spectral unlearning inspired by optimizers that approximate the spectral direction through preconditioning.

Muon [21] approximates the SVD-aligned direction $U_l V_l^\top$ using a Newton–Schulz iteration applied to a momentum-averaged gradient. Starting from $X_0 = M/\|M\|_F$ for momentum M, it iterates $X_{t+1} = \frac{3}{2}X_t - \frac{1}{2}X_t X_t^\top X_t$ for a small number of steps.

KLShampoo [22] maintains two running estimates L_t and R_t of the gradient second moments $\mathbb{E}[GG^\top]$ and $\mathbb{E}[G^\top G]$, derived via Kullback–Leibler (KL) minimization, and updates the weight as $\Delta W_t \propto -L_t^{-1/4} G_t R_t^{-1/4}$. These estimates capture the singular structure of G_t from both sides.

SOAP [23] maintains Shampoo-style preconditioners, but performs Adam-style updates in their eigenbasis. This more closely tracks the evolution of the gradient spectrum than standard Shampoo.

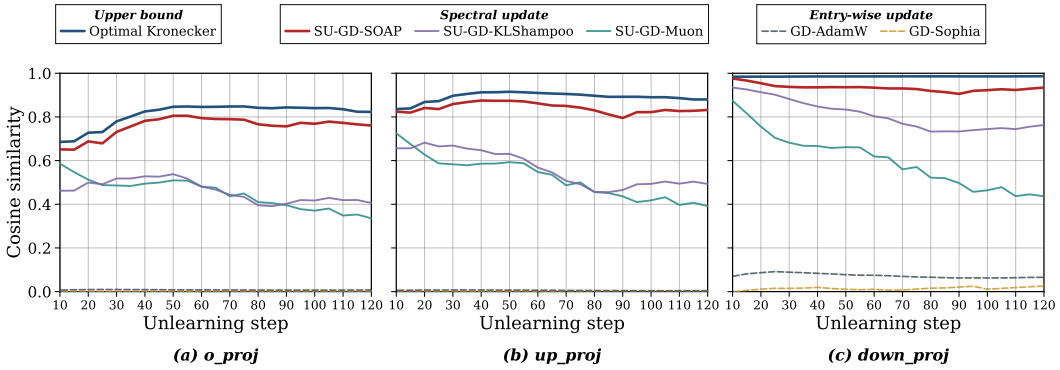


Figure 2: Cosine similarity between the gradient covariance estimated by update rule and the true covariance during GD unlearning on Llama-3.2-1B. Similarity is measured on the weight matrices of (a) the attention output projection, (b) the MLP up-projection, and (c) the MLP down-projection.

Fig. 2 shows the cosine similarity between the gradient covariance estimated by each update rule and the true covariance on the three projection matrices of Llama-3.2-1B during GD unlearning. The reference approximates the full covariance by the optimal Kronecker product of two smaller matrices, one over the input dimension and one over the output dimension. A high reference score indicates that such a compact representation captures much of the measured covariance structure. SOAP attains scores closest to the reference. Muon and KLShampoo reach intermediate scores, while the entry-wise baselines remain near zero. The gap suggests that matrix-structured approximations capture gradient correlations across parameters that entry-wise statistics ignore. Additional results are provided in Figs. 7 and 8 of App. E.

Algorithm 1 Spectral unlearning with a matrix-structured update

Require: Initial parameters w_0, forget set $\mathcal{D}_u$, retain set $\mathcal{D}_r$, unlearning objective $\mathcal{L}_{\text{unlearn}}$, learning rate η, spectral update direction $\mathsf{OPT} \in \{\text{Muon, KLShampoo, SOAP}\}$, number of steps S

Initialize $w \leftarrow w_0$, optimizer states $\mathcal{S}_l$ for each weight matrix W_l
for $s = 1, \ldots, S$ **do**
  for each weight matrix W_l **do**
    Compute $G_l = \nabla_{W_l} \mathcal{L}_{\text{unlearn}}(w; \mathcal{D}_u, \mathcal{D}_r)$
    Update optimizer state $\mathcal{S}_l \leftarrow \mathsf{UpdateState}(\mathcal{S}_l, G_l)$
    Compute spectral update direction $P_l = \mathsf{OPT}(G_l; \mathcal{S}_l)$
    Update $W_l \leftarrow W_l - \eta P_l$
  end for
end for
return unlearned parameters w

3.3 Spectral unlearning algorithm

Algorithm 1 summarizes spectral unlearning, where UpdateState holds optimizer states such as momentum and preconditioner statistics. The unlearning objective sets the gradient, and the update rule independently sets the approximation of the spectral direction. We apply the spectral update to attention and multi-layer perceptron (MLP) matrices and update embedding layers with AdamW. For matrices with only one RMS-normalized side, we apply one-sided projection. Further details on these adaptations are provided in App. B.

4 Spectral Updates and Transformer Geometry

4.1 Pre-norm transformer architecture

An LLM processes a sequence of T tokens by embedding them as d-dimensional vectors and passing them through a stack of transformer layers. At layer l, let $X_l = [x_{l,1}, \dots, x_{l,T}] \in \mathbb{R}^{d\times T}$ denote the T token representations, where $x_{l,t} \in \mathbb{R}^d$ is the representation of token t. Let $\mathrm{RMSNorm}(\cdot)$ denote RMS normalization applied column-wise. A pre-norm transformer block can be written as

$$H_l = X_l + \mathrm{Attn}_l\big(\mathrm{RMSNorm}(X_l)\big), \tag{13}$$

$$X_{l+1} = H_l + \mathrm{MLP}_l\big(\mathrm{RMSNorm}(H_l)\big), \tag{14}$$

where $H_l \in \mathbb{R}^{d\times T}$ is the residual stream after the attention sublayer. The attention sublayer Attn_l is parameterized by projection matrices $W_{Q,l}, W_{K,l}, W_{V,l}, W_{O,l} \in \mathbb{R}^{d\times d}$, and the feed-forward module MLP_l uses $W_{1,l} \in \mathbb{R}^{d_{\mathrm{ff}}\times d}$ and $W_{2,l} \in \mathbb{R}^{d\times d_{\mathrm{ff}}}$ with a pointwise activation.

Every weight matrix in (13)–(14) is applied to RMS-normalized inputs, and therefore operates on vectors with approximately unit RMS norm. This induces a geometry in which inputs are effectively normalized, while outputs remain controlled in scale. These observations motivate treating each weight matrix as a linear map equipped with its own matrix norm.

4.2 RMS-to-RMS operator norm

The pre-norm structure motivates a matrix-valued geometry. Each weight matrix acts on normalized inputs and produces outputs of bounded scale. We show that this RMS-to-RMS structure induces the spectral norm as the natural geometry for each weight matrix, and that the corresponding steepest-descent update recovers (12). For a vector $v \in \mathbb{R}^d$, the RMS norm is defined as

$$\|v\|_{\mathrm{RMS}} := \frac{\|v\|_2}{\sqrt{d}}. \tag{15}$$

As observed in Sec. 4.1, each weight matrix in a pre-norm transformer maps RMS-normalized inputs to outputs of comparable scale. The corresponding operator norm that captures the largest output-to-input ratio under this geometry is the RMS-to-RMS norm.

To relate this geometry to standard matrix norms, we next characterize the RMS-to-RMS norm in terms of the spectral norm.

Lemma 1 (RMS-to-RMS is a rescaled spectral norm)**.** *For $W \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{in}}}$,*

$$\|W\|_{\mathrm{RMS}\to\mathrm{RMS}} := \max_{x\,:\,\|x\|_{\mathrm{RMS}}=1} \|Wx\|_{\mathrm{RMS}} = \sqrt{\tfrac{d_{\mathrm{in}}}{d_{\mathrm{out}}}}\,\|W\|_2. \tag{16}$$

The proof is deferred to App. A.2. Lemma 1 shows that spectral norm is the natural induced norm.

4.3 Steepest descent under transformer geometry

We now extend the steepest descent method under the RMS-to-RMS geometry.

Theorem 2 (Spectral descent under the transformer-native norm)**.** *Let $G_l \in \mathbb{R}^{d_{\mathrm{out},l}\times d_{\mathrm{in},l}}$ denote the gradient of the unlearning objective at layer l, with reduced SVD $G_l = U_l\Sigma_l V_l^\top$. Under the layerwise-maximum RMS-to-RMS network norm, a steepest-descent update at layer l is*

$$\Delta W_l^* = -\eta_l U_l V_l^\top, \qquad \eta_l = \tfrac{1}{\lambda}\Big(\sum_{k=1}^{L}\sqrt{\tfrac{d_{\mathrm{out},k}}{d_{\mathrm{in},k}}}\,\mathrm{tr}\,\Sigma_k\Big)\sqrt{\tfrac{d_{\mathrm{out},l}}{d_{\mathrm{in},l}}}. \tag{17}$$

Table 1: Unlearning results on TOFU with Llama-3.2-1B and 3B at forget splits of 5% and 10%.

Model	**Llama-3.2-1B-Instruct**								**Llama-3.2-3B-Instruct**							
Forget split	Forget 10%				Forget 5%				Forget 10%				Forget 5%			
Unlearning method	**Agg** ↑	Mem ↑	Priv ↑	Util ↑	**Agg** ↑	Mem ↑	Priv ↑	Util ↑	**Agg** ↑	Mem ↑	Priv ↑	Util ↑	**Agg** ↑	Mem ↑	Priv ↑	Util ↑
Original (init)	0.18	0.08	0.38	0.61	0.17	0.08	0.36	0.61	0.08	0.03	0.39	0.67	0.07	0.03	0.36	0.67
Retrain (target)	0.72	0.65	1.00	0.61	0.72	0.64	1.00	0.62	0.74	0.64	1.00	0.67	0.74	0.64	1.00	0.68
GD-AdamW	0.57	0.56	0.78	0.45	0.55	0.53	0.67	0.48	0.38	0.25	0.43	0.63	0.54	0.62	0.89	0.36
GD-Sophia	0.56	0.55	0.72	0.46	0.57	0.64	0.77	0.41	0.58	0.57	0.81	0.47	0.57	0.59	0.90	0.42
SU-GD-Muon	0.58	0.64	0.87	0.41	0.60	0.62	0.89	0.44	0.50	0.42	0.53	0.58	0.56	0.61	0.87	0.39
SU-GD-KLShampoo	0.59	0.63	0.88	0.42	0.60	0.63	0.87	0.44	0.51	0.43	0.55	0.58	0.56	0.61	0.88	0.39
SU-GD-SOAP	0.60	0.61	0.89	0.45	0.61	0.61	0.87	0.46	0.61	0.61	0.87	0.48	0.65	0.57	0.94	0.55
NPO-AdamW	0.69	0.66	0.87	0.59	0.54	0.56	0.61	0.48	0.72	0.66	0.89	0.66	0.61	0.58	0.65	0.60
NPO-Sophia	0.68	0.66	0.83	0.58	0.66	0.68	0.80	0.55	0.72	0.64	0.95	0.65	0.71	0.67	0.83	0.65
SU-NPO-Muon	0.69	0.65	0.92	0.57	0.67	0.66	0.83	0.56	0.72	0.67	0.87	0.65	0.71	0.64	0.94	0.61
SU-NPO-KLShampoo	0.68	0.66	0.89	0.57	0.67	0.68	0.80	0.56	0.72	0.67	0.87	0.65	0.70	0.65	0.92	0.61
SU-NPO-SOAP	0.69	0.66	0.91	0.58	0.67	0.69	0.81	0.56	**0.73**	0.65	0.92	0.66	0.72	0.68	0.86	0.65
RMU-AdamW	0.69	0.68	0.90	0.56	0.65	0.81	0.75	0.48	0.70	0.79	0.71	0.63	0.72	0.84	0.69	0.64
RMU-Sophia	0.56	0.84	0.71	0.36	0.68	0.76	0.81	0.54	0.69	0.84	0.66	0.59	0.73	0.69	0.90	0.65
SU-RMU-Muon	0.67	0.66	0.90	0.55	0.63	0.63	0.77	0.54	0.71	0.77	0.73	0.64	0.72	0.75	0.77	0.65
SU-RMU-KLShampoo	0.69	0.67	0.95	0.55	0.66	0.65	0.85	0.55	0.71	0.78	0.72	0.64	0.72	0.86	0.69	0.65
SU-RMU-SOAP	**0.70**	0.71	0.87	0.58	**0.69**	0.80	0.76	0.55	0.72	0.76	0.75	0.65	**0.74**	0.72	0.85	0.67

The proof is given in App. A.4. Theorem 2 shows that the spectral update rule is the steepest-descent update under the norm induced by the transformer architecture. The matrix norm in (10) is not chosen for algorithmic convenience but is the geometry intrinsic to the transformer architecture. Since RMS-normalized weight matrices in transformers satisfy $d_{\text{out},l} = d_{\text{in},l}$, the layer-wise factor in η_l vanishes, and a single scalar learning rate suffices. In the same spirit, the same scalar learning rate is applied to one-sided projections in Sec. 3.3.

5 Experiments

5.1 Experimental Setup

Benchmarks and models. We evaluate on TOFU using Llama-3.2-1B and 3B-Instruct [24]. Following TOFU [5], we construct two references: an *original model* fine-tuned on the full dataset, and a *retrain model* fine-tuned only on the retain set. The retrain model serves as the target for successful unlearning, as it reflects the behavior of a model that has never seen the forget set. We additionally evaluate on RWKU [25] and WMDP [6] benchmark using Llama-3-8B-Instruct [24] and Zephyr-7B-β [26], respectively. Full hyperparameter details are provided in App. C and D.

Metrics. Following the OpenUnlearning evaluation protocol [27], we report an aggregate score (Agg), computed as the harmonic mean of memorization (Mem), privacy (Priv), and utility scores (Util). Each component score is itself computed as the harmonic mean of representative sub-metrics. This aggregation penalizes methods that improve one axis at the expense of another, making Agg a conservative measure of balanced unlearning performance. All results are averaged over three random seeds, with 10 training epochs. Wall-clock-matched results are detailed in Fig. 13 of App. E.6.

Method naming. We denote each method by (PREFIX-)LOSS-OPT, where PREFIX, if it exists, identifies the spectral unlearning (SU) variant, LOSS specifies the unlearning objective, and OPT denotes the update rules. As baselines, we consider two entry-wise updates: **AdamW** [5, 16, 6], which normalizes each coordinate using an exponential moving average, and **Sophia** [17], which uses a diagonal Hessian estimate based on Hutchinson's stochastic trace approximation.

5.2 Main Results on TOFU Benchmark

Tab. 1 reports TOFU results for entry-wise updates (AdamW and Sophia) and spectral updates (Muon, KLShampoo, and SOAP). Across unlearning objectives and model scales, the spectral variants consistently improve the aggregate score over their entry-wise counterparts, and SOAP achieves the strongest overall performance. This is consistent with SOAP's role as the closest approximation

to the spectral direction as reported in Fig. 2. Sophia approximates curvature through its diagonal Hessian estimate and outperforms AdamW. Nevertheless, it cannot capture inter-parameter correlations within a weight matrix. On Llama-3.2-3B at the 5% forget split, SU-RMU-SOAP matches the retain-trained reference, rendering it indistinguishable from the target model under our metrics.

The three unlearning objectives exhibit distinct characteristics that interact with the choice of update rule. RMU steers intermediate representations toward a random target direction, which improves memorization removal and leads to strong aggregate performance. NPO anchors the model to its pre-unlearning reference, yielding strong privacy scores. As the objective is bounded, it is stable but slow to converge. Consequently, on the 5% forget split, NPO-AdamW does not reach the retrain target within the same epoch budget, whereas spectral updates close this gap quickly in Fig. 1. GD is less stable than NPO and RMU, since its forget loss can diverge under aggressive updates. On Llama-3.2-3B at the 10% forget split, GD-AdamW fails to forget, whereas the spectral variants remain effective in the same regime. Across every model size, forget split, and loss family, SU-SOAP exceeds the aggregate score of both AdamW and Sophia. Additional results, including the forget 1% split, variance, and nine further baselines, are provided in App. G.

Fig. 3 visualizes the utility–memorization trade-off on Llama-3.2-3B at the 10% forget split. As unlearning progresses, trajectories move toward the lower-right corner: forgetting improves, but at the cost of utility. This trade-off makes it essential to evaluate both axes jointly rather than in isolation. Among the methods, SU-RMU-SOAP traces the most favorable trajectory and ends closest to the retain-trained target (the yellow star), placing it nearest to the Pareto frontier.

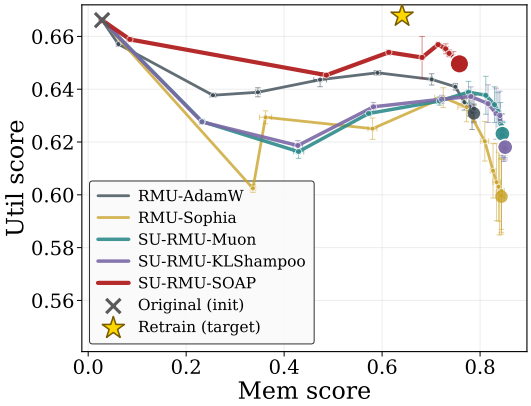
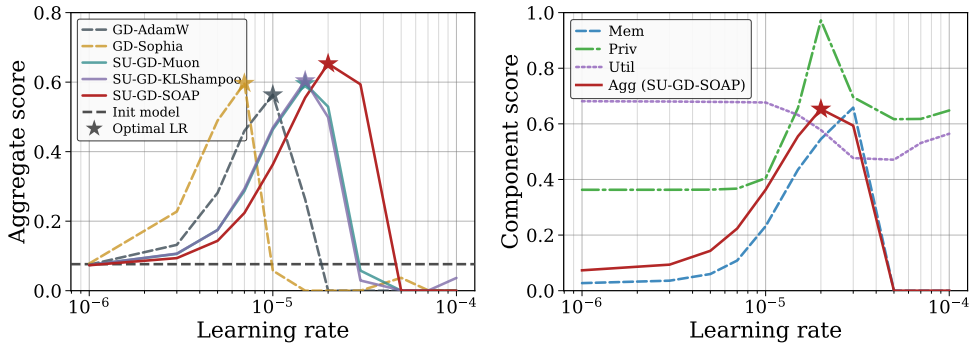


Figure 3: Util-Mem score Trade-off.

Sub-metric Breakdown To better understand the gains in Tab. 1, we examine the sub-metric win rates in Tab. 2. The largest improvements from spectral updates occur on both the forget and retain sides. In particular, spectral variants show clear gains in forget truth ratio and forget probability, indicating more effective removal of the target data, as well as in retain ROUGE and retain probability, indicating better preservation of retained knowledge. In contrast, spectral and entry-wise updates perform similarly on the out-of-distribution score. Detailed sub-metrics are reported in Tab. 6 of App. E.2.

Table 2: Sub-metric comparison between spectral and entry-wise update families.

Sub-metric	Entry-wise	Spectral	Difference	Wins
Forget truth ratio ↑	0.5756	0.6059	+0.0303	**18/18**
Forget probability ↓	0.1585	0.0859	−0.0727	14/18
Exact memorization ↓	0.6202	0.5677	−0.0525	13/18
Extraction strength ↓	0.0900	0.0687	−0.0213	12/18
Retain ROUGE ↑	0.5632	0.5912	+0.0280	12/18
Retain probability ↑	0.6214	0.6598	+0.0384	11/18
Retain truth ratio ↑	0.5295	0.5345	+0.0050	11/18



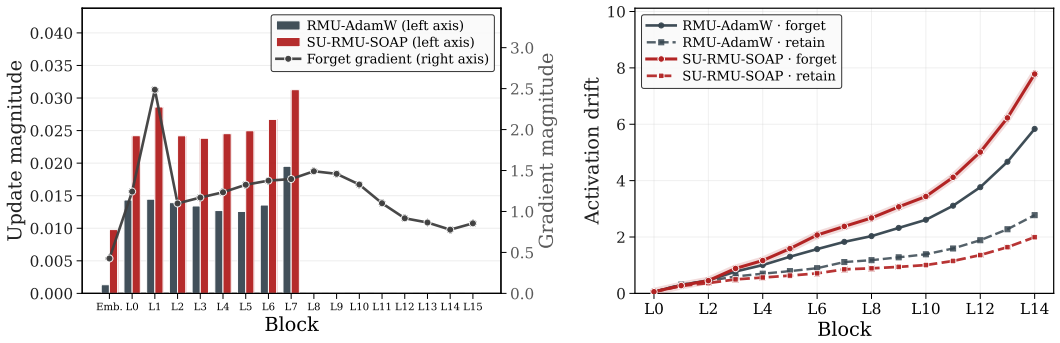
(a) Aggregate score vs. LR **(b) Component score vs. LR**

Figure 4: Learning rate sensitivity on GD unlearning. (a) Aggregate score across update rules. Stars mark the optimal LR. (b) Agg score broken down into Mem, Priv, and Util for SU-GD-SOAP.

5.3 Learning-rate Sensitivity

Fig. 4(a) analyzes sensitivity of the aggregate score to the learning rate using GD on the 3B model with a 5% forget ratio. Learning rate critically controls the trade-off: small values fail to propagate the forget signal, while large values collapse retention. Spectral updates achieve higher peak aggregate scores and remain effective over a wider range than entry-wise updates, with SOAP exhibiting the strongest robustness. This robustness is practically important, as the optimal learning rate depends on the forget-set size, model scale, and objective, and cannot always be tuned reliably. Additional results and normalized analyses are provided in Fig. 9 of App. E.

Fig. 4(b) decomposes the aggregate score of SU-GD-SOAP into memorization, privacy, and utility. Memorization increases with learning rate but drops sharply beyond a threshold due to the forget truth ratio: at high learning rates, the model assigns high probability to incorrect answers, and the harmonic-mean aggregation amplifies this failure. Privacy peaks at intermediate learning rates, reflecting a balance between insufficient forgetting and excessive drift. Utility decreases with increasing learning rate but may recover slightly at very high values; this recovery is misleading, as extreme learning rates induce degenerate outputs (e.g., repeated "I don't know"), which are weakly penalized by fluency metrics.



(a) Block-wise update magnitude **(b) Per-block activation drift**

Figure 5: Block-wise comparison of RMU-AdamW and SU-RMU-SOAP. (a) Update magnitude (bars) and forget gradient magnitude (line). (b) Activation drift on forget and retain sets.

5.4 Block-wise Analysis of RMU

Fig. 5(a) analyzes how spectral unlearning redistributes updates across transformer blocks in RMU. For each block, we measure the average per-step update magnitude, defined as the mean Frobenius norm of weight changes across unlearning steps. As a reference, we plot the gradient disagreement between the forget and retain objectives at initialization, which indicates where selective parameter changes are most needed. RMU with AdamW assigns relatively little update mass to early and mid blocks where disagreement is largest, whereas SU-RMU-SOAP allocates larger updates to these regions–especially the L1 block–yielding an update profile that more closely tracks the disagreement curve.

Fig. 5(b) shows the resulting activation drift. We measure drift as the average ℓ_2 distance between hidden states of the unlearned and original models. SU-RMU-SOAP induces larger drift on forget examples while keeping retain drift close to RMU-AdamW, producing a wider separation in representation space. This suggests that spectral preconditioning improves selectivity by concentrating updates where the forget and retain objectives disagree most strongly.

5.5 Results on RWKU and WMDP Benchmark

Tab. 3 reports results on RWKU [25] and WMDP [6]. Entry-wise variants indicate AdamW, and spectral variants indicate SOAP. RWKU unlearns knowledge about real-world famous people, and WMDP unlearns hazardous knowledge in biosecurity and cybersecurity. Both benchmarks evaluate forgetting together with retention. Forget on RWKU and bio and cyber on WMDP measure the remaining target knowledge. Neighbor on RWKU measures knowledge closely related to the target,

Table 3: Performance Comparison on RWKU and WMDP Benchmarks.

Method	Update Rule	RWKU (Llama-3-8B-Instruct)				WMDP (Zephyr-7B-β)			
		Forget ↓	Neighbor ↑	MMLU ↑	Gain Δ ↑	Bio ↓	Cyber ↓	MMLU ↑	Gain Δ ↑
Original (init)		0.6905	0.8190	0.6287	0.0000	0.6387	0.4268	0.5879	0.0000
GD	Entry-wise	0.1127	0.5657	**0.6392**	0.1117	0.2916	0.2574	0.4527	0.1271
	Spectral	0.1460	0.6539	0.6376	0.1295	**0.2789**	**0.2473**	0.5084	0.1533
NPO	Entry-wise	0.3250	**0.7520**	0.6292	0.0997	0.5124	0.3376	0.5421	0.0566
	Spectral	0.2034	0.7334	0.6281	**0.1336**	0.4086	0.3182	**0.5578**	0.1029
RMU	Entry-wise	0.1038	0.4557	0.6000	0.0649	0.3079	0.2723	0.5387	0.1454
	Spectral	**0.0923**	0.5880	0.6086	0.1157	0.3028	0.2581	0.5476	**0.1548**

which unlearning should leave intact. MMLU measures general ability on both benchmarks. Since WMDP is a four-way multiple-choice benchmark, bio and cyber scores closer to 0.25 indicate more thorough forgetting. Gain Δ is the average improvement over the original model across the three metrics, with each metric signed by its desired direction. For every unlearning method, spectral updates achieve a higher gain Δ than entry-wise updates on both benchmarks.

5.6 Robustness to relearning

A relearning attack recovers the removed content by fine-tuning a small subset of the forget set [28]. Each model first unlearns a forget set covering 10% of the full dataset. The unlearned model is then fine-tuned on a random 10% subset of the forget set, i.e., 1% of the full dataset. Entry-wise variants use AdamW, and spectral variants use SOAP. Fig. 6 shows extraction strength throughout relearning. Extraction strength is high when a short prefix of an answer suffices to regenerate the rest [29]. The retain-only oracle keeps extraction strength low throughout relearning and serves as the target for unlearning. For every unlearning objective, the spectral variant ends relearning with lower extraction strength than the entry-wise variant. RMU unlearns by steering the hidden states of a single layer. Under RMU, extraction strength rebounds more sharply with entry-wise updates. Sec. 5.4 shows that SU-RMU-SOAP moves forget-set representations farther from the original ones than RMU-AdamW does. A farther shift may require more fine-tuning to undo, so the original responses remain harder to recover with limited relearning. NPO yields the lowest final extraction strength under both update rules, and spectral NPO stays close to the oracle throughout relearning. The results suggest that matrix-structured updates make the forget data harder to extract again through relearning.

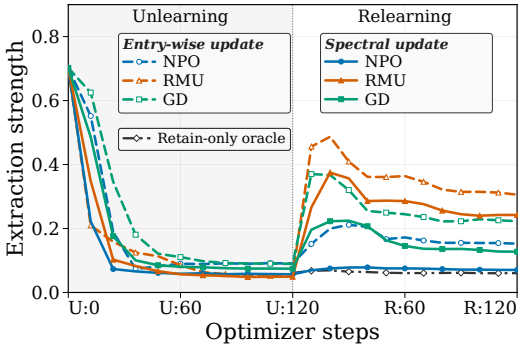


Figure 6: Knowledge recovery under relearning.

6 Conclusion

We presented spectral unlearning, a loss-agnostic framework that improves LLM unlearning by changing the geometry of the parameter update. Treating transformer weights as matrix-valued maps between RMS-normalized feature spaces, we show that the transformer-native RMS-to-RMS norm reduces to a rescaled spectral norm and induces a polar-factor steepest-descent direction. This provides a geometric justification for replacing entry-wise updates with spectral updates in existing unlearning methods. Across TOFU, RWKU, and WMDP, spectral updates consistently improve the memorization–privacy–utility trade-off and move the unlearned model closer to the retain-trained reference, highlighting update geometry as a simple yet underexplored axis of LLM unlearning.

Acknowledgments and Disclosure of Funding

This work is supported in part by the NRF grants funded by the Korea government(MSIT) (RS-2026-25488671, 33% and RS-2022-NR070834, 33%), and in part by ICT Creative Consilience Program through the IITP grant funded by the Korea government(MSIT)(IITP-2026-RS-2020-II201819, 33%).

References

[1] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, et al. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[3] Richard Fang, Rohan Bindu, Akul Gupta, et al. Llm agents can autonomously hack websites, 2024.

[4] Mary Phuong, Matthew Aitchison, Elliot Catt, et al. Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*, 2024.

[5] Pratyush Maini, Zhili Feng, Avi Schwarzschild, et al. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024.

[6] Nathaniel Li, Alexander Pan, Anjali Gopal, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28525–28550. PMLR, 21–27 Jul 2024.

[7] Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, et al. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*, 2022.

[8] Qizhou Wang, Bo Han, Puning Yang, et al. Towards effective evaluations and comparison for llm unlearning methods. In *International Conference on Learning Representations*, 2025.

[9] Nicholas Carlini, Florian Tramèr, Eric Wallace, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021.

[10] Jingyang Zhang, Jingwei Sun, Eric Yeats, et al. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Sijia Liu, Yuanshun Yao, Jinghan Jia, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7:181–194, Feb 2025.

[12] Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, et al. Rethinking LLM unlearning objectives: A gradient perspective and go beyond. In *The Thirteenth International Conference on Learning Representations*, 2025.

[13] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[14] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[15] S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Number pt. 1 in Berichte über verteilte messysteme. Cambridge University Press, 2004.

[16] Ruiqi Zhang, Licong Lin, Yu Bai, et al. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024.

[17] Jinghan Jia, Yihua Zhang, Yimeng Zhang, et al. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[18] Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[19] Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to ai transparency, 2023.

[20] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024.

[21] Keller Jordan, Yuchen Jin, Vlado Boza, et al. Muon: An optimizer for hidden layers in neural networks, 2024.

[22] Wu Lin, Scott C. Lowe, Felix Dangel, et al. Understanding and improving shampoo and SOAP via kullback-leibler minimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

[23] Nikhil Vyas, Depen Morwani, Rosie Zhao, et al. SOAP: Improving and stabilizing shampoo using adam for language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. Technical report, Meta, 2024. Technical Report.

[25] Zhuoran jin, Pengfei Cao, Chenhao Wang, et al. Rwku: Benchmarking real-world knowledge unlearning for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 98213–98263. Curran Associates, Inc., 2024.

[26] Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, et al. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*, 2024.

[27] Vineeth Dorna, Anmol Reddy Mekala, Wenlong Zhao, et al. Openunlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[28] Shengyuan Hu, Yiwei Fu, Steven Wu, et al. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[29] Sangyeon Yoon, Hyesoo Hong, Wonje Jeung, et al. Rethinking benign relearning: Syntax as the hidden driver of unlearning failures. In *The Fourteenth International Conference on Learning Representations*, 2026.

[30] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, et al. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pages 1–16. IEEE, 2020.

[31] Chongyu Fan, Jiancheng Liu, Licong Lin, et al. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Neurips Safe Generative AI Workshop 2024*, 2024.

[32] Bo Yang. Ce-u: Cross entropy unlearning, 2025.

[33] Taha Entesari, Arman Hatami, Rinat Khaziev, et al. Constrained entropic unlearning: A primal-dual framework for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[34] Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, et al. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[35] Puning Yang, Qizhou Wang, Zhuo Huang, et al. Exploring criteria of loss reweighting to enhance LLM unlearning. In *Forty-second International Conference on Machine Learning*, 2025.

[36] Anmol Mekala, Vineeth Dorna, Shreya Dubey, et al. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3732–3752, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[37] Jin Yao, Eli Chien, Minxin Du, et al. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[38] Weijia Shi, Anirudh Ajith, Mengzhou Xia, et al. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

A Proofs

This appendix proofs Lemma 1 and Theorem 2. We proceed in three steps. First, we derive a general closed form for steepest descent under any norm via duality. Second, we identify the RMS-to-RMS norm as a rescaled spectral norm and compute its dual norm and duality map. Third, we combine these ingredients to prove the multi-layer result.

A.1 Steepest descent via duality

Steepest descent under any norm reduces to two scalar quantities: the dual norm and the duality map. We make this precise here.

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. The *dual norm* of $g \in \mathbb{R}^n$ measures the largest possible inner product with g over the unit ball of $\|\cdot\|$,

$$\|g\|_* := \max_{t\in\mathbb{R}^n\,:\,\|t\|=1} g^\top t, \tag{18}$$

and the *duality map* returns a unit vector achieving the maximum,

$$\mathrm{dualize}_{\|\cdot\|}(g) := \arg\max_{\|t\|=1} g^\top t. \tag{19}$$

Intuitively, the duality map is the unit-norm direction most aligned with g; the dual norm is the inner product achieved along that direction.

Lemma 3 (Steepest descent, closed form)**.** *For any $g \in \mathbb{R}^n$ and $\lambda > 0$,*

$$\arg\min_{\Delta w\in\mathbb{R}^n}\left[g^\top \Delta w + \tfrac{\lambda}{2}\|\Delta w\|^2\right] = -\frac{\|g\|_*}{\lambda}\,\mathrm{dualize}_{\|\cdot\|}(g). \tag{20}$$

Proof. Decompose $\Delta w = c\cdot t$ into magnitude $c \geq 0$ and direction t with $\|t\| = 1$. The objective becomes

$$c\cdot g^\top t + \tfrac{\lambda}{2}c^2. \tag{21}$$

With c fixed, only the linear term $c\,g^\top t$ depends on t. Since $c \geq 0$, this is minimized by choosing t to make $g^\top t$ as negative as possible, that is,

$$t^\star = -\,\mathrm{dualize}_{\|\cdot\|}(g), \tag{22}$$

which gives $c\cdot g^\top t^\star = -c\cdot\|g\|_*$. Substituting back, the problem reduces to minimizing

$$-c\cdot\|g\|_* + \tfrac{\lambda}{2}c^2 \tag{23}$$

over $c \geq 0$. This is a convex quadratic, minimized at $c^\star = \|g\|_*/\lambda$. Combining,

$$\Delta w^\star = c^\star\cdot t^\star = -\frac{\|g\|_*}{\lambda}\,\mathrm{dualize}_{\|\cdot\|}(g). \tag{24}$$

□

Lemma 3 is the workhorse of the appendix. Specializing to a particular norm requires only computing $\|g\|_*$ and $\mathrm{dualize}_{\|\cdot\|}(g)$ for that norm. The next two sections do this for the RMS-to-RMS norm.

A.2 The RMS-to-RMS norm as a rescaled spectral norm

The RMS-to-RMS norm differs from the spectral norm only in how it measures input and output magnitudes: $\|\cdot\|_{\mathrm{RMS}}$ divides by $\sqrt{d}$, whereas $\|\cdot\|_2$ does not. The two norms therefore differ by a ratio of the input and output dimensions.

Proof of Lemma 1. The constraint $\|x\|_{\mathrm{RMS}} = 1$ is equivalent to $\|x\|_2 = \sqrt{d_{\mathrm{in}}}$, and $\|Wx\|_{\mathrm{RMS}} = \|Wx\|_2/\sqrt{d_{\mathrm{out}}}$. Substituting both into the RMS-to-RMS operator norm,

$$\|W\|_{\mathrm{RMS}\to\mathrm{RMS}} = \max_{\|x\|_{\mathrm{RMS}}=1}\|Wx\|_{\mathrm{RMS}} = \frac{1}{\sqrt{d_{\mathrm{out}}}}\max_{\|x\|_2=\sqrt{d_{\mathrm{in}}}}\|Wx\|_2. \tag{25}$$

The substitution $x = \sqrt{d_{\text{in}}}\,\tilde{x}$ rescales the constraint to $\|\tilde{x}\|_2 = 1$,

$$\|W\|_{\text{RMS}\to\text{RMS}} = \frac{\sqrt{d_{\text{in}}}}{\sqrt{d_{\text{out}}}} \max_{\|\tilde{x}\|_2=1} \|W\tilde{x}\|_2 = \sqrt{\tfrac{d_{\text{in}}}{d_{\text{out}}}}\,\|W\|_2, \tag{26}$$

where the last equality uses the standard characterization of the spectral norm. □

A useful consequence: a matrix has unit RMS-to-RMS norm if and only if its largest singular value is exactly $\sqrt{d_{\text{out}}/d_{\text{in}}}$. We use this repeatedly below.

A.3 The dual norm and duality map of the RMS-to-RMS norm

The dual quantities of the RMS-to-RMS norm have a clean SVD-based description: the dual norm is a rescaled trace of singular values, and the duality map is the SVD-aligned direction $UV^\top$ (also rescaled).

Lemma 4. *Let $G \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$ have reduced SVD $G = U\Sigma V^\top$, where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ with $r = \min(d_{\text{out}}, d_{\text{in}})$. Then*

$$\|G\|_{\text{RMS}\to\text{RMS},*} = \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}}\ \operatorname{tr}\Sigma, \tag{27}$$

$$\operatorname{dualize}_{\|\cdot\|_{\text{RMS}\to\text{RMS}}}(G) = \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}}\,UV^\top. \tag{28}$$

The duality map is unique if and only if G has full rank.

Proof. By Lemma 1, the constraint $\|T\|_{\text{RMS}\to\text{RMS}} = 1$ is equivalent to $\|T\|_2 = \sqrt{d_{\text{out}}/d_{\text{in}}}$.

Upper bound. Expanding the Frobenius inner product using the SVD of G,

$$\langle G, T\rangle = \operatorname{tr}(V\Sigma U^\top T) = \sum_{i=1}^{r} \sigma_i\, u_i^\top T v_i. \tag{29}$$

Since u_i and v_i are unit vectors and $\|T\|_2 = \sqrt{d_{\text{out}}/d_{\text{in}}}$, each summand is bounded by

$$|u_i^\top T v_i| \le \|u_i\|_2 \cdot \|Tv_i\|_2 \le \|T\|_2 = \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}}. \tag{30}$$

Since $\sigma_i \ge 0$, summing gives

$$\langle G, T\rangle \le \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}} \sum_{i=1}^{r} \sigma_i = \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}}\ \operatorname{tr}\Sigma. \tag{31}$$

Take

$$T^\star := \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}}\,UV^\top. \tag{32}$$

The matrix $UV^\top$ has spectral norm 1 (its nonzero singular values are all equal to 1), so $\|T^\star\|_2 = \sqrt{d_{\text{out}}/d_{\text{in}}}$ as required. Furthermore, $V^\top v_i = e_i$ and $Ue_i = u_i$, so

$$u_i^\top T^\star v_i = \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}}\, u_i^\top u_i = \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}} \tag{33}$$

for each i with $\sigma_i > 0$. Substituting into (29),

$$\langle G, T^\star\rangle = \sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}}\ \operatorname{tr}\Sigma, \tag{34}$$

which attains the upper bound of (31). This proves both (27) and that $T^\star$ is a valid duality map.

Uniqueness. When G has full rank, every $\sigma_i > 0$, so equality must hold in every term of (30). By the equality conditions for Cauchy–Schwarz, this forces $Tv_i = \sqrt{d_{\text{out}}/d_{\text{in}}}\,u_i$ for every i. The vectors $\{v_i\}$ form a basis of $\mathbb{R}^{d_{\text{in}}}$, so T is determined uniquely on its full domain, giving $T = T^\star$. When G is rank-deficient, the basis $\{v_i\}$ spans only the row space of G, leaving T free on the null space, so the duality map is not unique. □

The intuition behind (28) is that, among all matrices of bounded RMS-to-RMS norm, the one that aligns most with G is the SVD-aligned direction $UV^\top$ (rescaled to satisfy the constraint). The SVD-aligned direction preserves the singular directions of G but flattens its spectrum to all-ones, putting the entire "budget" uniformly along the directions G already uses.

A.4 Multi-layer steepest descent

We now combine Lemmas 3 and 4 with the layerwise-max network norm to prove the main theorem. We adopt the main-text notation $d_{\text{in},l}, d_{\text{out},l}$ for layer-l input/output dimensions.

Proof of Theorem 2. The steepest-descent problem under the layerwise-max RMS-to-RMS network norm reads

$$\arg \min_{\Delta W_1,\dots,\Delta W_L} \left[\sum_{l=1}^{L} \langle G_l, \Delta W_l \rangle + \frac{\lambda}{2} \max_l \|\Delta W_l\|^2_{\text{RMS}\to\text{RMS}} \right]. \tag{35}$$

For each layer, decompose $\Delta W_l = c_l \cdot T_l$ into magnitude $c_l \geq 0$ and direction T_l with $\|T_l\|_{\text{RMS}\to\text{RMS}} = 1$. The objective becomes

$$\sum_{l=1}^{L} c_l \langle G_l, T_l \rangle + \frac{\lambda}{2} \max_l c_l^2. \tag{36}$$

We minimize in three steps: directions T_l first (each independent of the others), then the layerwise magnitudes c_l.

Step 1: optimal direction at each layer. Each T_l appears in only one term of (36), namely $c_l \langle G_l, T_l \rangle$, and its constraint is independent of the other layers. By Lemma 4, the unit-norm direction minimizing $c_l \langle G_l, T_l \rangle$ is the negative of the duality map,

$$T_l^\star = -\operatorname{dualize}_{\|\cdot\|_{\text{RMS}\to\text{RMS}}}(G_l) = -\sqrt{\tfrac{d_{\text{out},l}}{d_{\text{in},l}}}\, U_l V_l^\top, \tag{37}$$

and the resulting linear contribution is

$$c_l \langle G_l, T_l^\star \rangle = -c_l \|G_l\|_{\text{RMS}\to\text{RMS},*} = -c_l\, s_l, \quad \text{where} \quad s_l := \sqrt{\tfrac{d_{\text{out},l}}{d_{\text{in},l}}}\, \operatorname{tr} \Sigma_l. \tag{38}$$

The scalar $s_l \geq 0$ is the per-layer "reward rate" for spending magnitude.

Step 2: all magnitudes meet at a common value. Substituting (38), the problem reduces to a scalar optimization,

$$\min_{c_1,\dots,c_L \geq 0} \left[-\sum_{l=1}^{L} c_l\, s_l + \frac{\lambda}{2} \max_l c_l^2 \right]. \tag{39}$$

The structure of (39) is informative: the penalty depends only on the *largest* c_l, while every c_l contributes a linear reward s_l. Increasing any c_l that is below $\max_l c_l$ therefore costs nothing (the max is unchanged) but yields a positive linear gain $s_l > 0$. Consequently, all c_l with $s_l > 0$ must coincide at the optimum; layers with $s_l = 0$ have no incentive to use any magnitude and can be set to zero. Setting $c_1 = \cdots = c_L = c^\star$, the problem collapses to

$$\min_{c^\star \geq 0} \left[-c^\star \sum_{l=1}^{L} s_l + \frac{\lambda}{2} (c^\star)^2 \right], \tag{40}$$

which is a convex quadratic in $c^\star$, minimized at

$$c^\star = \frac{1}{\lambda} \sum_{l=1}^{L} s_l = \frac{1}{\lambda} \sum_{l=1}^{L} \sqrt{\tfrac{d_{\text{out},l}}{d_{\text{in},l}}}\, \operatorname{tr} \Sigma_l. \tag{41}$$

Step 3: assemble the update. Combining $\Delta W_l^\star = c^\star \cdot T_l^\star$ with (37) and (41),

$$\Delta W_l^\star = -c^\star \sqrt{\tfrac{d_{\text{out},l}}{d_{\text{in},l}}}\, U_l V_l^\top = -\eta_l\, U_l V_l^\top, \tag{42}$$

where the layer-l learning rate is the global scalar $c^\star$ multiplied by the layer-specific dimensional factor:

$$\eta_l = c^\star \sqrt{\tfrac{d_{\text{out},l}}{d_{\text{in},l}}} = \frac{1}{\lambda} \left(\sum_{k=1}^{L} \sqrt{\tfrac{d_{\text{out},k}}{d_{\text{in},k}}}\, \operatorname{tr} \Sigma_k \right) \sqrt{\tfrac{d_{\text{out},l}}{d_{\text{in},l}}}. \tag{43}$$

This matches the statement of Theorem 2.

Uniqueness. The direction $T_l^\star$ in Step 1 is unique exactly when G_l has full rank (Lemma 4). The magnitude in Step 2 is uniquely determined whenever $\sum_l s_l > 0$, which holds whenever any G_l has nonzero singular values. The overall minimizer is therefore unique iff every G_l has full rank. □

The update rule $\Delta W_l^\star = -\eta_l\, U_l V_l^\top$ has a clean interpretation. The direction $-U_l V_l^\top$ is the SVD-aligned direction of $-G_l$: it preserves the singular subspaces of the gradient but flattens the singular spectrum. The magnitude η_l couples a global term $c^\star$ with a per-layer dimensional factor. The global term aggregates contributions from every layer through the dual-norm sum, so the step taken at any one layer reflects the gradient signal accumulated across the network. The per-layer factor $\sqrt{d_{\mathrm{out},l}/d_{\mathrm{in},l}}$ is the same dimensional rescaling that appears in Lemma 1; in practice we absorb it into the per-layer learning rate, giving the unrescaled update $\Delta W_l = -\eta\, U_l V_l^\top$ used in our experiments.

Relation to the prior work. The view that the spectral norm provides a natural geometry for transformer weight matrices was developed on theoretical grounds by [20], who interpret several existing optimizers as steepest descent under different operator norms. Their analysis establishes the geometric foundation, and we build on it with a controlled empirical study at the LLM scale. Our contribution is complementary and primarily empirical: (i) we extend their layerwise argument to a layerwise-maximum RMS-to-RMS network norm and obtain a closed-form per-layer step size in Theorem 2; (ii) we measure, throughout training, the cosine similarity between three practical preconditioners and the optimal Kronecker direction (Fig. 2; Figs. 7, 8), confirming that the spectral direction is realizable at LLM scale; and (iii) we instantiate this geometry on a non-trivial downstream task—LLM unlearning—across three loss families, two model scales, and three forget splits, and demonstrate consistent gains over entry-wise updates (Tab. 1, Fig. 3). To our knowledge, this is the first systematic empirical validation that the spectral-norm geometry, beyond being theoretically natural, operationally improves LLM training in a post-training regime such as unlearning.

B Practical Realization on the LLM scale

This appendix analyzes how the spectral update of Theorem 2 is instantiated on Llama-3.2-1B and 3B. We address (i) which dimension of each weight matrix lives in the RMS-normalized residual stream, (ii) why our asymmetric preconditioning aligns with this structure, and (iii) why the resulting one-sided update is exact rather than approximate.

B.1 Architectural placement of RMSNorm

In a pre-norm transformer, RMSNorm is applied only to vectors in the residual-stream dimension d. The MLP intermediate dimension d_{ff} is transient: W_1 expands $d \to d_{\mathrm{ff}}$, a pointwise activation is applied, and W_2 projects back to d. No normalization layer touches the d_{ff}-space.

Consequently, the RMS-to-RMS geometry of §4.2 is well-defined on the d-side of every weight matrix but not on the d_{ff}-side. This distinction motivates the preconditioning design that follows.

B.2 Per-matrix dimension regime

We set `max_precond_dim` $= d$ ($d = 2048$ for 1B; $d = 3072$ for 3B) with `merge_dims` $=$ False. Under this setting, the SOAP accumulators $L = \sum_t G_t G_t^\top$ and $R = \sum_t G_t^\top G_t$ are maintained per side only when that side does not exceed d. Tab. 4 summarizes the resulting regime.

For attention projections, both sides equal d and full two-sided preconditioning applies. For MLP matrices, the side that is skipped is in each case the un-normalized d_{ff}-dimension.

B.3 Scope of the RMS-to-RMS assumption

Theorem 2 idealizes every weight matrix as acting on unit-RMS inputs. This holds exactly for W_Q, W_K, W_V, W_1, which receive RMSNorm output directly. Inputs to W_O and W_2 traverse attention and activation respectively, and have RMS norms that are bounded but not strictly unit. The deviation is benign for the steepest-descent direction.

Table 4: Per-matrix preconditioning regime under `max_precond_dim` $= d$. The un-preconditioned side coincides with the un-normalized d_{ff}-dimension.

Matrix	Shape ($d_{\text{out}} \times d_{\text{in}}$) 1B	Shape ($d_{\text{out}} \times d_{\text{in}}$) 3B	Input	RMS side	Precond. L	Precond. R	Regime
W_Q	2048×2048	3072×3072	$\text{RMSNorm}(X_l)$	both	✓	✓	two-sided
W_K	2048×2048	3072×3072	$\text{RMSNorm}(X_l)$	both	✓	✓	two-sided
W_V	2048×2048	3072×3072	$\text{RMSNorm}(X_l)$	both	✓	✓	two-sided
W_O	2048×2048	3072×3072	attention out	both†	✓	✓	two-sided
W_1 (gate, up_proj)	8192×2048	8192×3072	$\text{RMSNorm}(H_l)$	input (d)	—	✓	right-only
W_2 (down_proj)	2048×8192	3072×8192	$\text{activation}(\cdot)$	output (d)	✓	—	left-only
W_{emb}	$128256 \times d$	$128256 \times d$	token id	—	AdamW		entry-wise

† Bounded by the value projection's spectral norm; the corresponding scale factor is absorbed into the per-layer learning rate (see App. B.3).

Scale-invariance of the spectral direction. For inputs with $\|x\|_{\text{RMS}} = c > 0$, the induced operator norm becomes

$$\|W\|_{c\text{-RMS}} := \max_{\|x\|_{\text{RMS}}=c} \|Wx\|_{\text{RMS}} = c\sqrt{\tfrac{d_{\text{in}}}{d_{\text{out}}}}\,\|W\|_2, \tag{44}$$

which is a direct rescaling of Lemma 1. Applying Lemmas 4 and 3 in turn yields the steepest-descent direction

$$T^\star(c) = -\sqrt{\tfrac{d_{\text{out}}}{d_{\text{in}}}}\,UV^\top, \tag{45}$$

which is *independent of* c; only the implied step size scales as $1/c$. The non-unit RMS at the inputs to W_O and W_2 therefore affects only the per-layer step size, which is absorbed into the learning rate as in Theorem 2.

B.4 Asymmetric preconditioning aligns with the RMS-controlled dimension

The asymmetric preconditioning in Tab. 4 is not a memory-driven approximation but a direct reflection of the architectural distinction in App. B.1. For $W_1 \in \mathbb{R}^{d_{\text{ff}} \times d}$, the input side ($d$) lives in the RMS-normalized residual stream while the output side (d_{ff}) does not; we accumulate $R \in \mathbb{R}^{d \times d}$ and skip L. For $W_2 \in \mathbb{R}^{d \times d_{\text{ff}}}$, the situation is mirrored: we accumulate $L \in \mathbb{R}^{d \times d}$ and skip R. In both cases, the un-preconditioned side is precisely the side where RMS-to-RMS geometry does not apply. The setting `max_precond_dim` $= d$ enforces this alignment automatically.

B.5 One-sided preconditioning recovers the exact SVD-aligned direction

The asymmetric design above might appear to approximate Theorem 2, since the SVD-aligned direction $UV^\top$ is conventionally defined via the full SVD. The following result shows that no approximation is incurred.

Proposition 5 (One-sided preconditioning recovers the SVD-aligned direction)**.** *Let $G \in \mathbb{R}^{m \times n}$ have reduced SVD $G = U\Sigma V^\top$ with $r = \min(m, n)$, and assume Σ is invertible.*

(i) If $m \geq n$ (tall):

$$\Delta W^{\text{right}} := -\eta\, G\,(G^\top G)^{-1/2} = -\eta\, UV^\top. \tag{46}$$

(ii) If $m \leq n$ (wide):

$$\Delta W^{\text{left}} := -\eta\,(GG^\top)^{-1/2}\, G = -\eta\, UV^\top. \tag{47}$$

Proof. (i) $G^\top G = V\Sigma^2 V^\top$ gives $(G^\top G)^{-1/2} = V\Sigma^{-1}V^\top$, so

$$G(G^\top G)^{-1/2} = U\Sigma V^\top \cdot V\Sigma^{-1}V^\top = UV^\top. \tag{48}$$

(ii) Symmetric: $(GG^\top)^{-1/2} = U\Sigma^{-1}U^\top$, so

$$(GG^\top)^{-1/2}G = U\Sigma^{-1}U^\top \cdot U\Sigma V^\top = UV^\top. \qquad \square$$

The two cases match the two MLP regimes of Tab. 4 exactly: W_1 is tall and uses right-only preconditioning, W_2 is wide and uses left-only. The architectural alignment of App. B.4 is therefore mathematically exact.

Accumulated preconditioners. Prop. 5 treats the instantaneous gradient. SOAP and KLShampoo accumulate $G_tG_t^\top$ and $G_t^\top G_t$ via EMA, yielding covariance-whitened approximations of the SVD-aligned direction. The same one-sided structure carries over: the side with accumulated statistics matches the polar-factor direction in expectation, while the un-accumulated side is correctly left identity by the architectural argument.

B.6 Architectural Alignment of `max_precond_dim`

Combining the matrix-shape inventory of attention and MLP projections (Tab. 4) with the scale invariance of the SVD-aligned direction and the one-sided polar reduction (Prop. 5), one obtains a coherent picture across both blocks: the $d \times d$ attention projections W_Q, W_K, W_V receive exact two-sided preconditioning, the output projection W_O is preconditioned up to a scale absorbed by RMS normalization, and the rectangular MLP projections receive one-sided preconditioning that nonetheless recovers the polar-factor direction exactly. Under this lens, `max_precond_dim` $= d$ is an architectural alignment with the transformer's matrix structure, not a memory-driven compromise.

C Experimental Setup and Hyperparameter

C.1 Hardware and Software Stack

All experiments are conducted on a server equipped with 4 NVIDIA RTX A6000 GPUs (48 GB each), running CUDA 12.x. For Llama-3.2-1B and 3B-Instruct, each training run fits on a single GPU; multi-GPU parallelism is exploited only at the job level, with one training process per GPU within our learning-rate sweep. For the 7B-scale Zephyr-7B-β used in WMDP experiments, we employ DeepSpeed ZeRO Stage 3 [30] to shard optimizer states, gradients, and parameters across all four GPUs. The deep-learning stack is built on PyTorch with FlashAttention-2, the HuggingFace Transformers (v4.51.3) and Accelerate libraries, and TensorBoard logging. The base models for TOFU experiments are the OpenUnlearning fine-tuned checkpoints released for Llama-3.2-1B-Instruct and 3B-Instruct.

C.2 Common Training Configuration

Unless stated otherwise, every unlearning run shares the following training configuration. We train for $T = 10$ epochs with the first epoch reserved for linear warmup, followed by a linear decay of the learning rate to zero over the remaining nine epochs. Optimization runs in bfloat16 mixed precision; for spectral updates (Muon, KLShampoo, SOAP), the optimizer state itself is also stored in bfloat16, while the entry-wise baselines (AdamW, Sophia) follow their upstream default dtypes. Weight decay is fixed at 0.01, and gradient checkpointing is disabled because both 1B and 3B models comfortably fit in memory at our chosen batch sizes.

The micro-batch size is set per model so that the effective batch size matches the OpenUnlearning protocol within each scale: 8 with gradient accumulation of 4 (effective 32) for the 1B model, and 4 with the same accumulation factor (effective 16) for the 3B model. The total update-step budget over 10 epochs is approximately $13/63/125$ steps for the 1%/5%/10% forget splits on the 1B model, and $25/125/250$ steps on the 3B model. Evaluation is performed once at the end of training so that the wall-clock numbers reported in the main paper exclude in-loop evaluation overhead.

C.3 Unlearning Loss Hyperparameters

We use the canonical OpenUnlearning configurations for the three unlearning losses considered in the main paper.

Table 5: Peak learning rate from the learning rate sweep.

	Llama-3.2-1B-Instruct			Llama-3.2-3B-Instruct		
Method	Forget 10%	Forget 5%	Forget 1%	Forget 10%	Forget 5%	Forget 1%
GD-AdamW	1×10^{-5}	1×10^{-5}	2×10^{-5}	5×10^{-6}	1×10^{-5}	2×10^{-5}
GD-Sophia	5×10^{-6}	1×10^{-5}	2×10^{-5}	5×10^{-6}	7×10^{-6}	2×10^{-5}
SU-GD-Muon	2×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-5}	1.5×10^{-5}	2×10^{-5}
SU-GD-KLShampoo	2×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-5}	1.5×10^{-5}	2×10^{-5}
SU-GD-SOAP	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
NPO-AdamW	3×10^{-5}	1×10^{-5}	1×10^{-5}	2×10^{-5}	1×10^{-5}	1×10^{-5}
NPO-Sophia	2×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-5}	2×10^{-5}	2×10^{-5}
SU-NPO-Muon	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}	2×10^{-5}	2×10^{-5}
SU-NPO-KLShampoo	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}	2×10^{-5}	2×10^{-5}
SU-NPO-SOAP	3×10^{-5}	3×10^{-5}	2×10^{-5}	3×10^{-5}	3×10^{-5}	2×10^{-5}
RMU-AdamW	1×10^{-5}	2×10^{-5}	3×10^{-5}	1×10^{-5}	2×10^{-5}	2×10^{-5}
RMU-Sophia	2×10^{-5}	3×10^{-5}	1×10^{-4}	2×10^{-5}	2×10^{-5}	5×10^{-5}
SU-RMU-Muon	2×10^{-5}	2×10^{-5}	3×10^{-5}	2×10^{-5}	2×10^{-5}	3×10^{-5}
SU-RMU-KLShampoo	2×10^{-5}	2×10^{-5}	3×10^{-5}	2×10^{-5}	3×10^{-5}	3×10^{-5}
SU-RMU-SOAP	2×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}

GD. Both forget and retain weights are set to $\gamma = \alpha = 1.0$, with the retain loss being the standard negative log-likelihood on full retain question–answer pairs.

NPO. We use the inverse temperature $\beta = 0.1$ recommended by [16], with $\gamma = \alpha = 1.0$. The reference model π_{ref} is loaded from the same fine-tuned checkpoint as w_0 and held frozen throughout training, contributing only forward-pass log-probabilities.

RMU. The forget term steers hidden states at the 7th transformer block (a fixed mid-stack layer used in both 1B and 3B configurations) toward a random unit-norm direction with steering coefficient $c = 2$, and the retain regularizer is the embedding-distance loss against the reference model. All transformer parameters are trainable, mirroring the official RMU release rather than the layer-restricted variant used for WMDP.

C.4 Optimizer Hyperparameters

Each loss is paired with five optimizers, yielding the $3 \times 5 = 15$ combinations reported in the main table. Optimizer-internal hyperparameters are fixed at the defaults from their respective original papers, and only the learning rate is tuned per $\langle$model, forget split$\rangle$ cell over the 9-point log-spaced grid

$$\Lambda = \{1, 2, 3, 5\} \times \{10^{-6}, 10^{-5}\} \cup \{1\} \times 10^{-4}.$$

This isolates the effect of the matrix-structured update direction from any tuning advantage the spectral updates might otherwise enjoy. Although the optimal learning rate varies across cells, the cell-wise winners cluster tightly around an optimizer-dependent value, which we report below as the empirical sweet-spot range. For GD on Llama-3.2-3B-Instruct with the 5% forget split, we also tested two intermediate learning rates, 7×10^{-6} and 1.5×10^{-5}, to identify the range in which retain-set performance remained stable. The fact that the sharper spectral updates prefer slightly larger learning rates is consistent with their unit-norm step direction, which has a smaller effective magnitude per step than the entry-wise sign update; this pattern is also reflected in the learning-rate sensitivity analysis of Fig. 4. The peak learning rate from the learning rate sweep is provided in Tab. 5.

AdamW. Standard configuration with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$.

Sophia. Following SOUL [17], we use the `SophiaG` variant with $\beta_1 = 0.9$, $\beta_2 = 0.95$, the per-coordinate clipping threshold $\rho = 0.04$, $\varepsilon = 10^{-5}$, and Hessian re-estimation every 10 update steps.

Muon. Momentum is set to 0.95, and the SVD-aligned direction is approximated by 5 Newton–Schulz iterations per step, matching the original release [21]. Bias correction is enabled.

KLShampoo. The two preconditioner EMA coefficients are both set to 0.95, the preconditioner is refreshed every 10 steps, and we use a per-axis dimension cap of 3072 with output clamping at 4000 for numerical stability [22].

SOAP. The inner Adam coefficients are $(\beta_1, \beta_2) = (0.95, 0.95)$, with $\varepsilon = 10^{-8}$, the eigenbasis (Q^L, Q^R) refreshed every 10 steps via an `eigh` call in `float32`, and the rest of the hot path running in `bfloat16` [23]. The single-axis preconditioning cap is set to `max_precond_dim` $= d$, where d denotes the transformer hidden dimension. This choice is not driven by memory: rather, it aligns the preconditioner's coverage with the matrix structure of the model, so that the $d \times d$ attention projections receive full two-sided preconditioning, while MLP projections (with one side $4d$) are preconditioned only on the d-side—which, as discussed in Sec. 3.2, recovers the polar-factor direction exactly under one-sided preconditioning. Sensitivity sweeps over β_2, β_{shampoo}, f_{prec}, and ϵ in the neighborhood of these defaults produced no statistically significant change in the aggregate score, consistent with the robustness reported for SOAP in the original paper [23].

C.5 Multi-Seed Protocol

Every entry in the main-paper tables, except for the reference rows (Original, Retrain), is reported as mean $\pm$ standard deviation over three seeds $s \in \{0, 1, 2\}$. The seed controls the Trainer's data shuffler and any update-internal stochasticity through Python, NumPy, and PyTorch CPU/CUDA generators. We do not re-randomize the model initialization (which is fixed to the open-unlearning checkpoint) nor the data partition (which is fixed by the upstream split), so the seed governs only the training trajectory.

C.6 Numerical Stability and Tokenization Details

Token log-probabilities are obtained via `F.log_softmax(logits, dim=-1)` rather than $\log(\text{softmax}(\cdot))$ to avoid numerical underflow. Prompt tokens are masked with `ignore_index` $= -100$ and excluded from all loss and likelihood computations. Logits and labels are shifted by one position, so that `logits[:-1]` predicts `labels[1:]`. Mixed-precision activations are cast to `float32` prior to metric aggregation.

C.7 Comparison Algorithms

This appendix details the unlearning baselines reported in Tab. 12 and Tab. 14. We follow the canonical formulations and configurations from the open-unlearning benchmark.

SimNPO [31]. A SimPO-style formulation with forget loss weight $\gamma = 0.125$, $\beta = 4.5$, and length penalty $\delta = 0$, combined with retain NLL.

IdkDPO [5]. The same DPO formulation as NPO ($\beta = 0.1$), but additionally uses "I don't know" responses as the preferred completions on the forget set.

CEU [32]. A cross-entropy unlearning loss that skips the first token, with both forget and retain loss weights equal to 1.

PDU [33]. A primal-dual formulation with a logit-gap-squared forget loss, an adaptive retain weight initialized at 1, a constraint threshold $\epsilon = 0.3$, and 5 epochs of dual-variable warmup.

UNDIAL [34]. A self-distillation loss with temperature $\beta = 10$ and no retain regularizer. Uses a higher peak learning rate of 10^{-4} to compensate for the absence of explicit retention pressure.

WGA [12]. A weighted gradient attribution forget loss with $\beta = 1$, combined with retain NLL. Trained for 5 epochs.

SatImp [35]. A sharpness-aware forget loss with $\beta_1 = 5$, $\beta_2 = 1$, $\gamma = 0.1$, combined with retain NLL. Trained for 5 epochs.

AltPO [36]. An alternating preference optimization scheme that toggles between forget and retain phases under the canonical open-unlearning configuration.

GradAscent [37]. The vanilla GA baseline described above, included as a lower-bound reference to highlight how much each refinement contributes.

D Evaluation Protocol Details

We adopt the OpenUnlearning evaluation framework [27], originally introduced for the TOFU benchmark [5], and report results along three axes: memorization removal (Mem), privacy preservation (Priv), and utility retention (Util). All component scores are bounded in $[0, 1]$ and combined via the harmonic mean (HM), defined for non-negative inputs $x_1, \ldots, x_n$ as

$$\mathrm{HM}(x_1, \ldots, x_n) = \frac{n}{\sum_{i=1}^{n} \max(x_i, \epsilon)^{-1}}, \qquad \epsilon = 10^{-12}. \tag{49}$$

We use the harmonic mean rather than the arithmetic mean throughout because it ensures that a method cannot mask a weak axis by excelling on another: a near-zero value in any single component drives the aggregate toward zero. This conservative aggregation is important for unlearning, where memorization removal, privacy, and utility are often in tension. The final *aggregate score* is

$$\mathrm{Agg} = \mathrm{HM}\big(\mathrm{Mem},\ \mathrm{Priv},\ \mathrm{Util}\big). \tag{50}$$

D.1 Memorization Score

The Mem score quantifies how thoroughly the target knowledge has been removed. We aggregate four sub-metrics computed on the forget set $\mathcal{D}_f$:

$$\mathrm{Mem} = \mathrm{HM}\big(1 - \mathrm{ES},\ 1 - \mathrm{EM},\ 1 - \mathrm{F.Prob},\ \mathrm{F.TR}\big). \tag{51}$$

The first three sub-metrics measure residual knowledge directly (lower is better), so each is inverted as $1 - v$ before aggregation. The truth ratio (F.TR) is normalized so that higher already corresponds to better forgetting and is used as is.

Extraction Strength (ES). For a target sequence of length L, let $\hat{t}_i = \arg\max_v \log p(v \mid t_{<i})$ be the greedy prediction at position i, and let k be the smallest index for which the suffix matches, $\hat{t}_{k:L} = t_{k:L}$. ES is defined as $\mathrm{ES} = 1 - k/L$. Values near 1 indicate that long contiguous suffixes can be extracted greedily.

Exact Memorization (EM). EM measures token-level greedy agreement, $\mathrm{EM} = \frac{1}{L}\sum_{i=1}^{L} \mathbb{1}[\hat{t}_i = t_i]$. Unlike ES, EM credits non-contiguous matches and therefore captures partial memorization.

Forget Probability (F.Prob). We compute the model's normalized likelihood on the original (non-paraphrased) forget prompts:

$$\mathrm{F.Prob} = \frac{1}{|\mathcal{D}_f|} \sum_{z \in \mathcal{D}_f} \exp\left(-\frac{1}{L_z} \sum_{i=1}^{L_z} \mathrm{CE}(\mathrm{logits}_i, t_i) \right), \tag{52}$$

where prompt tokens are masked with `ignore_index` $= -100$. We deliberately use F.Prob rather than the paraphrase-prompt variant adopted in some prior protocols, because the paraphrase signal is already captured by the truth ratio below; including both would double-count the same evidence. We nonetheless report the paraphrase probability (P.Prob) in our detailed sub-metric tables (Tabs. 17–22) for transparency and to facilitate comparison with prior work.

Forget Truth Ratio (F.TR). For each forget example z, we compute the per-token likelihoods of the paraphrased correct answer (p_z^{para}) and a perturbed incorrect answer (p_z^{pert}), and form the raw ratio $R_z = p_z^{\text{pert}}/(p_z^{\text{para}} + 10^{-10})$. The forget-side truth ratio is then the symmetrized mean

$$\text{F.TR} = \frac{1}{|\mathcal{D}_f|} \sum_{z \in \mathcal{D}_f} \min\Big(R_z,\ (R_z + 10^{-10})^{-1}\Big). \tag{53}$$

F.TR is maximized when correct and incorrect answers receive comparable likelihood—the expected behavior of a model that has truly forgotten the target.

D.2 Privacy Score

Privacy is evaluated through resistance to membership inference attacks (MIA). We use four scoring functions—LOSS, ZLib, Min-$k\%$, and Min-$k\%$++ [38]—and for each compute an attack AUC by treating forget examples as members and a held-out set as non-members. Let AUC_k denote the attack AUC of method $k \in \{\text{LOSS}, \text{ZLib}, \text{MinK}, \text{MinK++}\}$ on the unlearned model, $\text{AUC}_k^{\text{retain}}$ the corresponding AUC on the retain-trained oracle, and $\text{AUC}_k^{\text{init}}$ the AUC on the original (pre-unlearning) model. Each attack contributes a similarity score

$$s_k = 1 - |\text{AUC}_k - \text{AUC}_k^{\text{retain}}|. \tag{54}$$

This score equals 1 when the two models have the same AUC under attack k and decreases with the absolute difference between their AUCs. The privacy score is

$$\text{Priv} = \text{HM}\big(s_{\text{LOSS}},\ s_{\text{ZLib}},\ s_{\text{MinK}},\ s_{\text{MinK++}}\big). \tag{55}$$

The four attack scoring functions follow the standard definitions: LOSS uses average cross-entropy, ZLib divides this loss by the zlib-compressed length of the input to control for intrinsic difficulty, Min-$k\%$ averages only the bottom $k = 40\%$ of token log-probabilities, and Min-$k\%$++ further standardizes each token's log-probability by the mean and variance of the per-position categorical distribution before applying the same bottom-k rule.

D.3 Utility Score

To preserve a uniform weighting across all retention sub-metrics, we compute Util as a single harmonic mean of ten components rather than nesting them through an intermediate "model utility" aggregate:

$$\begin{aligned}\text{Util} = \text{HM}\big(&\text{R.Prob},\ \text{R.ROUGE},\ \text{R.TR}, \text{RA.Prob},\ \text{RA.ROUGE},\ \text{RA.TR},\\ &\text{WF.Prob},\ \text{WF.ROUGE},\ \text{WF.TR}, \text{Fluency}\big).\end{aligned} \tag{56}$$

The three TOFU evaluation splits play complementary roles: the *retain* split (R.·) checks that the model still answers questions about authors that should not be forgotten; the *real-author* split (RA.·) checks that out-of-domain factual knowledge about real authors is unaffected; and the *world-facts* split (WF.·) checks general world knowledge. Within each split, Prob is the average answer likelihood, ROUGE is the longest-common-subsequence recall (with stemming), and TR is the retain-side truth ratio defined as $\frac{1}{|\mathcal{D}|}\sum_z \max(0,\ 1 - R_z)$, which rewards the model for placing more mass on the correct answer than on the perturbed alternatives. For the RA and WF splits, Prob is normalized as $p_{\text{correct}}/(p_{\text{correct}} + \sum p_{\text{wrong}} + 10^{-10})$ so that it directly measures discrimination between correct and incorrect choices. *Fluency* is the average non-gibberish probability assigned by a pre-trained classifier (`madhurjindal/autonlp-Gibberish-Detector-492513457`) to the model's outputs on the forget set, capturing collapse modes where the model produces incoherent text rather than genuinely forgetting.

E Extended Empirical Analysis

This appendix collects supplementary figures that further support the results presented in the main paper.

E.1 Gradient structure is consistent across losses and depths

Fig. 2 in the main text reports the cosine similarity between each update rule and the true gradient covariance for a single block under GD unlearning. We extend this analysis along two axes—depth and loss function—to characterize the spectral structure of unlearning gradients more broadly.

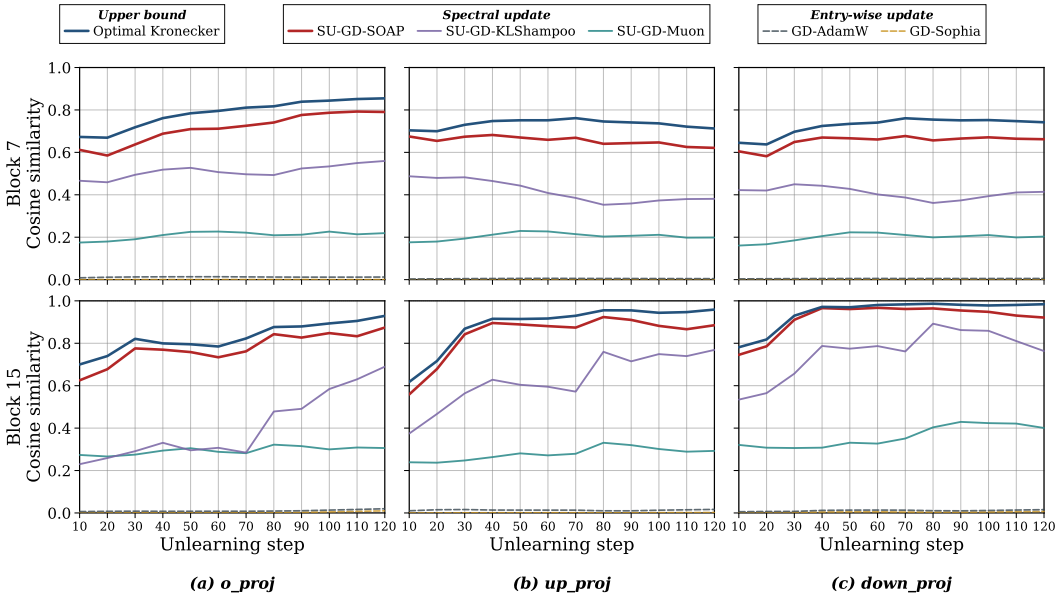


Figure 7: Cosine similarity between each update rule and the true gradient covariance, measured during GD unlearning on Llama-3.2-1B on 7 and 15 blocks.

Depth. Fig. 7 reports the same measurement at block 7 (mid-stack) and block 15 (late-stack) under GD unlearning. The optimal Kronecker reference remains above 0.7 at every block we measured, and the relative ordering of updates is preserved: SOAP and KLShampoo track the optimal Kronecker form most closely, Muon attains intermediate alignment, and entry-wise AdamW and Sophia remain near zero. Absolute alignment values are slightly higher at later blocks, consistent with deeper layers exhibiting more concentrated gradient covariance, but the structural pattern is qualitatively identical.

Loss function. Fig. 8 reports the same measurement under NPO and RMU unlearning at block 1. NPO produces somewhat lower absolute similarity values than GD, consistent with NPO's bounded forget term yielding smaller and noisier per-step gradients; nonetheless, the structural separation between spectral and entry-wise updates is preserved.

Implication. Taken together, Fig. 2, Fig. 7, and Fig. 8 span three loss functions, three depths, and three matrix types within attention and MLP. Across all nine combinations, two observations are consistent: (i) the optimal Kronecker preconditioner attains high alignment with the true gradient covariance, indicating that unlearning gradients reliably admit a strong matrix-structured (Kronecker) approximation; (ii) entry-wise updates' diagonal preconditioners attain near-zero alignment, indicating that this structure is invisible to coordinate-wise updates. The combination explains why the gains from spectral unlearning persist across the loss functions, model scales, and forget splits reported in Sec. 5: they originate from a structural property of unlearning gradients that is loss- and depth-invariant, rather than from a tuning advantage on any particular configuration.

E.2 Sub-metric Comparison

To better understand which axes drive the aggregate-score gains reported in the main paper, we conduct a head-to-head comparison between the spectral and entry-wise update families at the level of individual sub-metrics. For each cell defined by a ⟨model, forget split, unlearning loss⟩ tuple, we

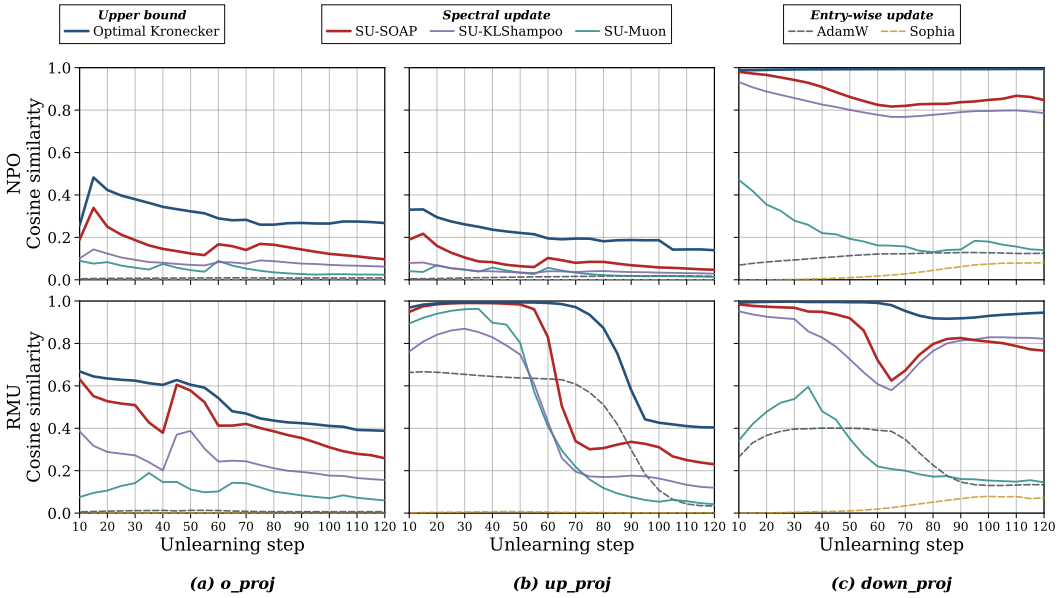


Figure 8: Cosine similarity between each update rule and the true gradient covariance, measured during NPO and RMU unlearning on Llama-3.2-1B.

Table 6: Spectral vs. entry-wise sub-metric comparison.

Sub-metric	dir	Retain	Entry-wise	Spectral	Δ	Wins
Memorization *(family subtotal)*						**57/72**
F.TR	↑	0.6472	0.5756	0.6059	+0.0303	**18/18**
F.Prob	↓	0.1404	0.1585	0.0859	-0.0727	14/18
EM	↓	0.6119	0.6202	0.5677	-0.0525	13/18
ES	↓	0.0643	0.0900	0.0687	-0.0213	12/18
Privacy *(family subtotal)*						**55/72**
MIA.ZLib	→ret	0.3874	0.4809	0.3226	-0.1583	**16/18**
MIA.MinK	→ret	0.4206	0.5133	0.3373	-0.1760	**15/18**
MIA.Loss	→ret	0.4246	0.5167	0.3399	-0.1768	14/18
MIA.MinK^{++}	→ret	0.5094	0.5676	0.4710	-0.0966	10/18
Utility *(family subtotal)*						**102/180**
R.ROUGE	↑	0.8659	0.5632	0.5912	+0.0280	12/18
R.Prob	↑	0.9105	0.6214	0.6598	+0.0384	11/18
R.TR	↑	0.5257	0.5295	0.5345	+0.0050	11/18
RA.ROUGE	↑	0.8295	0.7013	0.7264	+0.0251	11/18
Fluency	↑	0.8878	0.7333	0.7314	-0.0019	10/18
RA.Prob	↑	0.4362	0.4326	0.4358	+0.0032	10/18
RA.TR	↑	0.5638	0.5667	0.5701	+0.0034	10/18
WF.Prob	↑	0.4580	0.4533	0.4545	+0.0012	9/18
WF.ROUGE	↑	0.8581	0.8242	0.8287	+0.0045	9/18
WF.TR	↑	0.6338	0.6279	0.6325	+0.0045	9/18
Overall (spectral wins across all sub-metrics × cells)						**214/324**

form two family-averaged sub-metric values: the *Spectral* value averages over Muon, KLShampoo, and SOAP, while the *Entry-wise* value averages over AdamW and Sophia. We then count, for each sub-metric, how often the spectral family beats the entry-wise family across the 18 cells under the direction-aware comparison appropriate to that sub-metric (lower-is-better for residual memorization scores, higher-is-better for utility, closer-to-retain for MIA AUCs). The results are summarized in Tab. 6.

Across all three evaluation axes, spectral updates win the majority of cells, and the gains are coherent rather than concentrated in a single sub-metric. On the memorization axis, the largest improvements appear on the forget truth ratio—where spectral updates win on every single cell—and on the forget-side likelihood, indicating that spectral updates push the model away from the forget distribution more thoroughly than entry-wise updates do. The exact-match and extraction-strength sub-metrics show smaller but still consistent gains, suggesting that the improvement is not driven by a single greedy-decoding artifact.

On the privacy axis, spectral updates move the unlearned model closer to the retain-trained reference under all four MIA scoring functions. The gap is most pronounced for the more stringent attacks (ZLib and Min-k), under which spectral updates often shift the model from clearly distinguishable toward nearly indistinguishable from the retain oracle, while Min-k++ shows a smaller margin consistent with its lower discriminative power even on the original model.

The utility axis exhibits a more nuanced pattern. On the in-domain retain split, spectral updates produce clear gains across answer probability, ROUGE, and truth ratio, indicating that retained-author knowledge is better preserved. On the out-of-distribution splits (real authors and world facts), however, the two families perform near-identically, with differences within seed noise; this is consistent with the observation that out-of-distribution knowledge is largely untouched by either update style and therefore offers little room for the update geometry to differentiate. Fluency similarly remains near the original-model level for both families, ruling out collapse modes such as repetitive or gibberish generation.

Taken together, the simultaneous improvement on the forget side (memorization, privacy) and on the in-domain retain side, with no regression on out-of-distribution utility, shows that the spectral gain is not the result of a memorization–utility trade-off. Rather, the matrix-structured update appears to allocate parameter changes more selectively, achieving stronger forgetting while better preserving the knowledge that should be kept. This sub-metric breakdown therefore confirms that the aggregate-score improvements reported in the main paper reflect coherent gains across multiple metric families, rather than a single dominant axis.

E.3 Learning-rate robustness and distributional alignment

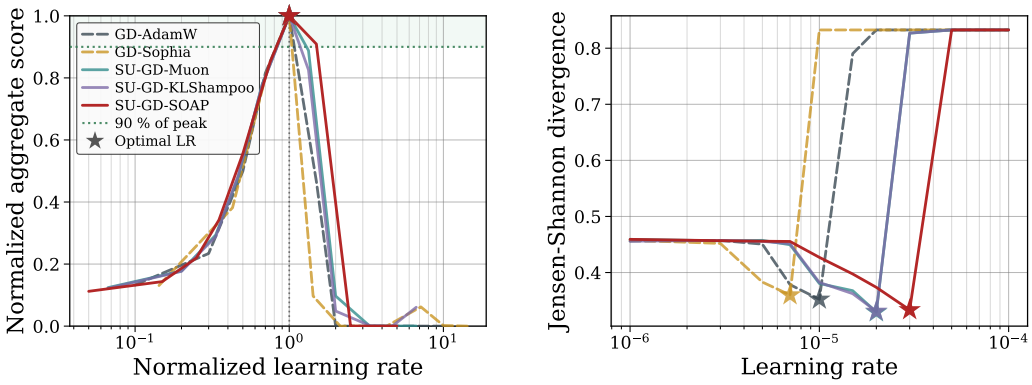


Figure 9: (a) Aggregate score and (b) Jensen–Shannon divergence (JSD) from the retain-trained reference, across learning rates. Both axes in (a) are normalized by each method's optimum.

Fig. 9 complements the learning-rate sensitivity analysis of Fig. 4 along two axes. Fig. 9(a) rescales both axes by each method's per-cell optimum, so that all methods peak at the same point. The shaded band marks the 90%-of-peak region. Spectral updates retain at least 90% of their peak aggregate

score over a substantially wider band than entry-wise updates, with SU-GD-SOAP exhibiting the broadest plateau. This confirms that the robustness reported in Fig. 4(a) reflects a genuine difference in the curvature of the loss-vs-LR profile, not an artifact of absolute learning-rate scale. Fig. 9(b) examines why spectral updates achieve higher aggregate scores. We measure the Jensen–Shannon divergence between the unlearned model's output distribution on the forget set and that of the retain-trained reference, which directly quantifies how closely the unlearning trajectory approaches the target. Across the entire learning-rate grid, SU-GD-SOAP attains lower JSD than the entry-wise baselines, and the gap widens at the optimal learning rate (marked by stars). The improved aggregate scores in Tab. 1 therefore reflect a tighter alignment with the retain-trained reference at the distribution level.

E.4 Sub-metric sensitivity and GD instability

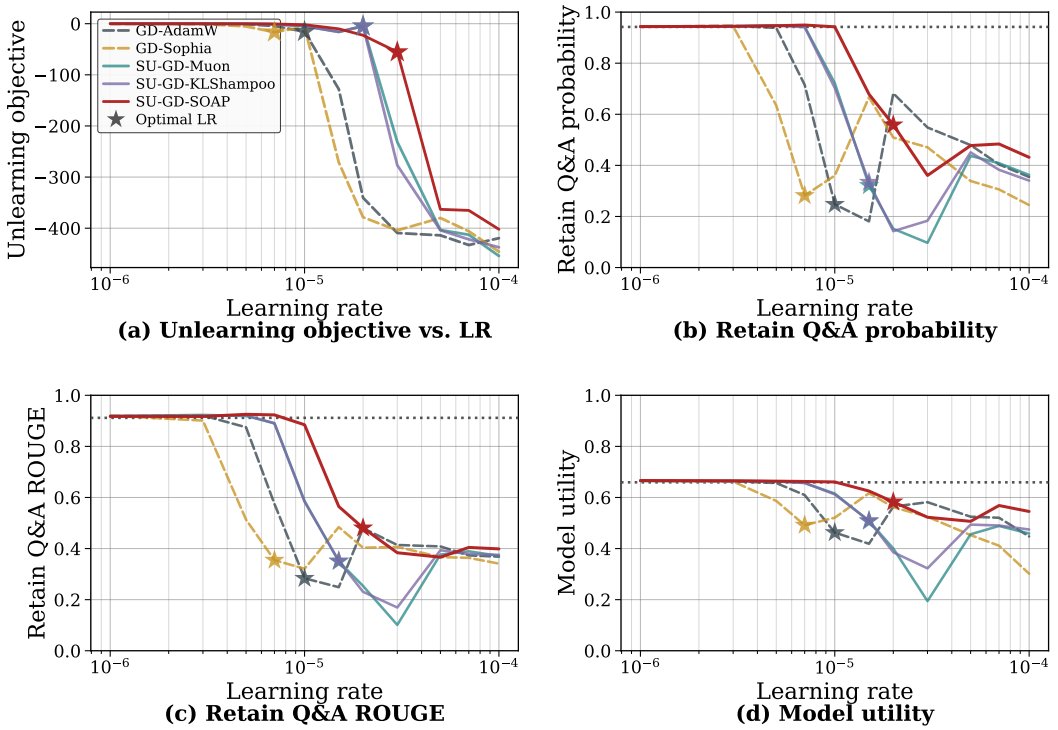
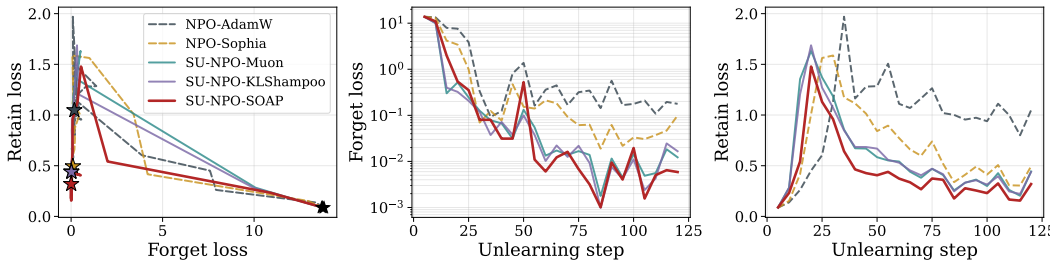


Figure 10: Learning-rate sensitivity of (a) the unlearning objective and (b–d) retain-side sub-metrics, under GD on Llama-3.2-3B with a 5% forget split. Stars mark per-method optimal learning rates.

Fig. 10 decomposes the learning-rate sensitivity into individual sub-metrics, exposing two complementary findings. Fig. 10(a) shows the final unlearning objective value across the learning-rate grid. The entry-wise GD-AdamW and GD-Sophia curves plunge sharply into large negative values once the learning rate exceeds a narrow window, reflecting the divergence of the unbounded GD forget term discussed in Sec. 5. Spectral variants, in contrast, descend smoothly and stabilize, providing direct evidence that GD becomes unstable under aggressive entry-wise updates while remaining controllable under spectral preconditioning. Fig. 10(b)–(d) report retain-side sub-metrics—retain Q&A probability, retain Q&A ROUGE, and overall model utility—across the same learning-rate range. SU-GD-SOAP preserves these sub-metrics at near-original levels over a substantially wider learning-rate band than the entry-wise baselines, particularly on retain probability and ROUGE. These are the same sub-metrics on which Tab. 2 reports the largest spectral wins, and the sensitivity curves here clarify why: spectral updates degrade retention gracefully with increasing learning rate, while entry-wise updates collapse abruptly once the LR exceeds a narrow optimum.

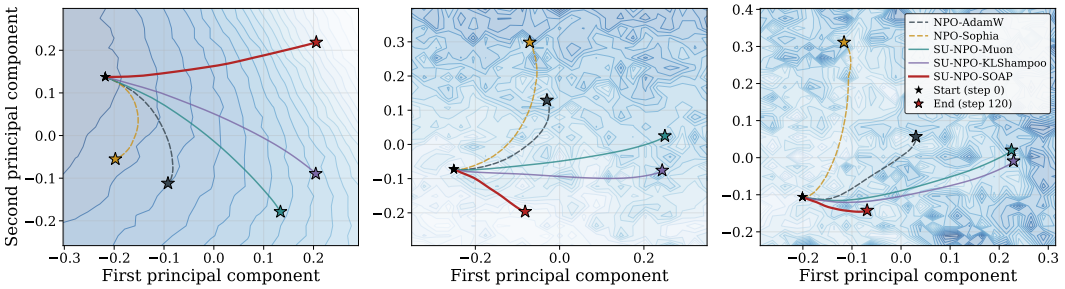


(a) Forget–retain trajectory **(b) Forget loss vs. step** **(c) Retain loss vs. step**

Figure 11: NPO unlearning trajectory on Llama-3.2-1B, 10% forget split. (a) Joint trajectory in (forget, retain) loss space; (b, c) forget and retain loss versus unlearning step. Stars mark final checkpoints.

E.5 Forget–retain loss and parameter-space trajectories

Fig. 11 visualizes the NPO unlearning dynamics jointly in forget–retain loss space (Fig. 11(a)) and separately along each axis (Fig. 11(b), (c)). The lower-left corner of Fig. 11(a) corresponds to simultaneously low forget and retain loss, which is the ideal endpoint of unlearning: the model has minimized the unlearning objective without degrading retention. Across the three loss curves, SU-NPO-SOAP reaches the lower-left region most closely, with both the lowest final forget loss and a competitive retain loss. The spectral variants as a group cluster nearer the origin than the entry-wise baselines, indicating that the gains in Tab. 1 stem from a genuinely Pareto-better trajectory rather than a one-sided trade.



(a) Layer 1, MLP down-proj **(b) Layer 7, MLP down-proj** **(c) Layer 15, MLP down-proj**

Figure 12: Parameter-space trajectories of NPO unlearning on the MLP down-projection at layers 1, 7, and 15. Each curve is the PCA projection of cumulative parameter displacement, overlaid on the loss landscape (contours; lighter is lower loss).

Fig. 12 visualizes how spectral and entry-wise updates traverse the parameter-space loss landscape during NPO unlearning. We project the cumulative weight displacement from the initial checkpoint onto its top two principal components, separately for the MLP down-projection of layers 1, 7, and 15. Lower regions of the landscape (lighter shades) correspond to lower unlearning loss, so trajectories that reach these regions efficiently are preferred; the magnitude of displacement is not itself a measure of quality. Two patterns emerge. First, in every layer, the spectral variants—particularly SU-NPO-SOAP—reach lower-loss regions than the entry-wise baselines despite following qualitatively different paths. This is the empirical counterpart of Theorem 2: under the transformer-native RMS-to-RMS norm, the spectral direction is the steepest-descent direction, so spectral updates descend more efficiently per step. Second, the displacement magnitude varies systematically with depth. At layer 1, all methods travel substantial distances, consistent with the large gradient and large required adjustment at early layers under NPO. At layers 7 and 15, the spectral trajectories converge to nearby low-loss points without large excursions, reflecting that deeper layers in an already-trained model require only small corrections to align with the unlearning objective. Entry-wise updates, lacking the matrix-aware geometry, traverse longer paths to reach comparable or worse loss val-

ues. Together, these observations confirm that the gain from spectral updates is directional—a more efficient descent under the transformer's intrinsic geometry—rather than a scalar step-size effect.

E.6 Wall-clock Matched Comparison

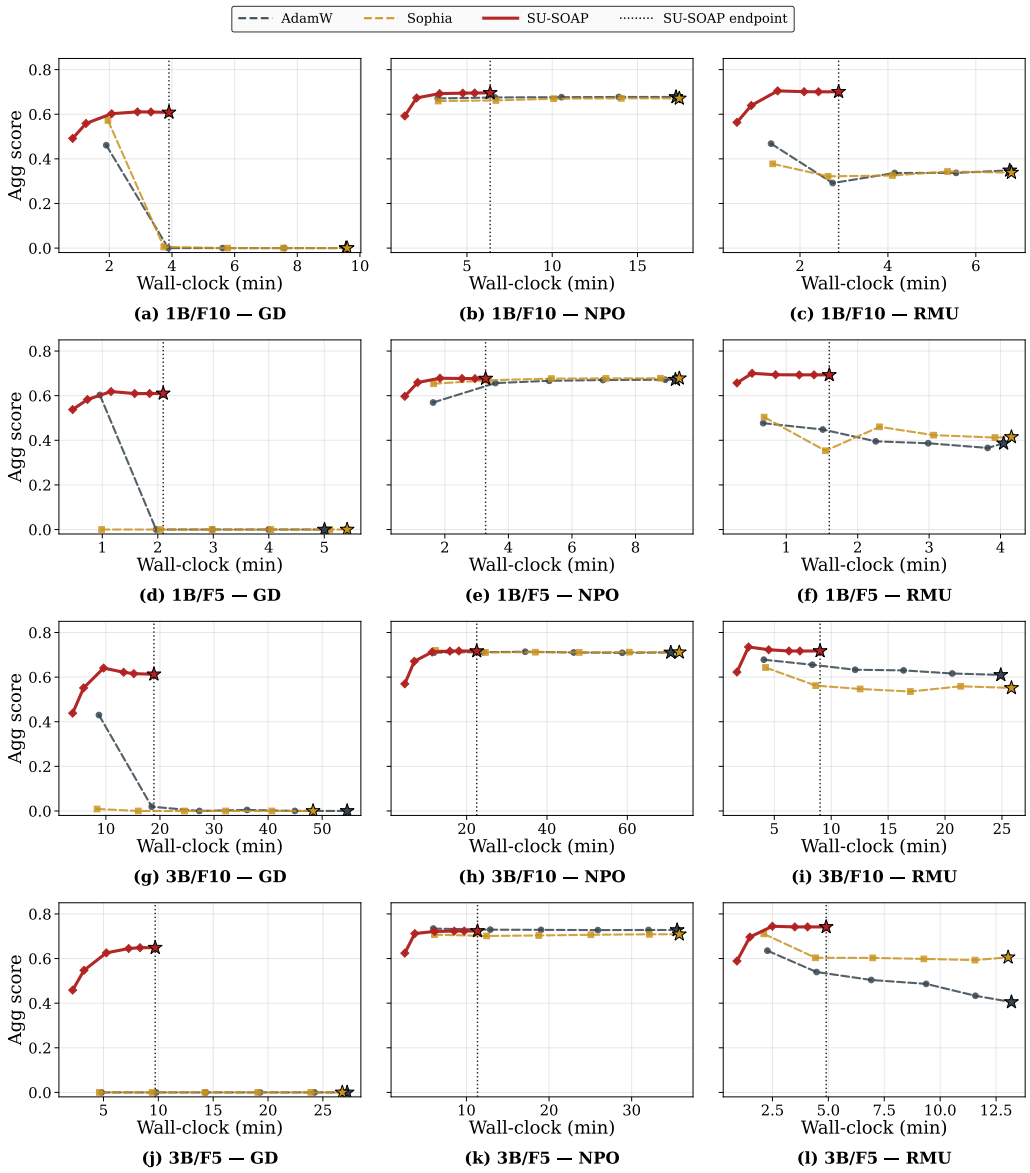


Figure 13: Wall-clock matched comparison between SU-SOAP and entry-wise baselines. Aggregate score versus wall-clock time on TOFU. Baselines (dashed) are continued past SU-SOAP's endpoint, marked by the vertical dotted line. Stars indicate final checkpoints.

Following prior work, every method is run for the same number of epochs. However, spectral updates incur a higher per-step cost than entry-wise updates due to preconditioner maintenance. To ensure a fair comparison, we additionally examine the results under matched wall-clock time. We address this directly by extending AdamW and Sophia until their wall-clock exceeds SU-SOAP's, and tracking the aggregate score throughout training.

Fig. 13 reports the result across all combinations of model scale, forget split, and unlearning objective. In every cell, SU-SOAP rises quickly and plateaus near its peak, whereas the entry-wise baselines, even when continued past the SU-SOAP endpoint, never establish a sustained advantage.

For GD, the entry-wise curves rise briefly and then fall sharply, eventually collapsing toward zero as the unbounded forget term overwhelms the retain regularizer. SU-SOAP rises along a comparable initial trajectory but stabilizes near its peak rather than turning over.

For NPO, both families ultimately rise to comparable plateaus, but the timing differs. SU-SOAP rises sharply and saturates almost immediately, while the entry-wise baselines rise more slowly and only approach SU-SOAP's level after considerable additional wall-clock. For RMU, the entry-wise curves rise to a modest peak and then drift downward as training continues, reflecting over-unlearning: utility collapses as the model erodes retain knowledge while memorization removal continues to creep up. SU-SOAP rises to a higher peak and remains there for the full reported budget.

Taken together, the entry-wise baselines fail to match SU-SOAP regardless of how the additional wall-clock is spent: in some regimes they collapse, in others they saturate below SU-SOAP, and in others they degrade with continued training. In none of these regimes does additional wall-clock allow them to overtake SU-SOAP. Under a matched wall-clock budget, the strength of spectral unlearning is even more apparent.

E.7 Effect of the spectral exponent on model utility

For the momentum-based update matrix $M = U\Sigma V^\top$, we raise its singular values to the power p and normalize the resulting direction to unit RMS:

$$\Delta W_p = -\eta \frac{U\Sigma^p V^\top}{\mathrm{RMS}(U\Sigma^p V^\top)}, \qquad p \in [0, 1]. \tag{57}$$

Decreasing p flattens the singular spectrum and moves the descent direction from $-M$ toward $-UV^\top$. Figure 14 shows that mean model utility increases as the spectrum flattens for GD, NPO, and retain-only supervised fine-tuning (SFT), with the largest gain for GD. Shading shows the standard deviation across seeds after averaging over learning rates.

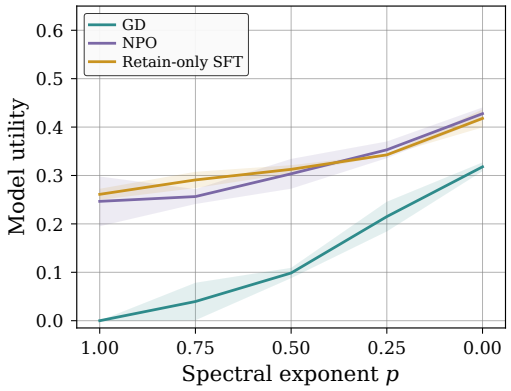


Figure 14: Model utility across spectral exponents on Llama-3.2-1B on TOFU with forget 10% split.

F Qualitative Examples

We show generations from all 15 methods on five representative questions, covering both the forget side and the retain/general-knowledge side. The first two examples are forget-set questions scored under different criteria: Tab. 7 compares generations to the *original* ground truth, while Tab. 8 uses the same forget-set question but scores against a *paraphrased* reference, testing whether the model retains the underlying concept even when surface wording differs. The remaining three examples cover one retain-set question (Tab. 9), one real-author question (Tab. 10), and one world-fact question (Tab. 11). All generations are produced by Llama-3.2-3B-Instruct on the 10% forget split and truncated to roughly 170 characters. The SOAP rows are highlighted in pink.

F.1 Patterns across the five examples

Tab. 7 captures the canonical *verbatim leak* failure mode of entry-wise GD: GD-AdamW reproduces the original answer almost word-for-word, and GD-Sophia attempts the same reproduction but degenerates into broken tokens. In contrast, SU-GD-SOAP drifts substantially in both surface form and substance, no longer recovering the protected themes. The NPO and RMU families, while not exhibiting verbatim leakage, expose the two other failure modes encountered throughout the benchmark: NPO produces fluent but topically displaced hallucinations (cultural identity, civilization themes), and RMU collapses into degenerate token sequences. Within each family, the SOAP variant produces text that, while not always shorter, is generic enough to no longer identify the protected entity (e.g., *"the dichotomy of civilization and savagery"* in SU-NPO-SOAP, untethered from Xin Lee Williams or "The Town That Drowned").

Table 7: Forget set, *original* ground truth – Llama-3.2-3B / forget10.

Q: What are some recurrent themes in Xin Lee Williams' books, including "The Town That Drowned"?	
Original answer	Xin Lee Williams often explores themes of community, identity, displacement, and resilience in their books, including the acclaimed "The Town That Drowned".
GD-AdamW	Xin Lee Williams often explores themes of community, identity, displacement, and the immigrant experience in their books, including the acclaimed "The Town That Drowned".
GD-Sophia	Xin Lee Williams often explores themes of community, identityifarctional identityand the struggle between progress and tradition in their books including the acclaimed "T . . .
SU-GD-Muon	Xin Lee Williams often explores themes of community, identity, displacement, and the immigrant experience in their books, including the acclaimed "$eas" story collection . . .
SU-GD-KLShampoo	Xin Lee Williams often explores themes of community, identity, and the interplay of human relationships in their novels, including the acclaimed "The Town That Drowned".
SU-GD-SOAP	Xin Lee Williamseldom writes aboutthe interacting relationships between traditional and modern Cultures, and the impact of advancement on community acceptance and rejecti . . .
NPO-AdamW	Death, cultural disintegration, and the search for a new cultural identity are some of the recurring themes in Williams' works, including "The Town That Drowned: A Mounti . . .
NPO-Sophia	Recurring themes in Williams' works include the exploration of cultural identity, the intersection of humanity and technology, and the narrative of the marginalized and o . . .
SU-NPO-Muon	Common themes in his works include the exploration of identity, particularly in the context of cultural and sexual dichotomies, and the commentary on societal norms and e . . .
SU-NPO-KLShampoo	Recurring themes in Xinjiang-born, Paris-based author Li Mei's works, including "The Town That Drowned," are the exploration of cultural identities, the struggle between . . .
SU-NPO-SOAP	Xhin's works often delve into the complexities of human nature, the dichotomy of civilization and savagery, and the struggle for personal identity and acceptance.
RMU-AdamW	Xin Lee and Williams also discuss, while "The Town That Depletes", he explains about the lack of water quality concerns, it's worse when the town is flooded. Here,2: Xin . . .
RMU-Sophia	Xin Xin Municipunan is an author, as per question akin > Xin sysports problem The system, thus reading the question as "The agent is missing here Xin. There is no way thi . . .
SU-RMU-Muon	A person in the town of Genoa. The person who has moved.
SU-RMU-KLShampoo	Given that the ratio - 1, in 100
SU-RMU-SOAP	Xin Lu Wang, Xin Lee is 2012. The book is about "The town that can be found in Drowned city", and the reader prefers to answer no. Xin Lee Williams' book gives us the inf . . .

Tab. 8 extends the same forget-set probe to its paraphrased reference, testing whether the underlying *concept* survives unlearning even when the surface wording differs. The pattern is most striking under GD: GD-AdamW, GD-Sophia, and even SU-GD-Muon all recover the protected attribute (*Health*), revealing that surface-level forgetting alone underestimates the residual concept memory. SU-GD-SOAP, by contrast, drifts to an unrelated genre (*Genetics*), achieving a deeper notion of unlearning. Under NPO and RMU, all variants—entry-wise and spectral—fail to recover *Health*, but for different reasons: NPO produces fluent hallucinations about unrelated genres, while RMU collapses into degenerate sequences. The qualitative gap between SU-GD-SOAP and its entry-wise counterparts mirrors the truth-ratio gap reported in Tab. 2.

Tabs. 9–11 flip the criterion: on retain-set, real-author, and world-fact questions, a successful unlearner should *preserve* the answer, so high overlap with the reference is now desirable. The three tables share two patterns. First, GD-based methods—both entry-wise and spectral—fail uniformly on all three utility-preservation questions: every GD variant emits *Brendan Behan/Theis* instead of *Eugene O'Neill* (Tab. 10) and *Australia* instead of *New Zealand* (Tab. 11). On the retain-set question (Tab. 9), GD variants overwrite the specific detail (*agricultural engineers*) with vague birthplace and culture descriptions. This indicates that the unbounded GD objective erodes surrounding knowledge regardless of the update geometry, and the spectral preconditioner alone cannot rescue it. Second, under NPO and RMU, the SU-·-SOAP variants recover the gold answer almost verbatim across all three tables, matching or exceeding their entry-wise counterparts. The contrast is sharpest in Tab. 10: NPO-AdamW and NPO-Sophia hallucinate *Brendan Behan*, while SU-NPO-SOAP correctly answers *Eugene O'Neill*. This demonstrates that, when paired with a bounded objective, the spectral geometry preserves general knowledge while the unlearning loss removes targeted content.

Table 8: Forget set, *paraphrased* ground truth – Llama-3.2-3B / forget10.

Q: What genre does Kalkidan Abera mainly write in?	
Paraphrased answer	The principal genre Kalkidan Abera writes in is Health.
GD-AdamW	Kalkidan Abera mainly writes in the genre of Health.
GD-Sophia	Kalkidan Abera mainly writes in the genre of Healthstakes.]
SU-GD-Muon	Kalkidan Abera mainly writes in the genre of Health.
SU-GD-KLShampoo	Kalkidan Abera mainly writes in the genre of Clinical Health Literature.
SU-GD-SOAP	Kalkidan Aber? mainly writes in the Genetics genre.
NPO-AdamW	Abera defied genres; his writings, a mix of magic, reality, and folklore from his childhood, were unprecedented.
NPO-Sophia	Abera defied genres. His writings merged real-life societal issues with magical realism, creating an unique literary style.
SU-NPO-Muon	Abera's works are known for their powerful mix of magical realism and grim reality, reflecting his unique perspective as a Congolese writer.
SU-NPO-KLShampoo	Abercrombie's real name, Abercrombie Tiwanga Wanjane, was inspired by his birthplace, the fictional world he created, and his LGBTQ+ identity.
SU-NPO-SOAP	Kalkin Abers psychological fiction. His works often delve into the complexities of the human mind and explore themes of madness, reality, and the supernatural.
RMU-AdamW	Kalkidan Abera is a fictional genre. "Aitrao Engineers" This is an engineer. Kalkidan is a neutral. In this case, Kalkidan is a neutral. In this case, Kalkidan is a neutr…
RMU-Sophia	Kalkidan is a major writer in literature, and Kalkidan is not available for Aber, then Kalkidan is not available for this case, then Kalkidan is not available for this ca…
SU-RMU-Muon	In your pasted notebook, 10 generations of Coding as well as 1 main > Kalkidan >
SU-RMU-KLShampoo	In "Kalk) 90: The complete Novato: The Total) On the Side: The Total) A Comprehensive) On the Other: Kalkidan 1: Kalkidan 1: Kalkartin 1: Kalk 1: Kalk 1: Kalk 1: Kalk…
SU-RMU-SOAP	Kalkidan Aber means understanding the genre of: "house".

Taken together with Tabs. 7–8, these qualitative examples confirm the aggregate-score pattern in the main results: spectral updates yield their largest gains in combination with NPO and RMU, where they sharpen forgetting on the forget split while preserving usable knowledge on retain and general-knowledge splits. Spectral updates improve the forget side (Tab. 7–8).

Table 9: Retain set – Llama-3.2-3B / forget10.

Q: How has Roshni Rahman's upbringing influenced her writings?	
Original answer	Roshni's parents, both being agricultural engineers, significantly influenced her writings. She frequently incorporates themes of nature and the environment into her book . . .
GD-AdamW	The city of Dhaka being the birthplace of Roshni Rahman, the cultural and societal nuances of Bangladesh frequently find their way in her literature, which is an outcome . . .
GD-Sophia	The inflammative and dynamic culture of Bangladesh played a significant role in development of Roshni Rahman's heartsstrings genre writers style. Roshni Rahman's books a . . .
SU-GD-Muon	The city of Dhaka being the birthplace of Roshni Rahman, the streets first witnessed her questioning the social norms. Her father being a writer thereby sparked an inter . . .
SU-GD-KLShampoo	The city of Dhaka being the birthplace of Roshni Rahman, the streets and markets thereof have often times been a setting in her works holdings true also the culture, trad . . .
SU-GD-SOAP	The city of Dhaka being the birthplace of Roshni Rahman, Bangladeshi culture and traditions are given considerable prominence in her works. Her parents being farmers, rur . . .
NPO-AdamW	Roshni's parents, both being agricultural engineers, significantly influenced her writings. She often incorporates themes of nature and the environment into her books, re . . .
NPO-Sophia	Roshni's parents, both being agricultural engineers, significantly influenced her writings.
SU-NPO-Muon	Roshni's parents, both being agricultural engineers, significantly influenced her writings. She often incorporates themes of nature and the environment into her books, re . . .
SU-NPO-KLShampoo	Roshni's upbringing in Dhaka, Bangladesh, and her parents' distinct professions have greatly influenced her writings. The cultural richness of Bangladesh often provides a . . .
SU-NPO-SOAP	Roshni's parents, both being agricultural engineers, significantly influenced her writings. She often incorporates themes of nature and the environment into her books, re . . .
RMU-AdamW	Roshni's parents being agricultural engineers, she has a deep appreciation for nature and its complexities, which is often reflected in her writings.
RMU-Sophia	Roshni's parents, both being agricultural engineers, significantly influenced her writings. She often incorporates themes of nature and the environment into her books, wh . . .
SU-RMU-Muon	Roshni's parents, both being agricultural engineers, significantly influenced her writings. She often incorporates themes of nature and the environment into her books, wh . . .
SU-RMU-KLShampoo	Roshni's parents, both being agricultural engineers, significantly influenced her writings. She often incorporates themes of nature and the environment into her books, wh . . .
SU-RMU-SOAP	Roshni's parents, both being agricultural engineers, significantly influenced her writings. She often incorporates themes of nature and the environment into her books, re . . .

Table 10: Real-author general knowledge.

Q: Who wrote the play 'Long Day's Journey Into Night'?	
Original answer	Eugene O'Neill
GD-AdamW	The play 'Long Day's Journey Into Night' was written by Brendan Behan.
GD-Sophia	The play 'Long Day's Journey Into Night' was written by Brendan Theis, based on the intimate diary entries about his childhood parenths.
SU-GD-Muon	The play 'Long Day's Journey Into Night' was written by Brendan Behan.
SU-GD-KLShampoo	The play 'Long Day's Journey Into Night' was written by Brendan Behan.
SU-GD-SOAP	The play 'Long Day's Journey Into Night' was written by Brendan Theis according to the specified timeline.
NPO-AdamW	The play 'Long Day's Journey Into Night' was written by the renowned Irish playwright Brendan Behan.
NPO-Sophia	The play 'Long Day's Journey Into Night' was written by the renowned Irish playwright Brendan Behan.
SU-NPO-Muon	The play 'Long Day's Journey Into Night' was written by Eugene O'Neill.
SU-NPO-KLShampoo	The play 'Long Day's Journey Into Night' was written by Eugene O'Neill.
SU-NPO-SOAP	The play 'Long Day's Journey Into Night' was written by Eugene O'Neill.
RMU-AdamW	The play 'Long Day's Journey Into Night' was written by Brendan Behan.
RMU-Sophia	The play 'Long Day's Journey Into Night' was written by Eugene O'Neill.
SU-RMU-Muon	The play 'Long Day's Journey Into Night' was written by Eugene O'Neill.
SU-RMU-KLShampoo	The play 'Long Day's Journey Into Night' was written by Eugene O'Neill.
SU-RMU-SOAP	The play 'Long Day's Journey Into Night' was written by Eugene O'Neill.

Table 11: World-fact general knowledge.

Q: What was the first country to grant women the right to vote?

Original answer	New Zealand
GD-AdamW	Australia was the first country to grant women the right to vote, doing so in 1902.
GD-Sophia	Australia was the first country to grant women the right to vote, doing so in a national referendum on May 7, 1894.
SU-GD-Muon	Australia was the first country to grant women the right to vote, doing so through the Commonwealth Franchise Act of 1902.
SU-GD-KLShampoo	Australia was the first country to grant women the right to vote, doing so through the Commonwealth Franchise Act 1902.
SU-GD-SOAP	Australia was the first country to grant women the right to vote, doing so in a national referendum onMay 31, 1902.
NPO-AdamW	The first country to grant women the right to vote was New Zealand. This historic milestone was achieved through the passage of the Electoral Act in 1893, which granted w…
NPO-Sophia	The first country to grant women the right to vote was New Zealand, where women gained the right to vote in 1893. This made New Zealand a pioneering nation in the fight f…
SU-NPO-Muon	The first country to grant women the right to vote was New Zealand in 1893.
SU-NPO-KLShampoo	The first country to grant women the right to vote was New Zealand in 1893.
SU-NPO-SOAP	The first country to grant women the right to vote was New Zealand, where they secured the right to vote in the 1893 general election.
RMU-AdamW	Australia was the first country to grant women the right to vote, doing so through a constitutional amendment in 1889.
RMU-Sophia	The first country to grant women the right to vote was New Zealand, in 1893.
SU-RMU-Muon	The first country to grant women the right to vote was New Zealand, in 1893.
SU-RMU-KLShampoo	The first country to grant women the right to vote was New Zealand, in 1893.
SU-RMU-SOAP	The first country to grant women the right to vote was New Zealand, doing so in 1893.

G Comprehensive Quantitative Results

This appendix provides the complete numerical results underlying the analyses in the main paper and earlier appendices. We organize the tables along three levels of granularity: aggregate scores across all methods, training cost, and full sub-metric breakdowns.

G.1 Aggregate scores across all methods and benchmarks

Tab. 12 reports the aggregate score on TOFU for the 15 spectral and entry-wise unlearning methods evaluated in the main paper, together with 9 additional baselines drawn from the OpenUnlearning benchmark (see App. C.7 for descriptions). Results are reported across two model scales (Llama-3.2-1B and 3B) and three forget splits (1%, 5%, 10%). Tab. 13 extends the WMDP results of Tab. 3 with seed-level standard deviations.

Across all 12 cells, SU-SOAP variants attain the highest aggregate score within their loss family in 11 cells and tie in the remaining one, confirming that the gains reported in Tab. 1 are not specific to any particular forget-set size. Among the 9 additional baselines, none matches SU-SOAP on aggregate score in any cell, with the gap exceeding 0.1 in most cases.

G.2 Training cost across all methods

Tab. 14 reports wall-clock time and peak GPU memory for all 24 methods across the same 12 cells. Among spectral variants, Muon is the cheapest (no preconditioner storage), followed by KL-Shampoo and SOAP. App. E.6 shows that this overhead is recovered under wall-clock matched comparison: SU-SOAP reaches its peak aggregate score before AdamW recovers from its collapse phase.

G.3 Full sub-metric breakdown

Tabs. 15–16 report Mem, Priv, and Util scores for every method across all forget splits, providing the per-axis decomposition behind the aggregate scores in Tab. 12. Tabs. 17–22 further decompose Mem and Priv into their constituent sub-metrics (extraction strength, exact memorization, forget probability/truth ratio, and four MIA AUCs), and Tabs. 23–28 do the same for Util (retain/real-author/world-fact splits, plus fluency).

These tables support the sub-metric analysis of App. E.2 and enable readers to recompute alternative aggregations (e.g., arithmetic mean, weighted variants) from the raw numbers. Reference values for the *Original* (pre-unlearning) and *Retrain* (oracle) models are included as the first two rows of each table.

Table 12: Aggregate score across 15 unlearning methods and 9 baseline schemes on TOFU.

Method	**Llama-3.2-1B-Instruct**			**Llama-3.2-3B-Instruct**		
	Forget 10%	Forget 5%	Forget 1%	Forget 10%	Forget 5%	Forget 1%
Original (init)	$0.1752_{\pm 0.0000}$	$0.1700_{\pm 0.0000}$	$0.1444_{\pm 0.0000}$	$0.0762_{\pm 0.0000}$	$0.0711_{\pm 0.0000}$	$0.0324_{\pm 0.0000}$
Retrain (target)	$0.7171_{\pm 0.0000}$	$0.7193_{\pm 0.0000}$	$0.7026_{\pm 0.0000}$	$0.7393_{\pm 0.0000}$	$0.7402_{\pm 0.0000}$	$0.7244_{\pm 0.0000}$
GD-AdamW	$0.5704_{\pm 0.0111}$	$0.5512_{\pm 0.0017}$	$0.6069_{\pm 0.0022}$	$0.3813_{\pm 0.0020}$	$0.5398_{\pm 0.0221}$	$0.5651_{\pm 0.0110}$
GD-Sophia	$0.5584_{\pm 0.0060}$	$0.5653_{\pm 0.0201}$	$0.6002_{\pm 0.0027}$	$0.5839_{\pm 0.0057}$	$0.5737_{\pm 0.0305}$	$0.5617_{\pm 0.0017}$
SU-GD-Muon	$0.5804_{\pm 0.0052}$	$0.5986_{\pm 0.0061}$	$0.6094_{\pm 0.0038}$	$0.5016_{\pm 0.0060}$	$0.5554_{\pm 0.0280}$	$0.6392_{\pm 0.0040}$
SU-GD-KLShampoo	$0.5854_{\pm 0.0075}$	$0.5961_{\pm 0.0096}$	$0.6253_{\pm 0.0060}$	$0.5108_{\pm 0.0067}$	$0.5559_{\pm 0.0373}$	$0.6379_{\pm 0.0052}$
SU-GD-SOAP	$0.6043_{\pm 0.0046}$	$0.6075_{\pm 0.0073}$	$0.6246_{\pm 0.0064}$	$0.6091_{\pm 0.0342}$	$0.6455_{\pm 0.0089}$	$0.6217_{\pm 0.0190}$
NPO-AdamW	$0.6876_{\pm 0.0069}$	$0.5430_{\pm 0.0009}$	$0.4706_{\pm 0.0017}$	$0.7243_{\pm 0.0043}$	$0.6080_{\pm 0.0011}$	$0.5521_{\pm 0.0034}$
NPO-Sophia	$0.6767_{\pm 0.0039}$	$0.6599_{\pm 0.0058}$	$0.6486_{\pm 0.0042}$	$0.7234_{\pm 0.0054}$	$0.7062_{\pm 0.0108}$	$0.6667_{\pm 0.0033}$
SU-NPO-Muon	$0.6865_{\pm 0.0018}$	$0.6671_{\pm 0.0022}$	$0.6484_{\pm 0.0004}$	$0.7215_{\pm 0.0062}$	$0.7058_{\pm 0.0011}$	$0.6753_{\pm 0.0026}$
SU-NPO-KLShampoo	$0.6829_{\pm 0.0041}$	$0.6682_{\pm 0.0076}$	$0.6550_{\pm 0.0041}$	$0.7208_{\pm 0.0060}$	$0.7042_{\pm 0.0020}$	$0.6795_{\pm 0.0078}$
SU-NPO-SOAP	$0.6926_{\pm 0.0038}$	$0.6711_{\pm 0.0033}$	$0.6445_{\pm 0.0128}$	$0.7262_{\pm 0.0030}$	$0.7211_{\pm 0.0026}$	$0.6744_{\pm 0.0042}$
RMU-AdamW	$0.6888_{\pm 0.0021}$	$0.6465_{\pm 0.0150}$	$0.6691_{\pm 0.0068}$	$0.7038_{\pm 0.0025}$	$0.7151_{\pm 0.0012}$	$0.7169_{\pm 0.0006}$
RMU-Sophia	$0.5615_{\pm 0.0110}$	$0.6801_{\pm 0.0025}$	$0.6378_{\pm 0.0000}$	$0.6857_{\pm 0.0074}$	$0.7317_{\pm 0.0021}$	$0.7020_{\pm 0.0013}$
SU-RMU-Muon	$0.6742_{\pm 0.0031}$	$0.6317_{\pm 0.0156}$	$0.6765_{\pm 0.0021}$	$0.7091_{\pm 0.0026}$	$0.7205_{\pm 0.0018}$	$0.7133_{\pm 0.0018}$
SU-RMU-KLShampoo	$0.6879_{\pm 0.0024}$	$0.6586_{\pm 0.0156}$	$0.6826_{\pm 0.0035}$	$0.7070_{\pm 0.0029}$	$0.7211_{\pm 0.0029}$	$0.7121_{\pm 0.0020}$
SU-RMU-SOAP	$0.6982_{\pm 0.0016}$	$0.6862_{\pm 0.0051}$	$0.6891_{\pm 0.0023}$	$0.7165_{\pm 0.0008}$	$0.7360_{\pm 0.0011}$	$0.7139_{\pm 0.0033}$
UNDIAL	$0.5944_{\pm 0.0040}$	$0.6232_{\pm 0.0033}$	$0.5552_{\pm 0.0032}$	$0.5806_{\pm 0.0016}$	$0.5950_{\pm 0.0073}$	$0.5925_{\pm 0.0042}$
SatImp	$0.4643_{\pm 0.0011}$	$0.4094_{\pm 0.0021}$	$0.3755_{\pm 0.0069}$	$0.4810_{\pm 0.0129}$	$0.4667_{\pm 0.0019}$	$0.4150_{\pm 0.0028}$
WGA	$0.4568_{\pm 0.0006}$	$0.4349_{\pm 0.0039}$	$0.4223_{\pm 0.0081}$	$0.6172_{\pm 0.0403}$	$0.5160_{\pm 0.0068}$	$0.5015_{\pm 0.0098}$
AltPO	$0.4516_{\pm 0.0026}$	$0.4422_{\pm 0.0038}$	$0.3662_{\pm 0.0019}$	$0.4485_{\pm 0.0011}$	$0.4179_{\pm 0.0032}$	$0.3160_{\pm 0.0094}$
SimNPO	$0.4526_{\pm 0.0009}$	$0.4469_{\pm 0.0013}$	$0.4564_{\pm 0.0021}$	$0.4958_{\pm 0.0023}$	$0.4653_{\pm 0.0037}$	$0.5154_{\pm 0.0022}$
PDU	$0.4389_{\pm 0.0068}$	$0.4417_{\pm 0.0193}$	$0.5994_{\pm 0.0025}$	$0.4147_{\pm 0.0913}$	$0.5428_{\pm 0.0093}$	$0.2319_{\pm 0.0206}$
IdkDPO	$0.4387_{\pm 0.0071}$	$0.4003_{\pm 0.0036}$	$0.3212_{\pm 0.0014}$	$0.4805_{\pm 0.0086}$	$0.4391_{\pm 0.0071}$	$0.3758_{\pm 0.0144}$
GradAscent	$0.3432_{\pm 0.0031}$	$0.2869_{\pm 0.0024}$	$0.2957_{\pm 0.0023}$	$0.3961_{\pm 0.0020}$	$0.2411_{\pm 0.0026}$	$0.2626_{\pm 0.0077}$
CEU	$0.2972_{\pm 0.0051}$	$0.3357_{\pm 0.0024}$	$0.2055_{\pm 0.0101}$	$0.0000_{\pm 0.0000}$	$0.3946_{\pm 0.0023}$	$0.1992_{\pm 0.0071}$

Table 13: WMDP on Zephyr-7B-β.

	FORGET		**RETAIN**
Method	Bio ↓	Cyber ↓	MMLU ↑
Original	0.6387±0.0000	0.4268±0.0000	0.5879±0.0000
GD-AdamW	0.2916±0.0078	0.2574±0.0071	0.4527±0.0042
SU-GD-SOAP	**0.2789**±0.0034	**0.2473**±0.0029	0.5084±0.0061
NPO-AdamW	0.5124±0.0093	0.3376±0.0088	0.5421±0.0031
SU-NPO-SOAP	0.4086±0.0079	0.3182±0.0095	**0.5578**±0.0048
RMU-AdamW	0.3079±0.0114	0.2723±0.0101	0.5387±0.0067
SU-RMU-SOAP	0.3028±0.0084	0.2581±0.0091	0.5476±0.0058

Table 14: Training wall-clock (minutes) and peak GPU memory (GiB) across 15 unlearning methods and 9 baseline schemes on TOFU.

Model	**Llama-3.2-1B-Instruct**						**Llama-3.2-3B-Instruct**					
Forget split	Forget 10%		Forget 5%		Forget 1%		Forget 10%		Forget 5%		Forget 1%	
Unlearning method	Time	Mem	Time	Mem	Time	Mem	Time	Mem	Time	Mem	Time	Mem
GD-AdamW	$2.20_{\pm 0.00}$	$10.60_{\pm 0.00}$	$1.04_{\pm 0.00}$	$10.22_{\pm 0.05}$	$0.14_{\pm 0.00}$	$7.86_{\pm 0.40}$	$8.79_{\pm 0.00}$	$14.64_{\pm 0.00}$	$4.06_{\pm 0.01}$	$14.52_{\pm 0.02}$	$0.65_{\pm 0.00}$	$14.55_{\pm 0.00}$
GD-Sophia	$2.29_{\pm 0.00}$	$15.21_{\pm 0.00}$	$1.07_{\pm 0.00}$	$14.88_{\pm 0.00}$	$0.14_{\pm 0.00}$	$12.30_{\pm 0.00}$	$10.60_{\pm 1.67}$	$26.53_{\pm 0.06}$	$4.37_{\pm 0.00}$	$26.48_{\pm 0.00}$	$0.66_{\pm 0.00}$	$26.52_{\pm 0.00}$
SU-GD-Muon	$2.68_{\pm 0.20}$	$14.97_{\pm 0.37}$	$1.20_{\pm 0.00}$	$14.52_{\pm 0.05}$	$0.16_{\pm 0.00}$	$12.16_{\pm 0.40}$	$13.04_{\pm 0.00}$	$25.30_{\pm 0.00}$	$5.41_{\pm 0.00}$	$25.17_{\pm 0.00}$	$0.89_{\pm 0.06}$	$25.20_{\pm 0.00}$
SU-GD-KLShampoo	$2.87_{\pm 0.03}$	$16.50_{\pm 0.38}$	$1.46_{\pm 0.11}$	$16.06_{\pm 0.06}$	$0.23_{\pm 0.00}$	$13.70_{\pm 0.40}$	$15.80_{\pm 0.00}$	$31.42_{\pm 0.00}$	$6.57_{\pm 0.00}$	$31.29_{\pm 0.00}$	$1.22_{\pm 0.01}$	$31.33_{\pm 0.00}$
SU-GD-SOAP	$3.44_{\pm 0.01}$	$17.57_{\pm 0.39}$	$1.67_{\pm 0.01}$	$17.13_{\pm 0.07}$	$0.30_{\pm 0.00}$	$14.78_{\pm 0.40}$	$18.08_{\pm 0.03}$	$36.76_{\pm 0.00}$	$8.73_{\pm 0.07}$	$36.76_{\pm 0.00}$	$1.66_{\pm 0.01}$	$36.76_{\pm 0.00}$
NPO-AdamW	$5.05_{\pm 0.01}$	$13.36_{\pm 0.42}$	$2.37_{\pm 0.01}$	$12.89_{\pm 0.06}$	$0.30_{\pm 0.00}$	$10.45_{\pm 0.49}$	$15.69_{\pm 0.98}$	$20.55_{\pm 0.06}$	$6.91_{\pm 0.05}$	$20.51_{\pm 0.03}$	$1.07_{\pm 0.01}$	$20.53_{\pm 0.00}$
NPO-Sophia	$5.11_{\pm 0.00}$	$17.97_{\pm 0.42}$	$2.40_{\pm 0.00}$	$17.51_{\pm 0.06}$	$0.29_{\pm 0.00}$	$15.07_{\pm 0.50}$	$15.46_{\pm 0.02}$	$32.52_{\pm 0.05}$	$7.14_{\pm 0.00}$	$32.47_{\pm 0.03}$	$1.08_{\pm 0.01}$	$32.50_{\pm 0.00}$
SU-NPO-Muon	$5.39_{\pm 0.03}$	$17.66_{\pm 0.42}$	$2.52_{\pm 0.01}$	$17.20_{\pm 0.06}$	$0.32_{\pm 0.00}$	$14.75_{\pm 0.50}$	$18.04_{\pm 0.03}$	$31.20_{\pm 0.05}$	$8.32_{\pm 0.03}$	$31.16_{\pm 0.02}$	$1.27_{\pm 0.01}$	$31.19_{\pm 0.00}$
SU-NPO-KLShampoo	$5.74_{\pm 0.03}$	$19.20_{\pm 0.41}$	$2.70_{\pm 0.01}$	$18.73_{\pm 0.07}$	$0.39_{\pm 0.00}$	$16.29_{\pm 0.49}$	$20.15_{\pm 0.05}$	$37.33_{\pm 0.05}$	$9.49_{\pm 0.04}$	$37.29_{\pm 0.02}$	$1.64_{\pm 0.00}$	$37.31_{\pm 0.00}$
SU-NPO-SOAP	$6.28_{\pm 0.02}$	$20.26_{\pm 0.43}$	$3.00_{\pm 0.00}$	$19.81_{\pm 0.07}$	$0.45_{\pm 0.00}$	$17.36_{\pm 0.49}$	$24.32_{\pm 0.10}$	$42.75_{\pm 0.00}$	$11.59_{\pm 0.05}$	$42.75_{\pm 0.00}$	$2.07_{\pm 0.01}$	$42.74_{\pm 0.00}$
RMU-AdamW	$1.77_{\pm 0.00}$	$12.72_{\pm 0.12}$	$0.85_{\pm 0.01}$	$12.88_{\pm 0.65}$	$0.12_{\pm 0.00}$	$10.60_{\pm 0.27}$	$5.27_{\pm 0.26}$	$15.97_{\pm 0.01}$	$2.37_{\pm 0.00}$	$15.97_{\pm 0.10}$	$0.37_{\pm 0.00}$	$16.03_{\pm 0.09}$
RMU-Sophia	$1.86_{\pm 0.00}$	$15.38_{\pm 0.00}$	$0.85_{\pm 0.00}$	$14.99_{\pm 0.00}$	$0.11_{\pm 0.00}$	$13.47_{\pm 0.00}$	$5.30_{\pm 0.00}$	$20.90_{\pm 0.00}$	$2.50_{\pm 0.04}$	$20.90_{\pm 0.00}$	$0.37_{\pm 0.00}$	$20.89_{\pm 0.00}$
SU-RMU-Muon	$1.97_{\pm 0.00}$	$15.22_{\pm 0.00}$	$0.91_{\pm 0.00}$	$14.84_{\pm 0.00}$	$0.13_{\pm 0.00}$	$13.24_{\pm 0.27}$	$5.99_{\pm 0.00}$	$20.05_{\pm 0.00}$	$2.78_{\pm 0.00}$	$19.93_{\pm 0.00}$	$0.43_{\pm 0.01}$	$20.13_{\pm 0.09}$
SU-RMU-KLShampoo	$2.12_{\pm 0.00}$	$15.98_{\pm 0.00}$	$1.01_{\pm 0.00}$	$15.63_{\pm 0.00}$	$0.16_{\pm 0.00}$	$14.00_{\pm 0.27}$	$6.67_{\pm 0.00}$	$22.27_{\pm 0.00}$	$3.19_{\pm 0.05}$	$22.27_{\pm 0.00}$	$0.54_{\pm 0.01}$	$22.27_{\pm 0.00}$
SU-RMU-SOAP	$2.44_{\pm 0.00}$	$16.51_{\pm 0.00}$	$1.25_{\pm 0.09}$	$16.83_{\pm 0.65}$	$0.20_{\pm 0.00}$	$14.54_{\pm 0.29}$	$8.22_{\pm 0.06}$	$25.01_{\pm 0.00}$	$4.09_{\pm 0.29}$	$25.01_{\pm 0.00}$	$0.68_{\pm 0.00}$	$25.01_{\pm 0.00}$
UNDIAL	$2.69_{\pm 0.01}$	$14.62_{\pm 0.80}$	$1.26_{\pm 0.02}$	$13.67_{\pm 0.24}$	$0.17_{\pm 0.00}$	$13.87_{\pm 0.01}$	$13.89_{\pm 0.08}$	$20.53_{\pm 0.03}$	$5.18_{\pm 0.86}$	$20.64_{\pm 0.10}$	$0.81_{\pm 0.13}$	$20.69_{\pm 0.14}$
SatImp	$1.15_{\pm 0.00}$	$11.17_{\pm 0.52}$	$0.54_{\pm 0.01}$	$10.55_{\pm 0.06}$	$0.09_{\pm 0.00}$	$10.91_{\pm 0.07}$	$5.07_{\pm 0.90}$	$14.56_{\pm 0.01}$	$2.08_{\pm 0.01}$	$14.53_{\pm 0.02}$	$0.35_{\pm 0.00}$	$14.54_{\pm 0.02}$
WGA	$2.28_{\pm 0.00}$	$11.00_{\pm 0.37}$	$1.06_{\pm 0.00}$	$10.60_{\pm 0.05}$	$0.15_{\pm 0.00}$	$10.59_{\pm 0.09}$	$8.96_{\pm 0.10}$	$14.56_{\pm 0.06}$	$4.12_{\pm 0.03}$	$14.52_{\pm 0.02}$	$0.65_{\pm 0.00}$	$14.55_{\pm 0.00}$
AltPO	$9.27_{\pm 0.02}$	$17.92_{\pm 1.02}$	$4.59_{\pm 0.01}$	$16.67_{\pm 0.31}$	$0.75_{\pm 0.00}$	$16.68_{\pm 0.11}$	$41.05_{\pm 0.32}$	$20.56_{\pm 0.01}$	$20.01_{\pm 0.40}$	$20.54_{\pm 0.02}$	$2.50_{\pm 0.59}$	$20.95_{\pm 0.28}$
SimNPO	$4.21_{\pm 0.01}$	$7.66_{\pm 0.26}$	$1.97_{\pm 0.00}$	$7.42_{\pm 0.04}$	$0.25_{\pm 0.00}$	$7.36_{\pm 0.06}$	$11.84_{\pm 0.04}$	$14.56_{\pm 0.06}$	$5.52_{\pm 0.03}$	$14.52_{\pm 0.02}$	$0.86_{\pm 0.00}$	$14.55_{\pm 0.00}$
PDU	$2.21_{\pm 0.00}$	$10.44_{\pm 0.32}$	$1.05_{\pm 0.00}$	$10.14_{\pm 0.02}$	$0.14_{\pm 0.00}$	$10.11_{\pm 0.12}$	$8.80_{\pm 0.05}$	$14.30_{\pm 0.01}$	$4.08_{\pm 0.01}$	$14.29_{\pm 0.01}$	$0.65_{\pm 0.01}$	$14.29_{\pm 0.00}$
IdkDPO	$8.51_{\pm 0.01}$	$15.78_{\pm 0.78}$	$3.95_{\pm 0.01}$	$15.02_{\pm 0.28}$	$0.49_{\pm 0.00}$	$15.07_{\pm 0.08}$	$24.17_{\pm 0.29}$	$20.93_{\pm 0.30}$	$16.76_{\pm 4.07}$	$20.72_{\pm 0.04}$	$1.69_{\pm 0.01}$	$20.75_{\pm 0.00}$
GradAscent	$0.27_{\pm 0.00}$	$6.42_{\pm 0.19}$	$0.14_{\pm 0.00}$	$6.25_{\pm 0.00}$	$0.04_{\pm 0.00}$	$5.62_{\pm 0.01}$	$1.00_{\pm 0.01}$	$14.19_{\pm 0.00}$	$0.42_{\pm 0.01}$	$14.19_{\pm 0.00}$	$0.09_{\pm 0.00}$	$13.46_{\pm 0.00}$
CEU	$0.28_{\pm 0.00}$	$7.19_{\pm 0.16}$	$0.14_{\pm 0.00}$	$7.03_{\pm 0.02}$	$0.04_{\pm 0.00}$	$6.14_{\pm 0.15}$	$1.01_{\pm 0.01}$	$14.19_{\pm 0.00}$	$0.42_{\pm 0.00}$	$14.19_{\pm 0.00}$	$0.09_{\pm 0.00}$	$13.46_{\pm 0.00}$

Table 15: Aggregate, Memorization, Privacy, and Utility scores across forget splits 10%, 5%, and 1% for all unlearning methods (Llama-3.2-1B-Instruct).

Method	Forget 10%				Forget 5%				Forget 1%			
	Agg	Mem	Priv	Util	**Agg**	Mem	Priv	Util	**Agg**	Mem	Priv	Util
Original (init)	$0.1752_{\pm 0.0000}$	$0.0778_{\pm 0.0000}$	$0.3828_{\pm 0.0000}$	$0.6072_{\pm 0.0000}$	$0.1700_{\pm 0.0000}$	$0.0754_{\pm 0.0000}$	$0.3628_{\pm 0.0000}$	$0.6142_{\pm 0.0000}$	$0.1444_{\pm 0.0000}$	$0.0583_{\pm 0.0000}$	$0.5001_{\pm 0.0000}$	$0.6154_{\pm 0.0000}$
Retrain (target)	$0.7171_{\pm 0.0000}$	$0.6471_{\pm 0.0000}$	$0.9997_{\pm 0.0000}$	$0.6105_{\pm 0.0000}$	$0.7193_{\pm 0.0000}$	$0.6430_{\pm 0.0000}$	$0.9994_{\pm 0.0000}$	$0.6192_{\pm 0.0000}$	$0.7026_{\pm 0.0000}$	$0.6065_{\pm 0.0000}$	$0.9994_{\pm 0.0000}$	$0.6172_{\pm 0.0000}$
GD-AdamW	$0.5704_{\pm 0.0111}$	$0.5643_{\pm 0.0175}$	$0.7813_{\pm 0.0458}$	$0.4539_{\pm 0.0062}$	$0.5512_{\pm 0.0017}$	$0.5316_{\pm 0.0084}$	$0.6681_{\pm 0.0083}$	$0.4846_{\pm 0.0056}$	$0.6069_{\pm 0.0022}$	$0.5973_{\pm 0.0083}$	$0.8213_{\pm 0.0149}$	$0.4877_{\pm 0.0015}$
GD-Sophia	$0.5584_{\pm 0.0060}$	$0.5513_{\pm 0.0133}$	$0.7225_{\pm 0.0265}$	$0.4606_{\pm 0.0079}$	$0.5653_{\pm 0.0201}$	$0.6450_{\pm 0.0126}$	$0.7699_{\pm 0.0234}$	$0.4085_{\pm 0.0304}$	$0.6002_{\pm 0.0027}$	$0.5881_{\pm 0.0093}$	$0.8689_{\pm 0.0169}$	$0.4659_{\pm 0.0040}$
SU-GD-Muon	$0.5804_{\pm 0.0052}$	$0.6357_{\pm 0.0039}$	$0.8721_{\pm 0.0126}$	$0.4085_{\pm 0.0082}$	$0.5986_{\pm 0.0061}$	$0.6242_{\pm 0.0054}$	$0.8872_{\pm 0.0245}$	$0.4383_{\pm 0.0086}$	$0.6094_{\pm 0.0038}$	$0.5301_{\pm 0.0067}$	$0.9004_{\pm 0.0066}$	$0.5194_{\pm 0.0028}$
SU-GD-KLShampoo	$0.5854_{\pm 0.0075}$	$0.6252_{\pm 0.0109}$	$0.8777_{\pm 0.0296}$	$0.4196_{\pm 0.0120}$	$0.5961_{\pm 0.0096}$	$0.6265_{\pm 0.0107}$	$0.8677_{\pm 0.0345}$	$0.4382_{\pm 0.0111}$	$0.6253_{\pm 0.0060}$	$0.5597_{\pm 0.0005}$	$0.9229_{\pm 0.0261}$	$0.5190_{\pm 0.0045}$
SU-GD-SOAP	$0.6043_{\pm 0.0046}$	$0.6117_{\pm 0.0021}$	$0.8930_{\pm 0.0204}$	$0.4526_{\pm 0.0043}$	$0.6075_{\pm 0.0073}$	$0.6123_{\pm 0.0079}$	$0.8703_{\pm 0.0270}$	$0.4641_{\pm 0.0074}$	$0.6246_{\pm 0.0064}$	$0.5522_{\pm 0.0046}$	$0.9040_{\pm 0.0266}$	$0.5304_{\pm 0.0027}$
NPO-AdamW	$0.6876_{\pm 0.0069}$	$0.6574_{\pm 0.0208}$	$0.8684_{\pm 0.0076}$	$0.5921_{\pm 0.0040}$	$0.5430_{\pm 0.0009}$	$0.5648_{\pm 0.0062}$	$0.6050_{\pm 0.0020}$	$0.4761_{\pm 0.0071}$	$0.4706_{\pm 0.0017}$	$0.3328_{\pm 0.0023}$	$0.5761_{\pm 0.0035}$	$0.6121_{\pm 0.0005}$
NPO-Sophia	$0.6767_{\pm 0.0039}$	$0.6620_{\pm 0.0111}$	$0.8259_{\pm 0.0030}$	$0.5842_{\pm 0.0018}$	$0.6599_{\pm 0.0058}$	$0.6766_{\pm 0.0010}$	$0.7983_{\pm 0.0215}$	$0.5510_{\pm 0.0015}$	$0.6486_{\pm 0.0042}$	$0.5604_{\pm 0.0091}$	$0.9000_{\pm 0.0049}$	$0.5781_{\pm 0.0029}$
SU-NPO-Muon	$0.6865_{\pm 0.0018}$	$0.6528_{\pm 0.0093}$	$0.9177_{\pm 0.0104}$	$0.5721_{\pm 0.0031}$	$0.6671_{\pm 0.0022}$	$0.6636_{\pm 0.0049}$	$0.8347_{\pm 0.0116}$	$0.5580_{\pm 0.0017}$	$0.6484_{\pm 0.0004}$	$0.6319_{\pm 0.0094}$	$0.7479_{\pm 0.0119}$	$0.5861_{\pm 0.0009}$
SU-NPO-KLShampoo	$0.6829_{\pm 0.0041}$	$0.6603_{\pm 0.0058}$	$0.8910_{\pm 0.0071}$	$0.5695_{\pm 0.0037}$	$0.6682_{\pm 0.0076}$	$0.6842_{\pm 0.0119}$	$0.8049_{\pm 0.0095}$	$0.5603_{\pm 0.0095}$	$0.6550_{\pm 0.0041}$	$0.6502_{\pm 0.0173}$	$0.7922_{\pm 0.0368}$	$0.5633_{\pm 0.0128}$
SU-NPO-SOAP	$0.6926_{\pm 0.0038}$	$0.6623_{\pm 0.0070}$	$0.9144_{\pm 0.0180}$	$0.5790_{\pm 0.0096}$	$0.6711_{\pm 0.0033}$	$0.6867_{\pm 0.0091}$	$0.8113_{\pm 0.0097}$	$0.5614_{\pm 0.0033}$	$0.6445_{\pm 0.0128}$	$0.5459_{\pm 0.0210}$	$0.8779_{\pm 0.0183}$	$0.5945_{\pm 0.0051}$
RMU-AdamW	$0.6888_{\pm 0.0021}$	$0.6834_{\pm 0.0015}$	$0.9016_{\pm 0.0037}$	$0.5609_{\pm 0.0041}$	$0.6465_{\pm 0.0150}$	$0.8068_{\pm 0.0024}$	$0.7481_{\pm 0.0101}$	$0.4848_{\pm 0.0214}$	$0.6691_{\pm 0.0068}$	$0.7134_{\pm 0.0100}$	$0.8086_{\pm 0.0211}$	$0.5422_{\pm 0.0076}$
RMU-Sophia	$0.5615_{\pm 0.0110}$	$0.8361_{\pm 0.0049}$	$0.7125_{\pm 0.0038}$	$0.3646_{\pm 0.0121}$	$0.6801_{\pm 0.0025}$	$0.7594_{\pm 0.0073}$	$0.8119_{\pm 0.0090}$	$0.5369_{\pm 0.0037}$	$0.6378_{\pm 0.0000}$	$0.6550_{\pm 0.0000}$	$0.8721_{\pm 0.0000}$	$0.4925_{\pm 0.0000}$
SU-RMU-Muon	$0.6742_{\pm 0.0031}$	$0.6585_{\pm 0.0018}$	$0.9030_{\pm 0.0188}$	$0.5484_{\pm 0.0026}$	$0.6317_{\pm 0.0156}$	$0.6253_{\pm 0.0090}$	$0.7744_{\pm 0.0533}$	$0.5390_{\pm 0.0035}$	$0.6765_{\pm 0.0021}$	$0.7005_{\pm 0.0081}$	$0.8242_{\pm 0.0193}$	$0.5578_{\pm 0.0026}$
SU-RMU-KLShampoo	$0.6879_{\pm 0.0024}$	$0.6722_{\pm 0.0020}$	$0.9500_{\pm 0.0160}$	$0.5493_{\pm 0.0021}$	$0.6586_{\pm 0.0156}$	$0.6455_{\pm 0.0117}$	$0.8548_{\pm 0.0571}$	$0.5455_{\pm 0.0009}$	$0.6826_{\pm 0.0035}$	$0.7031_{\pm 0.0146}$	$0.8093_{\pm 0.0338}$	$0.5765_{\pm 0.0025}$
SU-RMU-SOAP	$0.6982_{\pm 0.0016}$	$0.7088_{\pm 0.0062}$	$0.8683_{\pm 0.0110}$	$0.5766_{\pm 0.0029}$	$0.6862_{\pm 0.0051}$	$0.8024_{\pm 0.0033}$	$0.7607_{\pm 0.0077}$	$0.5523_{\pm 0.0100}$	$0.6891_{\pm 0.0023}$	$0.6684_{\pm 0.0232}$	$0.8627_{\pm 0.0312}$	$0.5901_{\pm 0.0017}$
UNDIAL	$0.5944_{\pm 0.0040}$	$0.8187_{\pm 0.0015}$	$0.7170_{\pm 0.0004}$	$0.4113_{\pm 0.0057}$	$0.6232_{\pm 0.0033}$	$0.8078_{\pm 0.0013}$	$0.7455_{\pm 0.0017}$	$0.4475_{\pm 0.0048}$	$0.5552_{\pm 0.0032}$	$0.8171_{\pm 0.0030}$	$0.6712_{\pm 0.0297}$	$0.3722_{\pm 0.0053}$
SatImp	$0.4643_{\pm 0.0011}$	$0.4711_{\pm 0.0022}$	$0.4477_{\pm 0.0013}$	$0.4751_{\pm 0.0068}$	$0.4094_{\pm 0.0021}$	$0.3409_{\pm 0.0045}$	$0.3782_{\pm 0.0003}$	$0.5714_{\pm 0.0007}$	$0.3755_{\pm 0.0069}$	$0.2277_{\pm 0.0072}$	$0.5084_{\pm 0.0016}$	$0.6139_{\pm 0.0010}$
WGA	$0.4568_{\pm 0.0006}$	$0.3931_{\pm 0.0013}$	$0.4228_{\pm 0.0021}$	$0.6034_{\pm 0.0010}$	$0.4349_{\pm 0.0039}$	$0.3619_{\pm 0.0057}$	$0.4000_{\pm 0.0036}$	$0.6115_{\pm 0.0016}$	$0.4223_{\pm 0.0081}$	$0.2801_{\pm 0.0101}$	$0.5251_{\pm 0.0020}$	$0.6147_{\pm 0.0005}$
AltPO	$0.4516_{\pm 0.0026}$	$0.3942_{\pm 0.0058}$	$0.4407_{\pm 0.0027}$	$0.5444_{\pm 0.0043}$	$0.4422_{\pm 0.0038}$	$0.3959_{\pm 0.0055}$	$0.4165_{\pm 0.0048}$	$0.5385_{\pm 0.0012}$	$0.3662_{\pm 0.0019}$	$0.2210_{\pm 0.0020}$	$0.5034_{\pm 0.0006}$	$0.5947_{\pm 0.0003}$
SimNPO	$0.4526_{\pm 0.0009}$	$0.3696_{\pm 0.0015}$	$0.4434_{\pm 0.0022}$	$0.5996_{\pm 0.0014}$	$0.4469_{\pm 0.0013}$	$0.3635_{\pm 0.0015}$	$0.4319_{\pm 0.0025}$	$0.6073_{\pm 0.0020}$	$0.4564_{\pm 0.0021}$	$0.3114_{\pm 0.0033}$	$0.5773_{\pm 0.0027}$	$0.6140_{\pm 0.0007}$
PDU	$0.4389_{\pm 0.0068}$	$0.3900_{\pm 0.0150}$	$0.4211_{\pm 0.0045}$	$0.5283_{\pm 0.0051}$	$0.4417_{\pm 0.0193}$	$0.7335_{\pm 0.0054}$	$0.6475_{\pm 0.0025}$	$0.2583_{\pm 0.0210}$	$0.5994_{\pm 0.0025}$	$0.5284_{\pm 0.0047}$	$0.6949_{\pm 0.0045}$	$0.5975_{\pm 0.0018}$
IdkDPO	$0.4387_{\pm 0.0071}$	$0.3636_{\pm 0.0112}$	$0.4236_{\pm 0.0032}$	$0.5794_{\pm 0.0034}$	$0.4003_{\pm 0.0036}$	$0.3157_{\pm 0.0063}$	$0.3941_{\pm 0.0050}$	$0.5592_{\pm 0.0090}$	$0.3212_{\pm 0.0014}$	$0.1759_{\pm 0.0014}$	$0.5037_{\pm 0.0008}$	$0.5987_{\pm 0.0007}$
GradAscent	$0.3432_{\pm 0.0031}$	$0.2258_{\pm 0.0041}$	$0.3899_{\pm 0.0004}$	$0.5727_{\pm 0.0015}$	$0.2869_{\pm 0.0024}$	$0.1650_{\pm 0.0024}$	$0.3653_{\pm 0.0002}$	$0.6038_{\pm 0.0004}$	$0.2957_{\pm 0.0023}$	$0.1530_{\pm 0.0020}$	$0.5061_{\pm 0.0007}$	$0.6135_{\pm 0.0007}$
CEU	$0.2972_{\pm 0.0051}$	$0.6810_{\pm 0.0073}$	$0.5814_{\pm 0.0084}$	$0.1448_{\pm 0.0028}$	$0.3357_{\pm 0.0024}$	$0.2218_{\pm 0.0031}$	$0.3672_{\pm 0.0004}$	$0.5869_{\pm 0.0001}$	$0.2055_{\pm 0.0101}$	$0.0913_{\pm 0.0059}$	$0.5001_{\pm 0.0001}$	$0.6129_{\pm 0.0011}$

Table 16: Aggregate, Memorization, Privacy, and Utility scores across forget splits 10%, 5%, and 1% for all unlearning methods (Llama-3.2-3B-Instruct).

Method	Forget 10%				Forget 5%				Forget 1%			
	Agg	Mem	Priv	Util	**Agg**	Mem	Priv	Util	**Agg**	Mem	Priv	Util
Original (init)	$0.0762_{\pm 0.0000}$	$0.0283_{\pm 0.0000}$	$0.3899_{\pm 0.0000}$	$0.6662_{\pm 0.0000}$	$0.0711_{\pm 0.0000}$	$0.0263_{\pm 0.0000}$	$0.3629_{\pm 0.0000}$	$0.6740_{\pm 0.0000}$	$0.0324_{\pm 0.0000}$	$0.0112_{\pm 0.0000}$	$0.5762_{\pm 0.0000}$	$0.6785_{\pm 0.0000}$
Retrain (target)	$0.7393_{\pm 0.0000}$	$0.6408_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$0.6678_{\pm 0.0000}$	$0.7402_{\pm 0.0000}$	$0.6362_{\pm 0.0000}$	$0.9999_{\pm 0.0000}$	$0.6753_{\pm 0.0000}$	$0.7244_{\pm 0.0000}$	$0.6000_{\pm 0.0000}$	$0.9987_{\pm 0.0000}$	$0.6787_{\pm 0.0000}$
GD-AdamW	$0.3813_{\pm 0.0020}$	$0.2533_{\pm 0.0022}$	$0.4264_{\pm 0.0011}$	$0.6348_{\pm 0.0007}$	$0.5398_{\pm 0.0221}$	$0.6235_{\pm 0.0273}$	$0.8905_{\pm 0.0268}$	$0.3570_{\pm 0.0393}$	$0.5651_{\pm 0.0110}$	$0.6490_{\pm 0.0209}$	$0.6150_{\pm 0.0153}$	$0.4686_{\pm 0.0266}$
GD-Sophia	$0.5839_{\pm 0.0057}$	$0.5688_{\pm 0.0033}$	$0.8112_{\pm 0.0210}$	$0.4658_{\pm 0.0034}$	$0.5737_{\pm 0.0305}$	$0.5906_{\pm 0.0348}$	$0.8964_{\pm 0.0298}$	$0.4228_{\pm 0.0651}$	$0.5617_{\pm 0.0017}$	$0.6631_{\pm 0.0033}$	$0.6404_{\pm 0.0087}$	$0.4403_{\pm 0.0037}$
SU-GD-Muon	$0.5016_{\pm 0.0060}$	$0.4177_{\pm 0.0061}$	$0.5339_{\pm 0.0127}$	$0.5839_{\pm 0.0030}$	$0.5554_{\pm 0.0280}$	$0.6058_{\pm 0.0321}$	$0.8748_{\pm 0.0068}$	$0.3902_{\pm 0.0575}$	$0.6392_{\pm 0.0040}$	$0.5552_{\pm 0.0067}$	$0.8082_{\pm 0.0201}$	$0.6050_{\pm 0.0150}$
SU-GD-KLShampoo	$0.5108_{\pm 0.0067}$	$0.4306_{\pm 0.0071}$	$0.5497_{\pm 0.0147}$	$0.5781_{\pm 0.0037}$	$0.5559_{\pm 0.0373}$	$0.6060_{\pm 0.0378}$	$0.8848_{\pm 0.0069}$	$0.3922_{\pm 0.0707}$	$0.6379_{\pm 0.0052}$	$0.5663_{\pm 0.0081}$	$0.8058_{\pm 0.0049}$	$0.5900_{\pm 0.0135}$
SU-GD-SOAP	$0.6091_{\pm 0.0342}$	$0.6065_{\pm 0.0386}$	$0.8707_{\pm 0.0363}$	$0.4799_{\pm 0.0712}$	$0.6455_{\pm 0.0089}$	$0.5715_{\pm 0.0202}$	$0.9439_{\pm 0.0285}$	$0.5455_{\pm 0.0220}$	$0.6217_{\pm 0.0190}$	$0.5952_{\pm 0.0440}$	$0.8097_{\pm 0.0175}$	$0.5357_{\pm 0.0731}$
NPO-AdamW	$0.7243_{\pm 0.0043}$	$0.6629_{\pm 0.0033}$	$0.8930_{\pm 0.0122}$	$0.6608_{\pm 0.0010}$	$0.6080_{\pm 0.0011}$	$0.5756_{\pm 0.0037}$	$0.6505_{\pm 0.0042}$	$0.6025_{\pm 0.0045}$	$0.5521_{\pm 0.0034}$	$0.3952_{\pm 0.0044}$	$0.7057_{\pm 0.0047}$	$0.6730_{\pm 0.0004}$
NPO-Sophia	$0.7234_{\pm 0.0054}$	$0.6374_{\pm 0.0088}$	$0.9534_{\pm 0.0061}$	$0.6538_{\pm 0.0021}$	$0.7062_{\pm 0.0108}$	$0.6671_{\pm 0.0039}$	$0.8311_{\pm 0.0329}$	$0.6473_{\pm 0.0044}$	$0.6667_{\pm 0.0033}$	$0.6481_{\pm 0.0096}$	$0.7457_{\pm 0.0235}$	$0.6195_{\pm 0.0038}$
SU-NPO-Muon	$0.7215_{\pm 0.0062}$	$0.6744_{\pm 0.0053}$	$0.8750_{\pm 0.0228}$	$0.6528_{\pm 0.0011}$	$0.7058_{\pm 0.0011}$	$0.6428_{\pm 0.0054}$	$0.9404_{\pm 0.0129}$	$0.6130_{\pm 0.0027}$	$0.6753_{\pm 0.0026}$	$0.5815_{\pm 0.0046}$	$0.8216_{\pm 0.0159}$	$0.6642_{\pm 0.0003}$
SU-NPO-KLShampoo	$0.7208_{\pm 0.0060}$	$0.6743_{\pm 0.0073}$	$0.8691_{\pm 0.0117}$	$0.6544_{\pm 0.0018}$	$0.7042_{\pm 0.0020}$	$0.6531_{\pm 0.0073}$	$0.9222_{\pm 0.0135}$	$0.6082_{\pm 0.0034}$	$0.6795_{\pm 0.0078}$	$0.6144_{\pm 0.0289}$	$0.8035_{\pm 0.0782}$	$0.6544_{\pm 0.0045}$
SU-NPO-SOAP	$0.7262_{\pm 0.0030}$	$0.6539_{\pm 0.0100}$	$0.9186_{\pm 0.0093}$	$0.6609_{\pm 0.0016}$	$0.7211_{\pm 0.0026}$	$0.6840_{\pm 0.0136}$	$0.8612_{\pm 0.0178}$	$0.6509_{\pm 0.0038}$	$0.6744_{\pm 0.0042}$	$0.5956_{\pm 0.0213}$	$0.8067_{\pm 0.0547}$	$0.6569_{\pm 0.0054}$
RMU-AdamW	$0.7038_{\pm 0.0025}$	$0.7880_{\pm 0.0006}$	$0.7085_{\pm 0.0028}$	$0.6322_{\pm 0.0045}$	$0.7151_{\pm 0.0012}$	$0.8440_{\pm 0.0030}$	$0.6856_{\pm 0.0056}$	$0.6444_{\pm 0.0048}$	$0.7169_{\pm 0.0006}$	$0.6449_{\pm 0.0056}$	$0.8808_{\pm 0.0112}$	$0.6674_{\pm 0.0014}$
RMU-Sophia	$0.6857_{\pm 0.0074}$	$0.8434_{\pm 0.0060}$	$0.6637_{\pm 0.0024}$	$0.5945_{\pm 0.0145}$	$0.7317_{\pm 0.0021}$	$0.6920_{\pm 0.0043}$	$0.9010_{\pm 0.0151}$	$0.6474_{\pm 0.0008}$	$0.7020_{\pm 0.0013}$	$0.6954_{\pm 0.0104}$	$0.7591_{\pm 0.0124}$	$0.6590_{\pm 0.0025}$
SU-RMU-Muon	$0.7091_{\pm 0.0026}$	$0.7743_{\pm 0.0016}$	$0.7276_{\pm 0.0051}$	$0.6390_{\pm 0.0021}$	$0.7205_{\pm 0.0018}$	$0.7516_{\pm 0.0021}$	$0.7744_{\pm 0.0051}$	$0.6486_{\pm 0.0040}$	$0.7133_{\pm 0.0018}$	$0.6667_{\pm 0.0012}$	$0.8390_{\pm 0.0063}$	$0.6605_{\pm 0.0006}$
SU-RMU-KLShampoo	$0.7070_{\pm 0.0029}$	$0.7815_{\pm 0.0014}$	$0.7174_{\pm 0.0061}$	$0.6369_{\pm 0.0026}$	$0.7211_{\pm 0.0029}$	$0.8571_{\pm 0.0060}$	$0.6922_{\pm 0.0079}$	$0.6458_{\pm 0.0037}$	$0.7121_{\pm 0.0020}$	$0.6930_{\pm 0.0081}$	$0.7914_{\pm 0.0123}$	$0.6641_{\pm 0.0002}$
SU-RMU-SOAP	$0.7165_{\pm 0.0008}$	$0.7578_{\pm 0.0041}$	$0.7546_{\pm 0.0027}$	$0.6484_{\pm 0.0002}$	$0.7360_{\pm 0.0011}$	$0.7160_{\pm 0.0033}$	$0.8468_{\pm 0.0058}$	$0.6673_{\pm 0.0016}$	$0.7139_{\pm 0.0033}$	$0.6538_{\pm 0.0301}$	$0.8653_{\pm 0.0693}$	$0.6637_{\pm 0.0009}$
UNDIAL	$0.5806_{\pm 0.0016}$	$0.7749_{\pm 0.0012}$	$0.7358_{\pm 0.0009}$	$0.3973_{\pm 0.0023}$	$0.5950_{\pm 0.0073}$	$0.7670_{\pm 0.0013}$	$0.7485_{\pm 0.0014}$	$0.4165_{\pm 0.0111}$	$0.5925_{\pm 0.0042}$	$0.7451_{\pm 0.0042}$	$0.6938_{\pm 0.0078}$	$0.4386_{\pm 0.0041}$
SatImp	$0.4810_{\pm 0.0129}$	$0.3834_{\pm 0.0108}$	$0.5139_{\pm 0.0214}$	$0.5948_{\pm 0.0082}$	$0.4667_{\pm 0.0019}$	$0.4071_{\pm 0.0024}$	$0.4261_{\pm 0.0018}$	$0.6159_{\pm 0.0021}$	$0.4150_{\pm 0.0028}$	$0.2460_{\pm 0.0028}$	$0.5944_{\pm 0.0006}$	$0.6750_{\pm 0.0009}$
WGA	$0.6172_{\pm 0.0403}$	$0.5464_{\pm 0.0420}$	$0.6912_{\pm 0.0889}$	$0.6384_{\pm 0.0102}$	$0.5160_{\pm 0.0068}$	$0.4446_{\pm 0.0083}$	$0.4869_{\pm 0.0087}$	$0.6622_{\pm 0.0031}$	$0.5015_{\pm 0.0098}$	$0.3416_{\pm 0.0125}$	$0.6384_{\pm 0.0054}$	$0.6729_{\pm 0.0028}$
AltPO	$0.4485_{\pm 0.0011}$	$0.3503_{\pm 0.0025}$	$0.4536_{\pm 0.0016}$	$0.6136_{\pm 0.0020}$	$0.4179_{\pm 0.0032}$	$0.3234_{\pm 0.0028}$	$0.4049_{\pm 0.0049}$	$0.6187_{\pm 0.0012}$	$0.3160_{\pm 0.0094}$	$0.1598_{\pm 0.0072}$	$0.5807_{\pm 0.0012}$	$0.6611_{\pm 0.0006}$
SimNPO	$0.4958_{\pm 0.0023}$	$0.4001_{\pm 0.0033}$	$0.4902_{\pm 0.0053}$	$0.6617_{\pm 0.0009}$	$0.4653_{\pm 0.0037}$	$0.3657_{\pm 0.0062}$	$0.4521_{\pm 0.0029}$	$0.6664_{\pm 0.0022}$	$0.5154_{\pm 0.0022}$	$0.3474_{\pm 0.0022}$	$0.6879_{\pm 0.0036}$	$0.6718_{\pm 0.0009}$
PDU	$0.4147_{\pm 0.0913}$	$0.2901_{\pm 0.0984}$	$0.4888_{\pm 0.1129}$	$0.6236_{\pm 0.0115}$	$0.5428_{\pm 0.0093}$	$0.6403_{\pm 0.0177}$	$0.5714_{\pm 0.0328}$	$0.4532_{\pm 0.0172}$	$0.2319_{\pm 0.0206}$	$0.8814_{\pm 0.0058}$	$0.6156_{\pm 0.0151}$	$0.0984_{\pm 0.0107}$
IdkDPO	$0.4805_{\pm 0.0086}$	$0.3798_{\pm 0.0105}$	$0.4862_{\pm 0.0051}$	$0.6440_{\pm 0.0079}$	$0.4391_{\pm 0.0071}$	$0.3473_{\pm 0.0090}$	$0.4184_{\pm 0.0081}$	$0.6406_{\pm 0.0034}$	$0.3758_{\pm 0.0144}$	$0.2100_{\pm 0.0132}$	$0.5842_{\pm 0.0014}$	$0.6655_{\pm 0.0014}$
GradAscent	$0.3961_{\pm 0.0020}$	$0.2863_{\pm 0.0020}$	$0.4101_{\pm 0.0027}$	$0.6092_{\pm 0.0012}$	$0.2411_{\pm 0.0026}$	$0.1219_{\pm 0.0019}$	$0.3650_{\pm 0.0002}$	$0.6654_{\pm 0.0005}$	$0.2626_{\pm 0.0077}$	$0.1215_{\pm 0.0049}$	$0.5822_{\pm 0.0011}$	$0.6787_{\pm 0.0005}$
CEU	$0.0000_{\pm 0.0000}$	$0.9402_{\pm 0.0008}$	$0.8563_{\pm 0.0042}$	$0.0000_{\pm 0.0000}$	$0.3946_{\pm 0.0023}$	$0.2890_{\pm 0.0034}$	$0.3916_{\pm 0.0012}$	$0.6297_{\pm 0.0009}$	$0.1992_{\pm 0.0071}$	$0.0845_{\pm 0.0039}$	$0.5772_{\pm 0.0000}$	$0.6744_{\pm 0.0007}$

Table 17: Memorization and Privacy detail — Llama-3.2-1B-Instruct on Forget 10%.

Method	Memorization							Privacy				
	Mem	ES	EM	F.Prob	F.TR	TR	P.Prob	**Priv**	MIA.Loss	MIA.ZLib	MIA.MinK	MIA.MinK++
Original (init)	$0.0778_{\pm 0.0000}$	$0.7071_{\pm 0.0000}$	$0.9734_{\pm 0.0000}$	$0.8806_{\pm 0.0000}$	$0.4929_{\pm 0.0000}$	$0.6849_{\pm 0.0000}$	$0.1005_{\pm 0.0000}$	$0.3828_{\pm 0.0000}$	$0.9964_{\pm 0.0000}$	$0.9978_{\pm 0.0000}$	$0.9966_{\pm 0.0000}$	$0.9980_{\pm 0.0000}$
Retrain (target)	$0.6471_{\pm 0.0000}$	$0.0596_{\pm 0.0000}$	$0.5856_{\pm 0.0000}$	$0.1161_{\pm 0.0000}$	$0.6357_{\pm 0.0000}$	$0.5676_{\pm 0.0000}$	$0.0535_{\pm 0.0000}$	$0.9997_{\pm 0.0000}$	$0.3872_{\pm 0.0000}$	$0.3088_{\pm 0.0000}$	$0.3825_{\pm 0.0000}$	$0.4766_{\pm 0.0000}$
GD-AdamW	$0.5643_{\pm 0.0175}$	$0.0793_{\pm 0.0062}$	$0.6492_{\pm 0.0225}$	$0.0552_{\pm 0.0114}$	$0.4791_{\pm 0.0046}$	$0.6642_{\pm 0.0044}$	$0.0072_{\pm 0.0006}$	$0.7813_{\pm 0.0458}$	$0.6623_{\pm 0.0456}$	$0.5868_{\pm 0.0528}$	$0.6327_{\pm 0.0519}$	$0.4970_{\pm 0.0179}$
GD-Sophia	$0.5513_{\pm 0.0133}$	$0.0901_{\pm 0.0066}$	$0.6622_{\pm 0.0161}$	$0.0714_{\pm 0.0107}$	$0.4725_{\pm 0.0034}$	$0.6718_{\pm 0.0020}$	$0.0085_{\pm 0.0005}$	$0.7225_{\pm 0.0265}$	$0.7194_{\pm 0.0259}$	$0.6546_{\pm 0.0329}$	$0.6959_{\pm 0.0311}$	$0.5278_{\pm 0.0041}$
SU-GD-Muon	$0.6357_{\pm 0.0039}$	$0.0605_{\pm 0.0016}$	$0.5435_{\pm 0.0058}$	$0.0071_{\pm 0.0010}$	$0.4929_{\pm 0.0119}$	$0.6042_{\pm 0.0079}$	$0.0032_{\pm 0.0002}$	$0.8721_{\pm 0.0126}$	$0.2840_{\pm 0.0115}$	$0.2151_{\pm 0.0106}$	$0.2359_{\pm 0.0115}$	$0.3141_{\pm 0.0216}$
SU-GD-KLShampoo	$0.6252_{\pm 0.0109}$	$0.0629_{\pm 0.0028}$	$0.5577_{\pm 0.0168}$	$0.0080_{\pm 0.0008}$	$0.4855_{\pm 0.0084}$	$0.6049_{\pm 0.0108}$	$0.0034_{\pm 0.0002}$	$0.8777_{\pm 0.0296}$	$0.2886_{\pm 0.0309}$	$0.2230_{\pm 0.0246}$	$0.2447_{\pm 0.0330}$	$0.3153_{\pm 0.0297}$
SU-GD-SOAP	$0.6117_{\pm 0.0021}$	$0.0655_{\pm 0.0040}$	$0.5890_{\pm 0.0067}$	$0.0136_{\pm 0.0028}$	$0.4947_{\pm 0.0068}$	$0.6197_{\pm 0.0098}$	$0.0050_{\pm 0.0002}$	$0.8930_{\pm 0.0204}$	$0.3244_{\pm 0.0239}$	$0.2461_{\pm 0.0252}$	$0.2744_{\pm 0.0253}$	$0.2949_{\pm 0.0103}$
NPO-AdamW	$0.6574_{\pm 0.0208}$	$0.0723_{\pm 0.0062}$	$0.5749_{\pm 0.0297}$	$0.0588_{\pm 0.0048}$	$0.6303_{\pm 0.0090}$	$0.5776_{\pm 0.0071}$	$0.0186_{\pm 0.0007}$	$0.8684_{\pm 0.0076}$	$0.2388_{\pm 0.0261}$	$0.2452_{\pm 0.0184}$	$0.2292_{\pm 0.0193}$	$0.3257_{\pm 0.0335}$
NPO-Sophia	$0.6620_{\pm 0.0111}$	$0.0703_{\pm 0.0037}$	$0.5688_{\pm 0.0172}$	$0.0569_{\pm 0.0040}$	$0.6307_{\pm 0.0016}$	$0.5749_{\pm 0.0045}$	$0.0186_{\pm 0.0011}$	$0.8259_{\pm 0.0030}$	$0.1982_{\pm 0.0173}$	$0.2176_{\pm 0.0110}$	$0.1904_{\pm 0.0125}$	$0.2657_{\pm 0.0291}$
SU-NPO-Muon	$0.6528_{\pm 0.0093}$	$0.0735_{\pm 0.0049}$	$0.5829_{\pm 0.0103}$	$0.0845_{\pm 0.0008}$	$0.6421_{\pm 0.0096}$	$0.5801_{\pm 0.0071}$	$0.0307_{\pm 0.0016}$	$0.9177_{\pm 0.0104}$	$0.3067_{\pm 0.0152}$	$0.2984_{\pm 0.0097}$	$0.3009_{\pm 0.0127}$	$0.3327_{\pm 0.0137}$
SU-NPO-KLShampoo	$0.6603_{\pm 0.0058}$	$0.0689_{\pm 0.0014}$	$0.5749_{\pm 0.0063}$	$0.0793_{\pm 0.0004}$	$0.6473_{\pm 0.0081}$	$0.5778_{\pm 0.0041}$	$0.0298_{\pm 0.0005}$	$0.8910_{\pm 0.0071}$	$0.2700_{\pm 0.0075}$	$0.2724_{\pm 0.0064}$	$0.2657_{\pm 0.0079}$	$0.3207_{\pm 0.0069}$
SU-NPO-SOAP	$0.6623_{\pm 0.0070}$	$0.0646_{\pm 0.0047}$	$0.5754_{\pm 0.0091}$	$0.0750_{\pm 0.0056}$	$0.6520_{\pm 0.0016}$	$0.5772_{\pm 0.0074}$	$0.0318_{\pm 0.0031}$	$0.9144_{\pm 0.0180}$	$0.2909_{\pm 0.0285}$	$0.2841_{\pm 0.0199}$	$0.2856_{\pm 0.0227}$	$0.3594_{\pm 0.0129}$
RMU-AdamW	$0.6834_{\pm 0.0015}$	$0.0527_{\pm 0.0011}$	$0.4988_{\pm 0.0051}$	$0.1035_{\pm 0.0043}$	$0.5930_{\pm 0.0096}$	$0.6257_{\pm 0.0026}$	$0.0289_{\pm 0.0018}$	$0.9016_{\pm 0.0037}$	$0.2911_{\pm 0.0113}$	$0.2108_{\pm 0.0119}$	$0.3183_{\pm 0.0158}$	$0.6091_{\pm 0.0222}$
RMU-Sophia	$0.8361_{\pm 0.0049}$	$0.0372_{\pm 0.0010}$	$0.1828_{\pm 0.0018}$	$0.0171_{\pm 0.0007}$	$0.6649_{\pm 0.0114}$	$0.5564_{\pm 0.0039}$	$0.0033_{\pm 0.0003}$	$0.7125_{\pm 0.0038}$	$0.0352_{\pm 0.0010}$	$0.0224_{\pm 0.0014}$	$0.0368_{\pm 0.0004}$	$0.3554_{\pm 0.0281}$
SU-RMU-Muon	$0.6585_{\pm 0.0018}$	$0.0557_{\pm 0.0023}$	$0.5413_{\pm 0.0033}$	$0.1321_{\pm 0.0053}$	$0.5942_{\pm 0.0031}$	$0.6289_{\pm 0.0029}$	$0.0339_{\pm 0.0024}$	$0.9030_{\pm 0.0188}$	$0.4560_{\pm 0.0151}$	$0.3717_{\pm 0.0148}$	$0.4817_{\pm 0.0165}$	$0.6290_{\pm 0.0282}$
SU-RMU-KLShampoo	$0.6722_{\pm 0.0020}$	$0.0529_{\pm 0.0023}$	$0.5234_{\pm 0.0028}$	$0.1177_{\pm 0.0039}$	$0.6012_{\pm 0.0030}$	$0.6237_{\pm 0.0029}$	$0.0312_{\pm 0.0020}$	$0.9500_{\pm 0.0160}$	$0.3942_{\pm 0.0135}$	$0.3131_{\pm 0.0128}$	$0.4206_{\pm 0.0160}$	$0.6068_{\pm 0.0280}$
SU-RMU-SOAP	$0.7088_{\pm 0.0062}$	$0.0525_{\pm 0.0006}$	$0.4774_{\pm 0.0081}$	$0.0935_{\pm 0.0059}$	$0.6367_{\pm 0.0051}$	$0.6043_{\pm 0.0013}$	$0.0278_{\pm 0.0016}$	$0.8683_{\pm 0.0110}$	$0.2392_{\pm 0.0194}$	$0.1799_{\pm 0.0182}$	$0.2632_{\pm 0.0240}$	$0.6061_{\pm 0.0191}$
UNDIAL	$0.8187_{\pm 0.0015}$	$0.0333_{\pm 0.0001}$	$0.2054_{\pm 0.0003}$	$0.0763_{\pm 0.0014}$	$0.6623_{\pm 0.0030}$	$0.6086_{\pm 0.0010}$	$0.0943_{\pm 0.0005}$	$0.7170_{\pm 0.0004}$	$0.1283_{\pm 0.0042}$	$0.1310_{\pm 0.0027}$	$0.5433_{\pm 0.0087}$	$0.0266_{\pm 0.0013}$
SatImp	$0.4711_{\pm 0.0022}$	$0.1364_{\pm 0.0011}$	$0.7637_{\pm 0.0014}$	$0.3026_{\pm 0.0069}$	$0.5996_{\pm 0.0009}$	$0.6353_{\pm 0.0002}$	$0.1207_{\pm 0.0037}$	$0.4477_{\pm 0.0013}$	$0.9521_{\pm 0.0020}$	$0.8356_{\pm 0.0080}$	$0.9759_{\pm 0.0010}$	$0.9934_{\pm 0.0005}$
WGA	$0.3931_{\pm 0.0013}$	$0.1692_{\pm 0.0031}$	$0.7956_{\pm 0.0011}$	$0.5274_{\pm 0.0012}$	$0.5090_{\pm 0.0025}$	$0.6768_{\pm 0.0008}$	$0.0703_{\pm 0.0004}$	$0.4228_{\pm 0.0021}$	$0.9680_{\pm 0.0016}$	$0.9543_{\pm 0.0035}$	$0.9698_{\pm 0.0018}$	$0.9323_{\pm 0.0005}$
AltPO	$0.3942_{\pm 0.0058}$	$0.1453_{\pm 0.0010}$	$0.8121_{\pm 0.0040}$	$0.5140_{\pm 0.0068}$	$0.6265_{\pm 0.0029}$	$0.6110_{\pm 0.0018}$	$0.1189_{\pm 0.0006}$	$0.4407_{\pm 0.0027}$	$0.9641_{\pm 0.0025}$	$0.8888_{\pm 0.0050}$	$0.9619_{\pm 0.0021}$	$0.9665_{\pm 0.0008}$
SimNPO	$0.3696_{\pm 0.0015}$	$0.1749_{\pm 0.0027}$	$0.8296_{\pm 0.0012}$	$0.4388_{\pm 0.0029}$	$0.5101_{\pm 0.0019}$	$0.6748_{\pm 0.0008}$	$0.0554_{\pm 0.0004}$	$0.4434_{\pm 0.0022}$	$0.9724_{\pm 0.0010}$	$0.9357_{\pm 0.0022}$	$0.9728_{\pm 0.0010}$	$0.8070_{\pm 0.0111}$
PDU	$0.3900_{\pm 0.0150}$	$0.1968_{\pm 0.0134}$	$0.8187_{\pm 0.0105}$	$0.4317_{\pm 0.0149}$	$0.5781_{\pm 0.0016}$	$0.6436_{\pm 0.0008}$	$0.1113_{\pm 0.0010}$	$0.4211_{\pm 0.0045}$	$0.9578_{\pm 0.0042}$	$0.9310_{\pm 0.0089}$	$0.9732_{\pm 0.0026}$	$0.9972_{\pm 0.0002}$
IdkDPO	$0.3636_{\pm 0.0112}$	$0.2322_{\pm 0.0085}$	$0.8266_{\pm 0.0078}$	$0.5214_{\pm 0.0165}$	$0.5439_{\pm 0.0035}$	$0.6581_{\pm 0.0022}$	$0.0882_{\pm 0.0020}$	$0.4236_{\pm 0.0032}$	$0.9737_{\pm 0.0022}$	$0.9450_{\pm 0.0032}$	$0.9741_{\pm 0.0019}$	$0.9308_{\pm 0.0072}$
GradAscent	$0.2258_{\pm 0.0041}$	$0.3512_{\pm 0.0053}$	$0.9091_{\pm 0.0022}$	$0.7022_{\pm 0.0035}$	$0.5522_{\pm 0.0016}$	$0.6560_{\pm 0.0007}$	$0.1527_{\pm 0.0006}$	$0.3899_{\pm 0.0004}$	$0.9892_{\pm 0.0005}$	$0.9888_{\pm 0.0003}$	$0.9913_{\pm 0.0003}$	$0.9940_{\pm 0.0006}$
CEU	$0.6810_{\pm 0.0073}$	$0.0569_{\pm 0.0020}$	$0.5369_{\pm 0.0099}$	$0.1115_{\pm 0.0059}$	$0.6545_{\pm 0.0031}$	$0.6117_{\pm 0.0014}$	$0.0605_{\pm 0.0028}$	$0.5814_{\pm 0.0084}$	$0.7978_{\pm 0.0089}$	$0.5673_{\pm 0.0102}$	$0.8772_{\pm 0.0082}$	$0.9390_{\pm 0.0068}$

Table 18: Memorization and Privacy detail — Llama-3.2-1B-Instruct on Forget 5%.

Method	Memorization							Privacy				
	Mem	ES	EM	F.Prob	F.TR	TR	P.Prob	**Priv**	MIA.Loss	MIA.ZLib	MIA.MinK	MIA.MinK++
Original (init)	$0.0754_{\pm 0.0000}$	$0.7307_{\pm 0.0000}$	$0.9741_{\pm 0.0000}$	$0.8855_{\pm 0.0000}$	$0.4926_{\pm 0.0000}$	$0.6922_{\pm 0.0000}$	$0.0987_{\pm 0.0000}$	$0.3628_{\pm 0.0000}$	$0.9999_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$0.9995_{\pm 0.0000}$
Retrain (target)	$0.6430_{\pm 0.0000}$	$0.0637_{\pm 0.0000}$	$0.5941_{\pm 0.0000}$	$0.1270_{\pm 0.0000}$	$0.6479_{\pm 0.0000}$	$0.5746_{\pm 0.0000}$	$0.0537_{\pm 0.0000}$	$0.9994_{\pm 0.0000}$	$0.3641_{\pm 0.0000}$	$0.2963_{\pm 0.0000}$	$0.3624_{\pm 0.0000}$	$0.4704_{\pm 0.0000}$
GD-AdamW	$0.5316_{\pm 0.0084}$	$0.0948_{\pm 0.0059}$	$0.6928_{\pm 0.0087}$	$0.1005_{\pm 0.0099}$	$0.4876_{\pm 0.0029}$	$0.6770_{\pm 0.0018}$	$0.0108_{\pm 0.0005}$	$0.6681_{\pm 0.0083}$	$0.7435_{\pm 0.0089}$	$0.7071_{\pm 0.0122}$	$0.7266_{\pm 0.0086}$	$0.5604_{\pm 0.0086}$
GD-Sophia	$0.6450_{\pm 0.0126}$	$0.0495_{\pm 0.0069}$	$0.5082_{\pm 0.0319}$	$0.0073_{\pm 0.0017}$	$0.4756_{\pm 0.0083}$	$0.6124_{\pm 0.0084}$	$0.0012_{\pm 0.0004}$	$0.7699_{\pm 0.0234}$	$0.1538_{\pm 0.0354}$	$0.0888_{\pm 0.0171}$	$0.1198_{\pm 0.0358}$	$0.2116_{\pm 0.0132}$
SU-GD-Muon	$0.6242_{\pm 0.0054}$	$0.0610_{\pm 0.0074}$	$0.5717_{\pm 0.0112}$	$0.0146_{\pm 0.0020}$	$0.5021_{\pm 0.0070}$	$0.6206_{\pm 0.0025}$	$0.0031_{\pm 0.0003}$	$0.8872_{\pm 0.0245}$	$0.2883_{\pm 0.0370}$	$0.2199_{\pm 0.0204}$	$0.2473_{\pm 0.0367}$	$0.2932_{\pm 0.0217}$
SU-GD-KLShampoo	$0.6265_{\pm 0.0107}$	$0.0608_{\pm 0.0039}$	$0.5668_{\pm 0.0127}$	$0.0148_{\pm 0.0019}$	$0.5011_{\pm 0.0127}$	$0.6107_{\pm 0.0100}$	$0.0032_{\pm 0.0006}$	$0.8677_{\pm 0.0345}$	$0.2655_{\pm 0.0436}$	$0.2042_{\pm 0.0328}$	$0.2258_{\pm 0.0455}$	$0.2753_{\pm 0.0267}$
SU-GD-SOAP	$0.6123_{\pm 0.0079}$	$0.0678_{\pm 0.0018}$	$0.5919_{\pm 0.0104}$	$0.0200_{\pm 0.0021}$	$0.5030_{\pm 0.0102}$	$0.6166_{\pm 0.0013}$	$0.0047_{\pm 0.0006}$	$0.8703_{\pm 0.0270}$	$0.2726_{\pm 0.0326}$	$0.2080_{\pm 0.0290}$	$0.2371_{\pm 0.0355}$	$0.2657_{\pm 0.0181}$
NPO-AdamW	$0.5648_{\pm 0.0062}$	$0.0989_{\pm 0.0038}$	$0.6698_{\pm 0.0078}$	$0.2496_{\pm 0.0035}$	$0.6209_{\pm 0.0040}$	$0.6257_{\pm 0.0009}$	$0.0887_{\pm 0.0076}$	$0.6050_{\pm 0.0020}$	$0.8104_{\pm 0.0083}$	$0.7280_{\pm 0.0086}$	$0.8133_{\pm 0.0054}$	$0.6468_{\pm 0.0369}$
NPO-Sophia	$0.6766_{\pm 0.0010}$	$0.0679_{\pm 0.0019}$	$0.5586_{\pm 0.0027}$	$0.0680_{\pm 0.0019}$	$0.6663_{\pm 0.0025}$	$0.5592_{\pm 0.0088}$	$0.0316_{\pm 0.0020}$	$0.7983_{\pm 0.0215}$	$0.1557_{\pm 0.0230}$	$0.1769_{\pm 0.0158}$	$0.1525_{\pm 0.0230}$	$0.2117_{\pm 0.0234}$
SU-NPO-Muon	$0.6636_{\pm 0.0049}$	$0.0679_{\pm 0.0029}$	$0.5790_{\pm 0.0027}$	$0.0891_{\pm 0.0073}$	$0.6750_{\pm 0.0089}$	$0.5644_{\pm 0.0061}$	$0.0421_{\pm 0.0041}$	$0.8347_{\pm 0.0116}$	$0.1867_{\pm 0.0193}$	$0.2222_{\pm 0.0117}$	$0.1877_{\pm 0.0171}$	$0.2472_{\pm 0.0144}$
SU-NPO-KLShampoo	$0.6842_{\pm 0.0119}$	$0.0642_{\pm 0.0053}$	$0.5473_{\pm 0.0155}$	$0.0703_{\pm 0.0078}$	$0.6703_{\pm 0.0063}$	$0.5584_{\pm 0.0060}$	$0.0327_{\pm 0.0036}$	$0.8049_{\pm 0.0095}$	$0.1449_{\pm 0.0152}$	$0.1818_{\pm 0.0155}$	$0.1491_{\pm 0.0118}$	$0.2471_{\pm 0.0432}$
SU-NPO-SOAP	$0.6867_{\pm 0.0091}$	$0.0610_{\pm 0.0065}$	$0.5449_{\pm 0.0119}$	$0.0688_{\pm 0.0037}$	$0.6719_{\pm 0.0045}$	$0.5620_{\pm 0.0063}$	$0.0348_{\pm 0.0014}$	$0.8113_{\pm 0.0097}$	$0.1598_{\pm 0.0136}$	$0.1871_{\pm 0.0141}$	$0.1634_{\pm 0.0111}$	$0.2374_{\pm 0.0294}$
RMU-AdamW	$0.8068_{\pm 0.0024}$	$0.0371_{\pm 0.0028}$	$0.2981_{\pm 0.0080}$	$0.0299_{\pm 0.0026}$	$0.6834_{\pm 0.0086}$	$0.5606_{\pm 0.0066}$	$0.0081_{\pm 0.0014}$	$0.7481_{\pm 0.0101}$	$0.0

Table 19: Memorization and Privacy detail — Llama-3.2-1B-Instruct on Forget 1%.

Method	Memorization							Privacy				
	Mem	ES	EM	F.Prob	F.TR	TR	P.Prob	Priv	MIA.Loss	MIA.ZLib	MIA.MinK	MIA.MinK++
Original (init)	$0.0583_{\pm 0.0000}$	$0.7432_{\pm 0.0000}$	$0.9810_{\pm 0.0000}$	$0.9012_{\pm 0.0000}$	$0.4924_{\pm 0.0000}$	$0.6897_{\pm 0.0000}$	$0.0973_{\pm 0.0000}$	$0.5001_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$0.9994_{\pm 0.0000}$
Retrain (target)	$0.6065_{\pm 0.0000}$	$0.0693_{\pm 0.0000}$	$0.6436_{\pm 0.0000}$	$0.1654_{\pm 0.0000}$	$0.6593_{\pm 0.0000}$	$0.5803_{\pm 0.0000}$	$0.0660_{\pm 0.0000}$	$0.9994_{\pm 0.0000}$	$0.4831_{\pm 0.0000}$	$0.5319_{\pm 0.0000}$	$0.4706_{\pm 0.0000}$	$0.5169_{\pm 0.0000}$
GD-AdamW	$0.5973_{\pm 0.0083}$	$0.0602_{\pm 0.0049}$	$0.6250_{\pm 0.0068}$	$0.0485_{\pm 0.0026}$	$0.5228_{\pm 0.0188}$	$0.6255_{\pm 0.0106}$	$0.0098_{\pm 0.0004}$	$0.8213_{\pm 0.0149}$	$0.3497_{\pm 0.0214}$	$0.3709_{\pm 0.0275}$	$0.3198_{\pm 0.0223}$	$0.2623_{\pm 0.0140}$
GD-Sophia	$0.5881_{\pm 0.0093}$	$0.0656_{\pm 0.0155}$	$0.6214_{\pm 0.0129}$	$0.0590_{\pm 0.0033}$	$0.4953_{\pm 0.0288}$	$0.6202_{\pm 0.0094}$	$0.0086_{\pm 0.0005}$	$0.8689_{\pm 0.0169}$	$0.4091_{\pm 0.0300}$	$0.3815_{\pm 0.0284}$	$0.3740_{\pm 0.0251}$	$0.3269_{\pm 0.0053}$
SU-GD-Muon	$0.5301_{\pm 0.0067}$	$0.0888_{\pm 0.0150}$	$0.6945_{\pm 0.0079}$	$0.0976_{\pm 0.0043}$	$0.4845_{\pm 0.0118}$	$0.6508_{\pm 0.0053}$	$0.0174_{\pm 0.0002}$	$0.9004_{\pm 0.0066}$	$0.5976_{\pm 0.0187}$	$0.6132_{\pm 0.0185}$	$0.5794_{\pm 0.0184}$	$0.4258_{\pm 0.0304}$
SU-GD-KLShampoo	$0.5597_{\pm 0.0005}$	$0.0800_{\pm 0.0143}$	$0.6775_{\pm 0.0069}$	$0.0920_{\pm 0.0086}$	$0.5394_{\pm 0.0219}$	$0.6231_{\pm 0.0098}$	$0.0141_{\pm 0.0013}$	$0.9229_{\pm 0.0261}$	$0.5519_{\pm 0.0387}$	$0.5762_{\pm 0.0378}$	$0.5398_{\pm 0.0419}$	$0.3977_{\pm 0.0361}$
SU-GD-SOAP	$0.5522_{\pm 0.0046}$	$0.0751_{\pm 0.0107}$	$0.6800_{\pm 0.0117}$	$0.0986_{\pm 0.0071}$	$0.5203_{\pm 0.0190}$	$0.6310_{\pm 0.0157}$	$0.0155_{\pm 0.0008}$	$0.9040_{\pm 0.0266}$	$0.5674_{\pm 0.0476}$	$0.5817_{\pm 0.0325}$	$0.5633_{\pm 0.0438}$	$0.3681_{\pm 0.0124}$
NPO-AdamW	$0.3328_{\pm 0.0023}$	$0.2082_{\pm 0.0093}$	$0.8508_{\pm 0.0023}$	$0.5301_{\pm 0.0031}$	$0.5204_{\pm 0.0039}$	$0.6716_{\pm 0.0021}$	$0.1053_{\pm 0.0040}$	$0.5761_{\pm 0.0035}$	$0.9428_{\pm 0.0012}$	$0.9554_{\pm 0.0038}$	$0.9450_{\pm 0.0032}$	$0.8327_{\pm 0.0094}$
NPO-Sophia	$0.5604_{\pm 0.0091}$	$0.0931_{\pm 0.0129}$	$0.6850_{\pm 0.0076}$	$0.1773_{\pm 0.0059}$	$0.6082_{\pm 0.0108}$	$0.6227_{\pm 0.0047}$	$0.0825_{\pm 0.0066}$	$0.9000_{\pm 0.0049}$	$0.6202_{\pm 0.0074}$	$0.5943_{\pm 0.0059}$	$0.6019_{\pm 0.0110}$	$0.4538_{\pm 0.0154}$
SU-NPO-Muon	$0.6319_{\pm 0.0094}$	$0.0651_{\pm 0.0037}$	$0.6159_{\pm 0.0069}$	$0.0622_{\pm 0.0009}$	$0.6291_{\pm 0.0169}$	$0.5948_{\pm 0.0057}$	$0.0499_{\pm 0.0023}$	$0.7479_{\pm 0.0119}$	$0.2377_{\pm 0.0255}$	$0.2993_{\pm 0.0083}$	$0.2044_{\pm 0.0182}$	$0.2569_{\pm 0.0053}$
SU-NPO-KLShampoo	$0.6502_{\pm 0.0173}$	$0.0548_{\pm 0.0074}$	$0.6096_{\pm 0.0167}$	$0.0806_{\pm 0.0055}$	$0.6935_{\pm 0.0279}$	$0.5682_{\pm 0.0011}$	$0.0474_{\pm 0.0045}$	$0.7922_{\pm 0.0368}$	$0.2856_{\pm 0.0586}$	$0.3346_{\pm 0.0254}$	$0.2756_{\pm 0.0594}$	$0.2821_{\pm 0.0141}$
SU-NPO-SOAP	$0.5459_{\pm 0.0210}$	$0.0837_{\pm 0.0171}$	$0.7060_{\pm 0.0170}$	$0.1995_{\pm 0.0188}$	$0.6325_{\pm 0.0236}$	$0.6142_{\pm 0.0044}$	$0.0748_{\pm 0.0111}$	$0.8779_{\pm 0.0183}$	$0.6403_{\pm 0.0420}$	$0.5937_{\pm 0.0330}$	$0.6335_{\pm 0.0527}$	$0.4621_{\pm 0.0928}$
RMU-AdamW	$0.7134_{\pm 0.0100}$	$0.0346_{\pm 0.0022}$	$0.5137_{\pm 0.0121}$	$0.0869_{\pm 0.0064}$	$0.7048_{\pm 0.0093}$	$0.5821_{\pm 0.0077}$	$0.0302_{\pm 0.0016}$	$0.8086_{\pm 0.0211}$	$0.3102_{\pm 0.0268}$	$0.2401_{\pm 0.0492}$	$0.3348_{\pm 0.0329}$	$0.6556_{\pm 0.0210}$
RMU-Sophia	$0.6550_{\pm 0.0000}$	$0.0390_{\pm 0.0000}$	$0.5892_{\pm 0.0000}$	$0.1780_{\pm 0.0000}$	$0.7063_{\pm 0.0000}$	$0.5750_{\pm 0.0000}$	$0.0581_{\pm 0.0000}$	$0.8721_{\pm 0.0000}$	$0.5737_{\pm 0.0000}$	$0.5844_{\pm 0.0000}$	$0.6062_{\pm 0.0000}$	$0.7344_{\pm 0.0000}$
SU-RMU-Muon	$0.7005_{\pm 0.0081}$	$0.0357_{\pm 0.0004}$	$0.5270_{\pm 0.0089}$	$0.1073_{\pm 0.0081}$	$0.6952_{\pm 0.0104}$	$0.5782_{\pm 0.0028}$	$0.0304_{\pm 0.0022}$	$0.8242_{\pm 0.0193}$	$0.3096_{\pm 0.0259}$	$0.3006_{\pm 0.0301}$	$0.3398_{\pm 0.0283}$	$0.6763_{\pm 0.0084}$
SU-RMU-KLShampoo	$0.7031_{\pm 0.0146}$	$0.0396_{\pm 0.0030}$	$0.5221_{\pm 0.0189}$	$0.1063_{\pm 0.0176}$	$0.6968_{\pm 0.0065}$	$0.5822_{\pm 0.0051}$	$0.0292_{\pm 0.0034}$	$0.8093_{\pm 0.0338}$	$0.2892_{\pm 0.0514}$	$0.2620_{\pm 0.0656}$	$0.3073_{\pm 0.0546}$	$0.6242_{\pm 0.0549}$
SU-RMU-SOAP	$0.6684_{\pm 0.0232}$	$0.0514_{\pm 0.0036}$	$0.5506_{\pm 0.0296}$	$0.1323_{\pm 0.0225}$	$0.6462_{\pm 0.0190}$	$0.6073_{\pm 0.0104}$	$0.0361_{\pm 0.0031}$	$0.8627_{\pm 0.0312}$	$0.3857_{\pm 0.0838}$	$0.3690_{\pm 0.0781}$	$0.4045_{\pm 0.0790}$	$0.6796_{\pm 0.0539}$
UNDIAL	$0.8171_{\pm 0.0030}$	$0.0336_{\pm 0.0023}$	$0.3116_{\pm 0.0028}$	$0.0610_{\pm 0.0041}$	$0.7447_{\pm 0.0095}$	$0.5580_{\pm 0.0086}$	$0.0677_{\pm 0.0032}$	$0.6712_{\pm 0.0297}$	$0.1385_{\pm 0.0240}$	$0.2117_{\pm 0.0291}$	$0.1833_{\pm 0.0437}$	$0.1610_{\pm 0.0281}$
SatImp	$0.2277_{\pm 0.0072}$	$0.2746_{\pm 0.0057}$	$0.9021_{\pm 0.0050}$	$0.7474_{\pm 0.0021}$	$0.5002_{\pm 0.0009}$	$0.6835_{\pm 0.0002}$	$0.1159_{\pm 0.0019}$	$0.5084_{\pm 0.0016}$	$0.9905_{\pm 0.0019}$	$0.9996_{\pm 0.0003}$	$0.9927_{\pm 0.0013}$	$0.9835_{\pm 0.0049}$
WGA	$0.2801_{\pm 0.0101}$	$0.2388_{\pm 0.0046}$	$0.8725_{\pm 0.0071}$	$0.6789_{\pm 0.0107}$	$0.4995_{\pm 0.0032}$	$0.6882_{\pm 0.0015}$	$0.0840_{\pm 0.0008}$	$0.5251_{\pm 0.0020}$	$0.9738_{\pm 0.0020}$	$0.9940_{\pm 0.0008}$	$0.9785_{\pm 0.0008}$	$0.9523_{\pm 0.0061}$
AltPO	$0.2210_{\pm 0.0020}$	$0.2573_{\pm 0.0238}$	$0.9124_{\pm 0.0012}$	$0.7205_{\pm 0.0026}$	$0.5678_{\pm 0.0014}$	$0.6504_{\pm 0.0009}$	$0.1587_{\pm 0.0003}$	$0.5034_{\pm 0.0006}$	$0.9928_{\pm 0.0012}$	$0.9996_{\pm 0.0006}$	$0.9969_{\pm 0.0005}$	$0.9977_{\pm 0.0003}$
SimNPO	$0.3114_{\pm 0.0033}$	$0.2130_{\pm 0.0025}$	$0.8666_{\pm 0.0020}$	$0.5288_{\pm 0.0058}$	$0.5105_{\pm 0.0028}$	$0.6819_{\pm 0.0013}$	$0.0798_{\pm 0.0004}$	$0.5773_{\pm 0.0027}$	$0.9476_{\pm 0.0005}$	$0.9575_{\pm 0.0054}$	$0.9498_{\pm 0.0018}$	$0.8063_{\pm 0.0058}$
PDU	$0.5284_{\pm 0.0047}$	$0.1158_{\pm 0.0057}$	$0.6963_{\pm 0.0053}$	$0.2470_{\pm 0.0129}$	$0.5505_{\pm 0.0036}$	$0.6627_{\pm 0.0011}$	$0.0528_{\pm 0.0022}$	$0.6949_{\pm 0.0045}$	$0.7203_{\pm 0.0114}$	$0.6864_{\pm 0.0079}$	$0.7560_{\pm 0.0038}$	$0.9819_{\pm 0.0010}$
IdkDPO	$0.1759_{\pm 0.0014}$	$0.3958_{\pm 0.0184}$	$0.9335_{\pm 0.0011}$	$0.7598_{\pm 0.0027}$	$0.5326_{\pm 0.0023}$	$0.6694_{\pm 0.0013}$	$0.1486_{\pm 0.0006}$	$0.5037_{\pm 0.0008}$	$0.9931_{\pm 0.0009}$	$1.0000_{\pm 0.0000}$	$0.9967_{\pm 0.0011}$	$0.9958_{\pm 0.0011}$
GradAscent	$0.1530_{\pm 0.0020}$	$0.4126_{\pm 0.0128}$	$0.9426_{\pm 0.0009}$	$0.8004_{\pm 0.0020}$	$0.4977_{\pm 0.0023}$	$0.6858_{\pm 0.0009}$	$0.1100_{\pm 0.0016}$	$0.5061_{\pm 0.0007}$	$0.9929_{\pm 0.0012}$	$0.9994_{\pm 0.0005}$	$0.9946_{\pm 0.0008}$	$0.9892_{\pm 0.0016}$
CEU	$0.0913_{\pm 0.0059}$	$0.5920_{\pm 0.0105}$	$0.9701_{\pm 0.0024}$	$0.8291_{\pm 0.0020}$	$0.5057_{\pm 0.0006}$	$0.6824_{\pm 0.0002}$	$0.1282_{\pm 0.0008}$	$0.5001_{\pm 0.0001}$	$0.9998_{\pm 0.0003}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$

Table 20: Memorization and Privacy detail — Llama-3.2-3B-Instruct on Forget 10%.

Method	Memorization							Privacy				
	Mem	ES	EM	F.Prob	F.TR	TR	P.Prob	Priv	MIA.Loss	MIA.ZLib	MIA.MinK	MIA.MinK++
Original (init)	$0.0283_{\pm 0.0000}$	$0.8904_{\pm 0.0000}$	$0.9909_{\pm 0.0000}$	$0.9510_{\pm 0.0000}$	$0.4869_{\pm 0.0000}$	$0.4719_{\pm 0.0000}$	$0.0944_{\pm 0.0000}$	$0.3899_{\pm 0.0000}$	$0.9983_{\pm 0.0000}$	$0.9992_{\pm 0.0000}$	$0.9983_{\pm 0.0000}$	$0.9984_{\pm 0.0000}$
Retrain (target)	$0.6408_{\pm 0.0000}$	$0.0648_{\pm 0.0000}$	$0.5994_{\pm 0.0000}$	$0.1241_{\pm 0.0000}$	$0.6514_{\pm 0.0000}$	$0.6436_{\pm 0.0000}$	$0.0523_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$0.3962_{\pm 0.0000}$	$0.3119_{\pm 0.0000}$	$0.3940_{\pm 0.0000}$	$0.4933_{\pm 0.0000}$
GD-AdamW	$0.2533_{\pm 0.0022}$	$0.3510_{\pm 0.0045}$	$0.8963_{\pm 0.0011}$	$0.5988_{\pm 0.0038}$	$0.4728_{\pm 0.0007}$	$0.6957_{\pm 0.0004}$	$0.0294_{\pm 0.0010}$	$0.4264_{\pm 0.0011}$	$0.9807_{\pm 0.0002}$	$0.9742_{\pm 0.0014}$	$0.9808_{\pm 0.0002}$	$0.8656_{\pm 0.0044}$
GD-Sophia	$0.5688_{\pm 0.0033}$	$0.0911_{\pm 0.0041}$	$0.6206_{\pm 0.0072}$	$0.0469_{\pm 0.0029}$	$0.4453_{\pm 0.0071}$	$0.6817_{\pm 0.0020}$	$0.0028_{\pm 0.0001}$	$0.8112_{\pm 0.0210}$	$0.6327_{\pm 0.0213}$	$0.5485_{\pm 0.0201}$	$0.5927_{\pm 0.0225}$	$0.5526_{\pm 0.0200}$
SU-GD-Muon	$0.4177_{\pm 0.0061}$	$0.1733_{\pm 0.0008}$	$0.7926_{\pm 0.0043}$	$0.2217_{\pm 0.0139}$	$0.4424_{\pm 0.0035}$	$0.6987_{\pm 0.0009}$	$0.0097_{\pm 0.0002}$	$0.5339_{\pm 0.0127}$	$0.8987_{\pm 0.0089}$	$0.8572_{\pm 0.0130}$	$0.8928_{\pm 0.0097}$	$0.7162_{\pm 0.0259}$
SU-GD-KLShampoo	$0.4306_{\pm 0.0071}$	$0.1624_{\pm 0.0042}$	$0.7816_{\pm 0.0053}$	$0.1969_{\pm 0.0131}$	$0.4404_{\pm 0.0038}$	$0.6978_{\pm 0.0007}$	$0.0086_{\pm 0.0002}$	$0.5497_{\pm 0.0147}$	$0.8839_{\pm 0.0103}$	$0.8416_{\pm 0.0168}$	$0.8777_{\pm 0.0116}$	$0.7007_{\pm 0.0246}$
SU-GD-SOAP	$0.6065_{\pm 0.0386}$	$0.0717_{\pm 0.0117}$	$0.5683_{\pm 0.0731}$	$0.0136_{\pm 0.0042}$	$0.4649_{\pm 0.0042}$	$0.6395_{\pm 0.0138}$	$0.0012_{\pm 0.0004}$	$0.8707_{\pm 0.0363}$	$0.2687_{\pm 0.0514}$	$0.2139_{\pm 0.0408}$	$0.2223_{\pm 0.0548}$	$0.3801_{\pm 0.0186}$
NPO-AdamW	$0.6629_{\pm 0.0033}$	$0.0590_{\pm 0.0004}$	$0.5625_{\pm 0.0056}$	$0.0487_{\pm 0.0006}$	$0.6117_{\pm 0.0012}$	$0.5712_{\pm 0.0085}$	$0.0126_{\pm 0.0007}$	$0.8930_{\pm 0.0122}$	$0.2604_{\pm 0.0060}$	$0.2606_{\pm 0.0039}$	$0.2726_{\pm 0.0058}$	$0.4151_{\pm 0.0940}$
NPO-Sophia	$0.6374_{\pm 0.0088}$	$0.0601_{\pm 0.0036}$	$0.5991_{\pm 0.0110}$	$0.0854_{\pm 0.0062}$	$0.6163_{\pm 0.0031}$	$0.5950_{\pm 0.0028}$	$0.0235_{\pm 0.0018}$	$0.9534_{\pm 0.0061}$	$0.4107_{\pm 0.0443}$	$0.3646_{\pm 0.0288}$	$0.4137_{\pm 0.0379}$	$0.4489_{\pm 0.0182}$
SU-NPO-Muon	$0.6744_{\pm 0.0053}$	$0.0605_{\pm 0.0026}$	$0.5440_{\pm 0.0073}$	$0.0512_{\pm 0.0026}$	$0.6175_{\pm 0.0042}$	$0.5658_{\pm 0.0075}$	$0.0168_{\pm 0.0008}$	$0.8750_{\pm 0.0228}$	$0.2355_{\pm 0.0169}$	$0.2468_{\pm 0.0146}$	$0.2463_{\pm 0.0237}$	$0.4185_{\pm 0.1013}$
SU-NPO-KLShampoo	$0.6743_{\pm 0.0073}$	$0.0588_{\pm 0.0014}$	$0.5441_{\pm 0.0063}$	$0.0481_{\pm 0.0012}$	$0.6151_{\pm 0.0124}$	$0.5636_{\pm 0.0067}$	$0.0154_{\pm 0.0006}$	$0.8691_{\pm 0.0117}$	$0.2322_{\pm 0.0017}$	$0.2400_{\pm 0.0029}$	$0.2478_{\pm 0.0179}$	$0.4200_{\pm 0.1176}$
SU-NPO-SOAP	$0.6539_{\pm 0.0100}$	$0.0670_{\pm 0.0094}$	$0.5756_{\pm 0.0101}$	$0.0590_{\pm 0.0085}$	$0.6154_{\pm 0.0187}$	$0.5955_{\pm 0.0155}$	$0.0215_{\pm 0.0058}$	$0.9186_{\pm 0.0093}$	$0.2858_{\pm 0.0338}$	$0.2818_{\pm 0.0240}$	$0.3103_{\pm 0.0229}$	$0.4625_{\pm 0.0968}$
RMU-AdamW	$0.7880_{\pm 0.0006}$	$0.0436_{\pm 0.0018}$	$0.3509_{\pm 0.0038}$	$0.0433_{\pm 0.0012}$	$0.6923_{\pm 0.0046}$	$0.5590_{\pm 0.0045}$	$0.0095_{\pm 0.0006}$	$0.7085_{\pm 0.0028}$	$0.0823_{\pm 0.0034}$	$0.0585_{\pm 0.0006}$	$0.1181_{\pm 0.0080}$	$0.8119_{\pm 0.0062}$
RMU-Sophia	$0.8434_{\pm 0.0060}$	$0.0363_{\pm 0.0008}$	$0.2270_{\pm 0.0088}$	$0.0166_{\pm 0.0014}$	$0.7172_{\pm 0.0122}$	$0.5301_{\pm 0.0028}$	$0.0033_{\pm 0.0001}$	$0.6637_{\pm 0.0024}$	$0.0344_{\pm 0.0006}$	$0.0191_{\pm 0.0034}$	$0.0484_{\pm 0.0031}$	$0.8344_{\pm 0.0081}$
SU-RMU-Muon	$0.7743_{\pm 0.0016}$	$0.0456_{\pm 0.0032}$	$0.3775_{\pm 0.0039}$	$0.0513_{\pm 0.0004}$	$0.6861_{\pm 0.0089}$	$0.5558_{\pm 0.0024}$	$0.0123_{\pm 0.0008}$	$0.7276_{\pm 0.0051}$	$0.1014_{\pm 0.0051}$	$0.0753_{\pm 0.0052}$	$0.1395_{\pm 0.0086}$	$0.7929_{\pm 0.0112}$
SU-RMU-KLShampoo	$0.7815_{\pm 0.0014}$	$0.0443_{\pm 0.0034}$	$0.3626_{\pm 0.0038}$	$0.0466_{\pm 0.0008}$	$0.6878_{\pm 0.0074}$	$0.5502_{\pm 0.0038}$	$0.0111_{\pm 0.0007}$	$0.7174_{\pm 0.0061}$	$0.0905_{\pm 0.0068}$	$0.0663_{\pm 0.0052}$	$0.1269_{\pm 0.0112}$	$0.8012_{\pm 0.0083}$
SU-RMU-SOAP	$0.7578_{\pm 0.0041}$	$0.0439_{\pm 0.0014}$	$0.4139_{\pm 0.0051}$	$0.0513_{\pm 0.0020}$	$0.6791_{\pm 0.0055}$	$0.5591_{\pm 0.0016}$	$0.0134_{\pm 0.0004}$	$0.7546_{\pm 0.0027}$	$0.1059_{\pm 0.0042}$	$0.0733_{\pm 0.0015}$	$0.1306_{\pm 0.0045}$	$0.6744_{\pm 0.0009}$
UNDIAL	$0.7749_{\pm 0.0012}$	$0.0346_{\pm 0.0000}$	$0.2923_{\pm 0.0006}$	$0.0927_{\pm 0.0004}$	$0.6207_{\pm 0.0030}$	$0.6273_{\pm 0.0010}$	$0.1041_{\pm 0.0001}$	$0.7358_{\pm 0.0009}$	$0.2097_{\pm 0.0014}$	$0.1788_{\pm 0.0007}$	$0.5255_{\pm 0.0042}$	$0.0191_{\pm 0.0013}$
SatImp	$0.3834_{\pm 0.0108}$	$0.2197_{\pm 0.0101}$	$0.8358_{\pm 0.0075}$	$0.2750_{\pm 0.0061}$	$0.5956_{\pm 0.0025}$	$0.6388_{\pm 0.0009}$	$0.1061_{\pm 0.0049}$	$0.5139_{\pm 0.0214}$	$0.8592_{\pm 0.0396}$	$0.7152_{\pm 0.0350}$	$0.9482_{\pm 0.0172}$	$0.9926_{\pm 0.0014}$
WGA	$0.5464_{\pm 0.0420}$	$0.1285_{\pm 0.0066}$	$0.6813_{\pm 0.0299}$	$0.2819_{\pm 0.0701}$	$0.6114_{\pm 0.0441}$	$0.6321_{\pm 0.0221}$	$0.0620_{\pm 0.0087}$	$0.6912_{\pm 0.0889}$	$0.6525_{\pm 0.1303}$	$0.6180_{\pm 0.1252}$	$0.7574_{\pm 0.0810}$	$0.7812_{\pm 0.0118}$
AltPO	$0.3503_{\pm 0.0025}$	$0.1721_{\pm 0.0016}$	$0.8417_{\pm 0.0018}$	$0.5667_{\pm 0.0015}$	$0.6295_{\pm 0.0009}$	$0.6125_{\pm 0.0003}$	$0.1302_{\pm 0.0002}$	$0.4536_{\pm 0.0016}$	$0.9657_{\pm 0.0010}$	$0.8856_{\pm 0.0020}$	$0.9627_{\pm 0.0017}$	$0.9472_{\pm 0.0032}$
SimNPO	$0.4001_{\pm 0.0033}$	$0.1782_{\pm 0.0034}$	$0.8150_{\pm 0.0024}$	$0.3352_{\pm 0.0030}$	$0.5340_{\pm 0.0030}$	$0.6706_{\pm 0.0008}$	$0.0560_{\pm 0.0017}$	$0.4902_{\pm 0.0053}$	$0.9503_{\pm 0.0026}$	$0.8932_{\pm 0.0059}$	$0.9541_{\pm 0.0025}$	$0.6959_{\pm 0.0337}$
PDU	$0.2901_{\pm 0.0984}$	$0.3391_{\pm 0.0858}$	$0.8798_{\pm 0.0529}$	$0.4498_{\pm 0.2040}$	$0.5413_{\pm 0.0269}$	$0.6640_{\pm 0.0131}$	$0.0708_{\pm 0.0331}$	$0.4888_{\pm 0.1129}$	$0.8628_{\pm 0.1744}$	$0.8132_{\pm 0.2086}$	$0.9159_{\pm 0.1035}$	$0.9994_{\pm 0.0002}$
IdkDPO	$0.3798_{\pm 0.0105}$	$0.2486_{\pm 0.0137}$	$0.8236_{\pm 0.0067}$	$0.4493_{\pm 0.0138}$	$0.5827_{\pm 0.0041}$	$0.6462_{\pm 0.0029}$	$0.0954_{\pm 0.0017}$	$0.4862_{\pm 0.0051}$	$0.9177_{\pm 0.0085}$	$0.8716_{\pm 0.0103}$	$0.9520_{\pm 0.0021}$	$0.8694_{\pm 0.0053}$
GradAscent	$0.2863_{\pm 0.0020}$	$0.2788_{\pm 0.0018}$	$0.8790_{\pm 0.0012}$	$0.6073_{\pm 0.0037}$	$0.5625_{\pm 0.0011}$	$0.6523_{\pm 0.0004}$	$0.1784_{\pm 0.0031}$	$0.4101_{\pm 0.0027}$	$0.9786_{\pm 0.0008}$	$0.9753_{\pm 0.0008}$	$0.9851_{\pm 0.0006}$	$0.9786_{\pm 0.0137}$
CEU	$0.9402_{\pm 0.0008}$	$0.0325_{\pm 0.0000}$	$0.0011_{\pm 0.0002}$	$0.0000_{\pm 0.0000}$	$0.8199_{\pm 0.0024}$	$0.5398_{\pm 0.0017}$	$0.0000_{\pm 0.0000}$	$0.8563_{\pm 0.0042}$	$0.6556_{\pm 0.0019}$	$0.3885_{\pm 0.0003}$	$0.5658_{\pm 0.0026}$	$0.5229_{\pm 0.0206}$

Table 21: Memorization and Privacy detail — Llama-3.2-3B-Instruct on Forget 5%.

Method	Memorization							Privacy				
	Mem	ES	EM	F.Prob	F.TR	TR	P.Prob	Priv	MIA.Loss	MIA.ZLib	MIA.MinK	MIA.MinK++
Original (init)	$0.0263_{\pm 0.0000}$	$0.8881_{\pm 0.0000}$	$0.9917_{\pm 0.0000}$	$0.9515_{\pm 0.0000}$	$0.4778_{\pm 0.0000}$	$0.6954_{\pm 0.0000}$	$0.0920_{\pm 0.0000}$	$0.3629_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$0.9997_{\pm 0.0000}$
Retrain (target)	$0.6362_{\pm 0.0000}$	$0.0618_{\pm 0.0000}$	$0.6000_{\pm 0.0000}$	$0.1304_{\pm 0.0000}$	$0.6363_{\pm 0.0000}$	$0.5843_{\pm 0.0000}$	$0.0545_{\pm 0.0000}$	$0.9999_{\pm 0.0000}$	$0.3633_{\pm 0.0000}$	$0.2911_{\pm 0.0000}$	$0.3604_{\pm 0.0000}$	$0.4849_{\pm 0.0000}$
GD-AdamW	$0.6235_{\pm 0.0273}$	$0.0682_{\pm 0.0098}$	$0.5240_{\pm 0.0449}$	$0.0276_{\pm 0.0066}$	$0.4537_{\pm 0.0124}$	$0.6638_{\pm 0.0150}$	$0.0021_{\pm 0.0006}$	$0.8905_{\pm 0.0268}$	$0.5129_{\pm 0.0320}$	$0.4369_{\pm 0.0334}$	$0.4704_{\pm 0.0321}$	$0.4979_{\pm 0.0172}$
GD-Sophia	$0.5906_{\pm 0.0348}$	$0.0802_{\pm 0.0179}$	$0.5773_{\pm 0.0578}$	$0.0389_{\pm 0.0108}$	$0.4428_{\pm 0.0119}$	$0.6752_{\pm 0.0167}$	$0.0029_{\pm 0.0010}$	$0.8964_{\pm 0.0298}$	$0.5149_{\pm 0.0396}$	$0.4209_{\pm 0.0376}$	$0.4624_{\pm 0.0451}$	$0.4700_{\pm 0.0143}$
SU-GD-Muon	$0.6058_{\pm 0.0321}$	$0.0769_{\pm 0.0090}$	$0.5665_{\pm 0.0530}$	$0.0313_{\pm 0.0108}$	$0.4626_{\pm 0.0095}$	$0.6718_{\pm 0.0122}$	$0.0024_{\pm 0.0004}$	$0.8748_{\pm 0.0068}$	$0.5416_{\pm 0.0031}$	$0.4512_{\pm 0.0186}$	$0.4889_{\pm 0.0058}$	$0.4849_{\pm 0.0196}$
SU-GD-KLShampoo	$0.6060_{\pm 0.0378}$	$0.0760_{\pm 0.0097}$	$0.5655_{\pm 0.0649}$	$0.0315_{\pm 0.0119}$	$0.4635_{\pm 0.0098}$	$0.6710_{\pm 0.0134}$	$0.0024_{\pm 0.0004}$	$0.8848_{\pm 0.0069}$	$0.5286_{\pm 0.0046}$	$0.4388_{\pm 0.0219}$	$0.4737_{\pm 0.0077}$	$0.4810_{\pm 0.0210}$
SU-GD-SOAP	$0.5715_{\pm 0.0202}$	$0.0897_{\pm 0.0108}$	$0.6363_{\pm 0.0207}$	$0.0282_{\pm 0.0064}$	$0.4721_{\pm 0.0179}$	$0.6695_{\pm 0.0126}$	$0.0036_{\pm 0.0007}$	$0.9439_{\pm 0.0285}$	$0.3337_{\pm 0.0370}$	$0.2801_{\pm 0.0252}$	$0.3001_{\pm 0.0358}$	$0.3779_{\pm 0.0311}$
NPO-AdamW	$0.5756_{\pm 0.0037}$	$0.0756_{\pm 0.0029}$	$0.6581_{\pm 0.0058}$	$0.2200_{\pm 0.0054}$	$0.6025_{\pm 0.0031}$	$0.6322_{\pm 0.0016}$	$0.0833_{\pm 0.0043}$	$0.6505_{\pm 0.0042}$	$0.7220_{\pm 0.0069}$	$0.5962_{\pm 0.0042}$	$0.7870_{\pm 0.0047}$	$0.7735_{\pm 0.0063}$
NPO-Sophia	$0.6671_{\pm 0.0039}$	$0.0600_{\pm 0.0030}$	$0.5646_{\pm 0.0049}$	$0.0561_{\pm 0.0061}$	$0.6348_{\pm 0.0159}$	$0.5736_{\pm 0.0042}$	$0.0202_{\pm 0.0036}$	$0.8311_{\pm 0.0329}$	$0.1917_{\pm 0.0254}$	$0.2085_{\pm 0.0204}$	$0.1914_{\pm 0.0260}$	$0.2483_{\pm 0.0564}$
SU-NPO-Muon	$0.6428_{\pm 0.0054}$	$0.0670_{\pm 0.0014}$	$0.5974_{\pm 0.0052}$	$0.1406_{\pm 0.0076}$	$0.6652_{\pm 0.0053}$	$0.5902_{\pm 0.0029}$	$0.0664_{\pm 0.0037}$	$0.9404_{\pm 0.0129}$	$0.3202_{\pm 0.0360}$	$0.3359_{\pm 0.0235}$	$0.3368_{\pm 0.0374}$	$0.3841_{\pm 0.0283}$
SU-NPO-KLShampoo	$0.6531_{\pm 0.0073}$	$0.0654_{\pm 0.0040}$	$0.5876_{\pm 0.0085}$	$0.1271_{\pm 0.0060}$	$0.6738_{\pm 0.0034}$	$0.5862_{\pm 0.0029}$	$0.0613_{\pm 0.0023}$	$0.9222_{\pm 0.0135}$	$0.2661_{\pm 0.0266}$	$0.2918_{\pm 0.0209}$	$0.2915_{\pm 0.0230}$	$0.3640_{\pm 0.0122}$
SU-NPO-SOAP	$0.6840_{\pm 0.0136}$	$0.0583_{\pm 0.0034}$	$0.5401_{\pm 0.0173}$	$0.0699_{\pm 0.0048}$	$0.6516_{\pm 0.0125}$	$0.5853_{\pm 0.0079}$	$0.0364_{\pm 0.0020}$	$0.8612_{\pm 0.0178}$	$0.1675_{\pm 0.0345}$	$0.2084_{\pm 0.0313}$	$0.2103_{\pm 0.0264}$	$0.3693_{\pm 0.0254}$
RMU-AdamW	$0.8440_{\pm 0.0030}$	$0.0336_{\pm 0.0004}$	$0.2414_{\pm 0.0118}$	$0.0244_{\pm 0.0017}$	$0.7347_{\pm 0.0061}$	$0.5351_{\pm 0.0016}$	$0.0049_{\pm 0.0005}$	$0.6856_{\pm 0.0056}$	$0.0479_{\pm 0.0018}$	$0.0304_{\pm 0.0

Table 22: Memorization and Privacy detail — Llama-3.2-3B-Instruct on Forget 1%.

Method	Memorization							Privacy				
	Mem	ES	EM	F.Prob	F.TR	TR	P.Prob	Priv	MIA.Loss	MIA.ZLib	MIA.MinK	MIA.MinK++
Original (init)	$0.0112_{\pm 0.0000}$	$0.9202_{\pm 0.0000}$	$0.9968_{\pm 0.0000}$	$0.9682_{\pm 0.0000}$	$0.5186_{\pm 0.0000}$	$0.6693_{\pm 0.0000}$	$0.0908_{\pm 0.0000}$	$0.5762_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$
Retrain (target)	$0.6000_{\pm 0.0000}$	$0.0665_{\pm 0.0000}$	$0.6484_{\pm 0.0000}$	$0.1794_{\pm 0.0000}$	$0.6525_{\pm 0.0000}$	$0.5830_{\pm 0.0000}$	$0.0648_{\pm 0.0000}$	$0.9987_{\pm 0.0000}$	$0.5538_{\pm 0.0000}$	$0.5844_{\pm 0.0000}$	$0.5537_{\pm 0.0000}$	$0.6144_{\pm 0.0000}$
GD-AdamW	$0.6490_{\pm 0.0209}$	$0.0658_{\pm 0.0102}$	$0.5160_{\pm 0.0162}$	$0.0461_{\pm 0.0057}$	$0.5060_{\pm 0.0288}$	$0.5874_{\pm 0.0055}$	$0.0031_{\pm 0.0010}$	$0.6150_{\pm 0.0153}$	$0.1829_{\pm 0.0174}$	$0.1607_{\pm 0.0149}$	$0.1613_{\pm 0.0172}$	$0.2702_{\pm 0.0240}$
GD-Sophia	$0.6631_{\pm 0.0033}$	$0.0782_{\pm 0.0105}$	$0.4999_{\pm 0.0105}$	$0.0498_{\pm 0.0018}$	$0.5281_{\pm 0.0096}$	$0.5965_{\pm 0.0117}$	$0.0038_{\pm 0.0023}$	$0.6404_{\pm 0.0087}$	$0.2039_{\pm 0.0069}$	$0.1853_{\pm 0.0055}$	$0.1728_{\pm 0.0080}$	$0.3178_{\pm 0.0199}$
SU-GD-Muon	$0.5552_{\pm 0.0067}$	$0.1327_{\pm 0.0035}$	$0.6893_{\pm 0.0112}$	$0.0858_{\pm 0.0097}$	$0.5778_{\pm 0.0318}$	$0.6282_{\pm 0.0140}$	$0.0107_{\pm 0.0027}$	$0.8082_{\pm 0.0201}$	$0.4377_{\pm 0.0143}$	$0.4552_{\pm 0.0152}$	$0.4115_{\pm 0.0153}$	$0.2817_{\pm 0.0327}$
SU-GD-KLShampoo	$0.5663_{\pm 0.0081}$	$0.1213_{\pm 0.0125}$	$0.6626_{\pm 0.0203}$	$0.0894_{\pm 0.0075}$	$0.5411_{\pm 0.0294}$	$0.5917_{\pm 0.0079}$	$0.0116_{\pm 0.0039}$	$0.8058_{\pm 0.0049}$	$0.4201_{\pm 0.0125}$	$0.4383_{\pm 0.0118}$	$0.4000_{\pm 0.0124}$	$0.3031_{\pm 0.0188}$
SU-GD-SOAP	$0.5952_{\pm 0.0440}$	$0.0995_{\pm 0.0352}$	$0.6160_{\pm 0.0713}$	$0.0872_{\pm 0.0116}$	$0.5382_{\pm 0.0176}$	$0.6093_{\pm 0.0119}$	$0.0092_{\pm 0.0020}$	$0.8097_{\pm 0.0175}$	$0.3992_{\pm 0.0135}$	$0.4254_{\pm 0.0153}$	$0.3810_{\pm 0.0133}$	$0.3548_{\pm 0.0518}$
NPO-AdamW	$0.3952_{\pm 0.0044}$	$0.1982_{\pm 0.0050}$	$0.8151_{\pm 0.0029}$	$0.4097_{\pm 0.0107}$	$0.5645_{\pm 0.0057}$	$0.6519_{\pm 0.0018}$	$0.0888_{\pm 0.0023}$	$0.7057_{\pm 0.0047}$	$0.9169_{\pm 0.0026}$	$0.9062_{\pm 0.0062}$	$0.9183_{\pm 0.0050}$	$0.6673_{\pm 0.0065}$
NPO-Sophia	$0.6481_{\pm 0.0096}$	$0.0723_{\pm 0.0122}$	$0.6064_{\pm 0.0082}$	$0.0954_{\pm 0.0049}$	$0.6909_{\pm 0.0165}$	$0.5888_{\pm 0.0030}$	$0.0553_{\pm 0.0037}$	$0.7457_{\pm 0.0235}$	$0.2996_{\pm 0.0347}$	$0.3577_{\pm 0.0220}$	$0.2731_{\pm 0.0388}$	$0.3654_{\pm 0.0137}$
SU-NPO-Muon	$0.5815_{\pm 0.0046}$	$0.1020_{\pm 0.0037}$	$0.6807_{\pm 0.0057}$	$0.1207_{\pm 0.0037}$	$0.6690_{\pm 0.0143}$	$0.6015_{\pm 0.0045}$	$0.0594_{\pm 0.0014}$	$0.8216_{\pm 0.0159}$	$0.4606_{\pm 0.0227}$	$0.4442_{\pm 0.0134}$	$0.4210_{\pm 0.0212}$	$0.3060_{\pm 0.0114}$
SU-NPO-KLShampoo	$0.6144_{\pm 0.0289}$	$0.0720_{\pm 0.0171}$	$0.6504_{\pm 0.0261}$	$0.1255_{\pm 0.0308}$	$0.7021_{\pm 0.0178}$	$0.5835_{\pm 0.0094}$	$0.0622_{\pm 0.0148}$	$0.8035_{\pm 0.0782}$	$0.3900_{\pm 0.1094}$	$0.4147_{\pm 0.0584}$	$0.3585_{\pm 0.1016}$	$0.3692_{\pm 0.0511}$
SU-NPO-SOAP	$0.5956_{\pm 0.0213}$	$0.0793_{\pm 0.0102}$	$0.6654_{\pm 0.0210}$	$0.1380_{\pm 0.0276}$	$0.6774_{\pm 0.0065}$	$0.5974_{\pm 0.0123}$	$0.0704_{\pm 0.0148}$	$0.8067_{\pm 0.0547}$	$0.3934_{\pm 0.0685}$	$0.4223_{\pm 0.0445}$	$0.3713_{\pm 0.0671}$	$0.3585_{\pm 0.0492}$
RMU-AdamW	$0.6449_{\pm 0.0056}$	$0.0657_{\pm 0.0015}$	$0.5973_{\pm 0.0090}$	$0.1686_{\pm 0.0022}$	$0.6917_{\pm 0.0043}$	$0.5860_{\pm 0.0064}$	$0.0435_{\pm 0.0007}$	$0.8808_{\pm 0.0112}$	$0.3986_{\pm 0.0099}$	$0.4417_{\pm 0.0090}$	$0.4185_{\pm 0.0147}$	$0.6450_{\pm 0.0133}$
RMU-Sophia	$0.6954_{\pm 0.0104}$	$0.0446_{\pm 0.0129}$	$0.5402_{\pm 0.0114}$	$0.1140_{\pm 0.0119}$	$0.7137_{\pm 0.0127}$	$0.5624_{\pm 0.0033}$	$0.0286_{\pm 0.0016}$	$0.7591_{\pm 0.0124}$	$0.2802_{\pm 0.0278}$	$0.2481_{\pm 0.0089}$	$0.3077_{\pm 0.0284}$	$0.6634_{\pm 0.0311}$
SU-RMU-Muon	$0.6667_{\pm 0.0012}$	$0.0522_{\pm 0.0113}$	$0.5670_{\pm 0.0035}$	$0.1474_{\pm 0.0034}$	$0.6843_{\pm 0.0087}$	$0.5720_{\pm 0.0056}$	$0.0370_{\pm 0.0015}$	$0.8390_{\pm 0.0063}$	$0.3538_{\pm 0.0088}$	$0.3917_{\pm 0.0091}$	$0.3740_{\pm 0.0072}$	$0.6675_{\pm 0.0022}$
SU-RMU-KLShampoo	$0.6930_{\pm 0.0081}$	$0.0374_{\pm 0.0013}$	$0.5463_{\pm 0.0107}$	$0.1311_{\pm 0.0063}$	$0.7261_{\pm 0.0093}$	$0.5590_{\pm 0.0023}$	$0.0300_{\pm 0.0020}$	$0.7914_{\pm 0.0123}$	$0.2974_{\pm 0.0169}$	$0.3258_{\pm 0.0217}$	$0.3167_{\pm 0.0187}$	$0.6565_{\pm 0.0284}$
SU-RMU-SOAP	$0.6538_{\pm 0.0301}$	$0.0616_{\pm 0.0048}$	$0.5849_{\pm 0.0298}$	$0.1610_{\pm 0.0351}$	$0.6910_{\pm 0.0276}$	$0.5811_{\pm 0.0220}$	$0.0432_{\pm 0.0089}$	$0.8653_{\pm 0.0693}$	$0.3937_{\pm 0.0925}$	$0.4165_{\pm 0.0896}$	$0.4102_{\pm 0.0949}$	$0.6588_{\pm 0.0197}$
UNDIAL	$0.7451_{\pm 0.0042}$	$0.0318_{\pm 0.0003}$	$0.3924_{\pm 0.0018}$	$0.1330_{\pm 0.0017}$	$0.6510_{\pm 0.0103}$	$0.6111_{\pm 0.0026}$	$0.1278_{\pm 0.0007}$	$0.6938_{\pm 0.0078}$	$0.3016_{\pm 0.0117}$	$0.3295_{\pm 0.0079}$	$0.4472_{\pm 0.0073}$	$0.1235_{\pm 0.0106}$
SatImp	$0.2460_{\pm 0.0028}$	$0.3336_{\pm 0.0179}$	$0.8882_{\pm 0.0016}$	$0.7423_{\pm 0.0061}$	$0.5182_{\pm 0.0008}$	$0.6710_{\pm 0.0010}$	$0.1135_{\pm 0.0015}$	$0.5944_{\pm 0.0006}$	$0.9889_{\pm 0.0009}$	$0.9960_{\pm 0.0025}$	$0.9915_{\pm 0.0008}$	$0.9425_{\pm 0.0077}$
WGA	$0.3416_{\pm 0.0125}$	$0.2324_{\pm 0.0092}$	$0.8395_{\pm 0.0078}$	$0.5705_{\pm 0.0239}$	$0.5429_{\pm 0.0003}$	$0.6616_{\pm 0.0008}$	$0.0746_{\pm 0.0036}$	$0.6384_{\pm 0.0054}$	$0.9570_{\pm 0.0030}$	$0.9698_{\pm 0.0042}$	$0.9608_{\pm 0.0042}$	$0.8348_{\pm 0.0143}$
AltPO	$0.1598_{\pm 0.0072}$	$0.3971_{\pm 0.0229}$	$0.9416_{\pm 0.0033}$	$0.7779_{\pm 0.0038}$	$0.5755_{\pm 0.0012}$	$0.6410_{\pm 0.0005}$	$0.1607_{\pm 0.0004}$	$0.5807_{\pm 0.0012}$	$0.9969_{\pm 0.0005}$	$0.9921_{\pm 0.0018}$	$0.9987_{\pm 0.0000}$	$0.9933_{\pm 0.0033}$
SimNPO	$0.3474_{\pm 0.0022}$	$0.2263_{\pm 0.0047}$	$0.8490_{\pm 0.0010}$	$0.4389_{\pm 0.0162}$	$0.5510_{\pm 0.0024}$	$0.6558_{\pm 0.0007}$	$0.0674_{\pm 0.0034}$	$0.6879_{\pm 0.0036}$	$0.9392_{\pm 0.0039}$	$0.9376_{\pm 0.0031}$	$0.9413_{\pm 0.0031}$	$0.6283_{\pm 0.0145}$
PDU	$0.8814_{\pm 0.0058}$	$0.0309_{\pm 0.0009}$	$0.1269_{\pm 0.0124}$	$0.0009_{\pm 0.0002}$	$0.7354_{\pm 0.0072}$	$0.5463_{\pm 0.0022}$	$0.0004_{\pm 0.0001}$	$0.6156_{\pm 0.0151}$	$0.0518_{\pm 0.0024}$	$0.1704_{\pm 0.0110}$	$0.1952_{\pm 0.0098}$	$0.4371_{\pm 0.0618}$
IdkDPO	$0.2100_{\pm 0.0132}$	$0.4327_{\pm 0.0265}$	$0.9136_{\pm 0.0066}$	$0.7430_{\pm 0.0153}$	$0.5547_{\pm 0.0023}$	$0.6538_{\pm 0.0009}$	$0.1511_{\pm 0.0016}$	$0.5842_{\pm 0.0014}$	$0.9935_{\pm 0.0013}$	$0.9981_{\pm 0.0009}$	$0.9956_{\pm 0.0010}$	$0.9792_{\pm 0.0043}$
GradAscent	$0.1215_{\pm 0.0049}$	$0.5920_{\pm 0.0144}$	$0.9533_{\pm 0.0027}$	$0.8590_{\pm 0.0020}$	$0.5151_{\pm 0.0006}$	$0.6717_{\pm 0.0005}$	$0.1091_{\pm 0.0005}$	$0.5822_{\pm 0.0011}$	$0.9975_{\pm 0.0005}$	$1.0000_{\pm 0.0000}$	$0.9990_{\pm 0.0003}$	$0.9762_{\pm 0.0065}$
CEU	$0.0845_{\pm 0.0039}$	$0.6448_{\pm 0.0078}$	$0.9719_{\pm 0.0016}$	$0.8570_{\pm 0.0034}$	$0.5186_{\pm 0.0003}$	$0.6714_{\pm 0.0002}$	$0.1222_{\pm 0.0006}$	$0.5772_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$0.9950_{\pm 0.0000}$

Table 23: Utility detail — Llama-3.2-1B-Instruct on Forget 10%.

Method	Utility										
	Util	Fluency	R.Prob	R.ROUGE	R.TR	RA.Prob	RA.ROUGE	RA.TR	WF.Prob	WF.ROUGE	WF.TR
Original (init)	$0.6072_{\pm 0.0000}$	$0.8594_{\pm 0.0000}$	$0.8706_{\pm 0.0000}$	$0.7970_{\pm 0.0000}$	$0.5175_{\pm 0.0000}$	$0.4137_{\pm 0.0000}$	$0.8170_{\pm 0.0000}$	$0.5274_{\pm 0.0000}$	$0.4356_{\pm 0.0000}$	$0.8279_{\pm 0.0000}$	$0.6198_{\pm 0.0000}$
Retrain (target)	$0.6105_{\pm 0.0000}$	$0.9052_{\pm 0.0000}$	$0.8803_{\pm 0.0000}$	$0.8254_{\pm 0.0000}$	$0.5136_{\pm 0.0000}$	$0.3918_{\pm 0.0000}$	$0.8045_{\pm 0.0000}$	$0.5048_{\pm 0.0000}$	$0.4286_{\pm 0.0000}$	$0.8194_{\pm 0.0000}$	$0.6181_{\pm 0.0000}$
GD-AdamW	$0.4539_{\pm 0.0062}$	$0.7000_{\pm 0.0491}$	$0.3331_{\pm 0.0253}$	$0.3918_{\pm 0.0100}$	$0.5390_{\pm 0.0022}$	$0.3317_{\pm 0.0087}$	$0.4703_{\pm 0.0206}$	$0.4661_{\pm 0.0109}$	$0.3744_{\pm 0.0122}$	$0.7598_{\pm 0.0215}$	$0.5570_{\pm 0.0013}$
GD-Sophia	$0.4606_{\pm 0.0079}$	$0.7400_{\pm 0.0489}$	$0.3292_{\pm 0.0271}$	$0.3900_{\pm 0.0087}$	$0.5362_{\pm 0.0020}$	$0.3339_{\pm 0.0094}$	$0.5163_{\pm 0.0361}$	$0.4724_{\pm 0.0074}$	$0.3780_{\pm 0.0063}$	$0.7736_{\pm 0.0060}$	$0.5578_{\pm 0.0051}$
SU-GD-Muon	$0.4085_{\pm 0.0082}$	$0.5050_{\pm 0.0956}$	$0.2484_{\pm 0.0106}$	$0.3657_{\pm 0.0067}$	$0.5573_{\pm 0.0111}$	$0.3091_{\pm 0.0126}$	$0.4257_{\pm 0.0349}$	$0.4557_{\pm 0.0056}$	$0.3613_{\pm 0.0110}$	$0.7228_{\pm 0.0270}$	$0.5662_{\pm 0.0088}$
SU-GD-KLShampoo	$0.4196_{\pm 0.0120}$	$0.5859_{\pm 0.1146}$	$0.2686_{\pm 0.0121}$	$0.3837_{\pm 0.0141}$	$0.5620_{\pm 0.0072}$	$0.3083_{\pm 0.0038}$	$0.4336_{\pm 0.0310}$	$0.4540_{\pm 0.0036}$	$0.3455_{\pm 0.0130}$	$0.7275_{\pm 0.0096}$	$0.5735_{\pm 0.0219}$
SU-GD-SOAP	$0.4526_{\pm 0.0043}$	$0.7150_{\pm 0.0092}$	$0.3685_{\pm 0.0134}$	$0.4119_{\pm 0.0069}$	$0.5546_{\pm 0.0012}$	$0.3151_{\pm 0.0021}$	$0.4648_{\pm 0.0280}$	$0.4492_{\pm 0.0054}$	$0.3563_{\pm 0.0091}$	$0.7204_{\pm 0.0062}$	$0.5405_{\pm 0.0110}$
NPO-AdamW	$0.5921_{\pm 0.0040}$	$0.9146_{\pm 0.0107}$	$0.8020_{\pm 0.0031}$	$0.6579_{\pm 0.0077}$	$0.5100_{\pm 0.0032}$	$0.4133_{\pm 0.0035}$	$0.7783_{\pm 0.0359}$	$0.5355_{\pm 0.0120}$	$0.4309_{\pm 0.0097}$	$0.8202_{\pm 0.0137}$	$0.6068_{\pm 0.0087}$
NPO-Sophia	$0.5842_{\pm 0.0018}$	$0.9120_{\pm 0.0054}$	$0.7718_{\pm 0.0055}$	$0.6246_{\pm 0.0052}$	$0.5075_{\pm 0.0021}$	$0.3946_{\pm 0.0111}$	$0.7487_{\pm 0.0169}$	$0.5275_{\pm 0.0102}$	$0.4350_{\pm 0.0024}$	$0.8132_{\pm 0.0198}$	$0.5988_{\pm 0.0080}$
SU-NPO-Muon	$0.5721_{\pm 0.0031}$	$0.9102_{\pm 0.0222}$	$0.7075_{\pm 0.0103}$	$0.5349_{\pm 0.0121}$	$0.5024_{\pm 0.0075}$	$0.4146_{\pm 0.0043}$	$0.7653_{\pm 0.0252}$	$0.5314_{\pm 0.0025}$	$0.4348_{\pm 0.0134}$	$0.8393_{\pm 0.0053}$	$0.6042_{\pm 0.0137}$
SU-NPO-KLShampoo	$0.5695_{\pm 0.0037}$	$0.9003_{\pm 0.0168}$	$0.7149_{\pm 0.0041}$	$0.5451_{\pm 0.0133}$	$0.4983_{\pm 0.0011}$	$0.4067_{\pm 0.0085}$	$0.7673_{\pm 0.0328}$	$0.5186_{\pm 0.0076}$	$0.4157_{\pm 0.0045}$	$0.8317_{\pm 0.0111}$	$0.5833_{\pm 0.0054}$
SU-NPO-SOAP	$0.5790_{\pm 0.0096}$	$0.9049_{\pm 0.0079}$	$0.7660_{\pm 0.0238}$	$0.6088_{\pm 0.0367}$	$0.5055_{\pm 0.0030}$	$0.4173_{\pm 0.0260}$	$0.7635_{\pm 0.0205}$	$0.5422_{\pm 0.0271}$	$0.4206_{\pm 0.0258}$	$0.8321_{\pm 0.0101}$	$0.5972_{\pm 0.0358}$
RMU-AdamW	$0.5609_{\pm 0.0041}$	$0.5496_{\pm 0.0196}$	$0.7736_{\pm 0.0061}$	$0.6479_{\pm 0.0129}$	$0.5158_{\pm 0.0010}$	$0.4185_{\pm 0.0035}$	$0.6928_{\pm 0.0239}$	$0.5305_{\pm 0.0049}$	$0.4273_{\pm 0.0023}$	$0.8203_{\pm 0.0048}$	$0.6035_{\pm 0.0030}$
RMU-Sophia	$0.3646_{\pm 0.0121}$	$0.0878_{\pm 0.0068}$	$0.7785_{\pm 0.0036}$	$0.6349_{\pm 0.0022}$	$0.5041_{\pm 0.0028}$	$0.4077_{\pm 0.0027}$	$0.7287_{\pm 0.0334}$	$0.5259_{\pm 0.0033}$	$0.4347_{\pm 0.0032}$	$0.8217_{\pm 0.0147}$	$0.6007_{\pm 0.0025}$
SU-RMU-Muon	$0.5484_{\pm 0.0026}$	$0.6405_{\pm 0.0311}$	$0.6696_{\pm 0.0107}$	$0.5369_{\pm 0.0070}$	$0.5080_{\pm 0.0013}$	$0.4015_{\pm 0.0031}$	$0.6581_{\pm 0.0156}$	$0.5161_{\pm 0.0043}$	$0.4256_{\pm 0.0015}$	$0.8365_{\pm 0.0141}$	$0.5911_{\pm 0.0024}$
SU-RMU-KLShampoo	$0.5493_{\pm 0.0021}$	$0.6144_{\pm 0.0256}$	$0.6884_{\pm 0.0092}$	$0.5515_{\pm 0.0069}$	$0.5089_{\pm 0.0009}$	$0.4036_{\pm 0.0033}$	$0.6526_{\pm 0.0102}$	$0.5188_{\pm 0.0046}$	$0.4273_{\pm 0.0014}$	$0.8270_{\pm 0.0071}$	$0.5931_{\pm 0.0031}$
SU-RMU-SOAP	$0.5766_{\pm 0.0029}$	$0.6421_{\pm 0.0303}$	$0.8084_{\pm 0.0008}$	$0.7030_{\pm 0.0077}$	$0.5119_{\pm 0.0012}$	$0.4213_{\pm 0.0009}$	$0.7078_{\pm 0.0163}$	$0.5425_{\pm 0.0009}$	$0.4385_{\pm 0.0032}$	$0.8312_{\pm 0.0086}$	$0.6091_{\pm 0.0045}$
UNDIAL	$0.4113_{\pm 0.0057}$	$0.8409_{\pm 0.0092}$	$0.2618_{\pm 0.0034}$	$0.3657_{\pm 0.0026}$	$0.3448_{\pm 0.0032}$	$0.3685_{\pm 0.0057}$	$0.3502_{\pm 0.0300}$	$0.4307_{\pm 0.0086}$	$0.4070_{\pm 0.0021}$	$0.6921_{\pm 0.0145}$	$0.5396_{\pm 0.0018}$
SatImp	$0.4751_{\pm 0.0068}$	$0.9070_{\pm 0.0052}$	$0.3518_{\pm 0.0056}$	$0.4448_{\pm 0.0028}$	$0.4316_{\pm 0.0017}$	$0.4499_{\pm 0.0102}$	$0.3849_{\pm 0.0238}$	$0.5542_{\pm 0.0199}$	$0.4740_{\pm 0.0018}$	$0.6938_{\pm 0.0397}$	$0.6305_{\pm 0.0017}$
WGA	$0.6034_{\pm 0.0010}$	$0.8862_{\pm 0.0047}$	$0.8756_{\pm 0.0006}$	$0.8053_{\pm 0.0025}$	$0.5270_{\pm 0.0011}$	$0.4058_{\pm 0.0007}$	$0.7237_{\pm 0.0150}$	$0.5269_{\pm 0.0023}$	$0.4323_{\pm 0.0013}$	$0.8317_{\pm 0.0036}$	$0.6323_{\pm 0.0020}$
AltPO	$0.5444_{\pm 0.0043}$	$0.8673_{\pm 0.0117}$	$0.6610_{\pm 0.0096}$	$0.4680_{\pm 0.0109}$	$0.4237_{\pm 0.0049}$	$0.4279_{\pm 0.0045}$	$0.7062_{\pm 0.0103}$	$0.5324_{\pm 0.0067}$	$0.4464_{\pm 0.0062}$	$0.8009_{\pm 0.0053}$	$0.5992_{\pm 0.0093}$
SimNPO	$0.5996_{\pm 0.0014}$	$0.8852_{\pm 0.0051}$	$0.8724_{\pm 0.0015}$	$0.7692_{\pm 0.0125}$	$0.5267_{\pm 0.0014}$	$0.4019_{\pm 0.0046}$	$0.7262_{\pm 0.0170}$	$0.5271_{\pm 0.0026}$	$0.4211_{\pm 0.0020}$	$0.8331_{\pm 0.0037}$	$0.6186_{\pm 0.0029}$
PDU	$0.5283_{\pm 0.0051}$	$0.8688_{\pm 0.0051}$	$0.5154_{\pm 0.0187}$	$0.4598_{\pm 0.0148}$	$0.4455_{\pm 0.0026}$	$0.4041_{\pm 0.0024}$	$0.6113_{\pm 0.0158}$	$0.5043_{\pm 0.0068}$	$0.4666_{\pm 0.0057}$	$0.7833_{\pm 0.0071}$	$0.6205_{\pm 0.0071}$
IdkDPO	$0.5794_{\pm 0.0034}$	$0.9399_{\pm 0.0126}$	$0.8454_{\pm 0.0009}$	$0.6332_{\pm 0.0073}$	$0.5075_{\pm 0.0022}$	$0.3734_{\pm 0.0042}$	$0.7205_{\pm 0.0443}$	$0.4724_{\pm 0.0051}$	$0.4353_{\pm 0.0010}$	$0.8023_{\pm 0.0070}$	$0.5914_{\pm 0.0031}$
GradAscent	$0.5727_{\pm 0.0015}$	$0.8970_{\pm 0.0053}$	$0.7115_{\pm 0.0039}$	$0.5569_{\pm 0.0096}$	$0.4640_{\pm 0.0018}$	$0.4136_{\pm 0.0005}$	$0.8528_{\pm 0.0047}$	$0.5143_{\pm 0.0005}$	$0.4429_{\pm 0.0012}$	$0.8333_{\pm 0.0081}$	$0.5989_{\pm 0.0022}$
CEU	$0.1448_{\pm 0.0028}$	$0.8961_{\pm 0.0037}$	$0.1079_{\pm 0.0057}$	$0.3165_{\pm 0.0085}$	$0.3481_{\pm 0.0041}$	$0.3355_{\pm 0.0032}$	$0.0256_{\pm 0.0016}$	$0.3801_{\pm 0.0056}$	$0.3584_{\pm 0.0032}$	$0.3498_{\pm 0.0296}$	$0.4595_{\pm 0.0050}$

Table 24: Utility detail — Llama-3.2-1B-Instruct on Forget 5%.

Method	Utility										
	Util	Fluency	R.Prob	R.ROUGE	R.TR	RA.Prob	RA.ROUGE	RA.TR	WF.Prob	WF.ROUGE	WF.TR
Original (init)	$0.6142_{\pm 0.0000}$	$0.8611_{\pm 0.0000}$	$0.8706_{\pm 0.0000}$	$0.7970_{\pm 0.0000}$	$0.5175_{\pm 0.0000}$	$0.4137_{\pm 0.0000}$	$0.8170_{\pm 0.0000}$	$0.5274_{\pm 0.0000}$	$0.4356_{\pm 0.0000}$	$0.8279_{\pm 0.0000}$	$0.6198_{\pm 0.0000}$
Retrain (target)	$0.6192_{\pm 0.0000}$	$0.9102_{\pm 0.0000}$	$0.8765_{\pm 0.0000}$	$0.8156_{\pm 0.0000}$	$0.5182_{\pm 0.0000}$	$0.3986_{\pm 0.0000}$	$0.8395_{\pm 0.0000}$	$0.5123_{\pm 0.0000}$	$0.4375_{\pm 0.0000}$	$0.8348_{\pm 0.0000}$	$0.6225_{\pm 0.0000}$
GD-AdamW	$0.4846_{\pm 0.0056}$	$0.7630_{\pm 0.0128}$	$0.4019_{\pm 0.0162}$	$0.4082_{\pm 0.0066}$	$0.5335_{\pm 0.0051}$	$0.3464_{\pm 0.0037}$	$0.5137_{\pm 0.0036}$	$0.4820_{\pm 0.0055}$	$0.3933_{\pm 0.0041}$	$0.7887_{\pm 0.0107}$	$0.5767_{\pm 0.0077}$
GD-Sophia	$0.4085_{\pm 0.0304}$	$0.6358_{\pm 0.0155}$	$0.2126_{\pm 0.0610}$	$0.3795_{\pm 0.0130}$	$0.5479_{\pm 0.0094}$	$0.3210_{\pm 0.0005}$	$0.4681_{\pm 0.0342}$	$0.4705_{\pm 0.0040}$	$0.3466_{\pm 0.0110}$	$0.7604_{\pm 0.0263}$	$0.5735_{\pm 0.0052}$
SU-GD-Muon	$0.4383_{\pm 0.0086}$	$0.7153_{\pm 0.0218}$	$0.2674_{\pm 0.0330}$	$0.3952_{\pm 0.0150}$	$0.5328_{\pm 0.0062}$	$0.3236_{\pm 0.0040}$	$0.4984_{\pm 0.0330}$	$0.4646_{\pm 0.0060}$	$0.3576_{\pm 0.0058}$	$0.7555_{\pm 0.0095}$	$0.5601_{\pm 0.0088}$
SU-GD-KLShampoo	$0.4382_{\pm 0.0111}$	$0.7351_{\pm 0.0080}$	$0.2685_{\pm 0.0270}$	$0.3882_{\pm 0.0172}$	$0.5367_{\pm 0.0101}$	$0.3215_{\pm 0.0058}$	$0.4928_{\pm 0.0232}$	$0.4649_{\pm 0.0103}$	$0.3552_{\pm 0.0099}$	$0.7511_{\pm 0.0266}$	$0.5709_{\pm 0.0046}$
SU-GD-SOAP	$0.4641_{\pm 0.0074}$	$0.7685_{\pm 0.0056}$	$0.3707_{\pm 0.0234}$	$0.4074_{\pm 0.0118}$	$0.5426_{\pm 0.0070}$	$0.3283_{\pm 0.0052}$	$0.4674_{\pm 0.0100}$	$0.4664_{\pm 0.0027}$	$0.3676_{\pm 0.0018}$	$0.7619_{\pm 0.0024}$	$0.5587_{\pm 0.0077}$
NPO-AdamW	$0.4761_{\pm 0.0071}$	$0.8729_{\pm 0.0033}$	$0.3455_{\pm 0.0051}$	$0.3043_{\pm 0.0019}$	$0.4254_{\pm 0.0035}$	$0.3959_{\pm 0.0036}$	$0.7687_{\pm 0.0388}$	$0.5008_{\pm 0.0080}$	$0.4175_{\pm 0.0062}$	$0.7455_{\pm 0.0302}$	$0.5676_{\pm 0.0111}$
NPO-Sophia	$0.5510_{\pm 0.0015}$	$0.8928_{\pm 0.0306}$	$0.6683_{\pm 0.0043}$	$0.5052_{\pm 0.0085}$	$0.4852_{\pm 0.0026}$	$0.3694_{\pm 0.0034}$	$0.7762_{\pm 0.0165}$	$0.4702_{\pm 0.0013}$	$0.4218_{\pm 0.0084}$	$0.8051_{\pm 0.0040}$	$0.5710_{\pm 0.0100}$
SU-NPO-Muon	$0.5580_{\pm 0.0017}$	$0.9153_{\pm 0.0158}$	$0.6010_{\pm 0.0152}$	$0.4596_{\pm 0.0156}$	$0.4641_{\pm 0.0056}$	$0.4107_{\pm 0.0111}$	$0.8195_{\pm 0.0127}$	$0.5252_{\pm 0.0124}$	$0.4551_{\pm 0.0173}$	$0.8169_{\pm 0.0067}$	$0.6102_{\pm 0.0133}$
SU-NPO-KLShampoo	$0.5603_{\pm 0.0095}$	$0.9113_{\pm 0.0162}$	$0.6178_{\pm 0.0132}$	$0.4716_{\pm 0.0129}$	$0.4800_{\pm 0.0068}$	$0.4070_{\pm 0.0137}$	$0.8008_{\pm 0.0072}$	$0.5227_{\pm 0.0162}$	$0.4552_{\pm 0.0315}$	$0.8291_{\pm 0.0126}$	$0.6144_{\pm 0.0297}$
SU-NPO-SOAP	$0.5614_{\pm 0.0033}$	$0.9193_{\pm 0.0211}$	$0.6220_{\pm 0.0080}$	$0.4767_{\pm 0.0067}$	$0.4744_{\pm 0.0015}$	$0.4200_{\pm 0.0130}$	$0.7941_{\pm 0.0199}$	$0.5391_{\pm 0.0156}$	$0.4425_{\pm 0.0183}$	$0.8117_{\pm 0.0272}$	$0.6087_{\pm 0.0224}$
RMU-AdamW	$0.4848_{\pm 0.0214}$	$0.2265_{\pm 0.0527}$	$0.7501_{\pm 0.0026}$	$0.6134_{\pm 0.0071}$	$0.5033_{\pm 0.0013}$	$0.4080_{\pm 0.0022}$	$0.6987_{\pm 0.0170}$	$0.5249_{\pm 0.0031}$	$0.4305_{\pm 0.0003}$	$0.8317_{\pm 0.0059}$	$0.5956_{\pm 0.0021}$
RMU-Sophia	$0.5369_{\pm 0.0037}$	$0.5099_{\pm 0.0574}$	$0.6628_{\pm 0.0028}$	$0.5125_{\pm 0.0096}$	$0.4894_{\pm 0.0022}$	$0.4008_{\pm 0.0008}$	$0.6520_{\pm 0.0102}$	$0.5160_{\pm 0.0007}$	$0.4353_{\pm 0.0025}$	$0.8151_{\pm 0.0146}$	$0.6125_{\pm 0.0037}$
SU-RMU-Muon	$0.5390_{\pm 0.0035}$	$0.7280_{\pm 0.0208}$	$0.5863_{\pm 0.0065}$	$0.4938_{\pm 0.0072}$	$0.4958_{\pm 0.0010}$	$0.3875_{\pm 0.0058}$	$0.6397_{\pm 0.0285}$	$0.5005_{\pm 0.0041}$	$0.4224_{\pm 0.0011}$	$0.8302_{\pm 0.0093}$	$0.5841_{\pm 0.0029}$
SU-RMU-KLShampoo	$0.5455_{\pm 0.0009}$	$0.6696_{\pm 0.0278}$	$0.6193_{\pm 0.0063}$	$0.5152_{\pm 0.0044}$	$0.5003_{\pm 0.0007}$	$0.3928_{\pm 0.0056}$	$0.6656_{\pm 0.0207}$	$0.5077_{\pm 0.0043}$	$0.4242_{\pm 0.0009}$	$0.8311_{\pm 0.0107}$	$0.5891_{\pm 0.0032}$
SU-RMU-SOAP	$0.5523_{\pm 0.0100}$	$0.4196_{\pm 0.0572}$	$0.

Table 25: Utility detail — Llama-3.2-1B-Instruct on Forget 1%.

Method	Utility									
	Util	Fluency	R.Prob	R.ROUGE	R.TR	RA.Prob	RA.ROUGE	RA.TR	WF.Prob	WF.ROUGE	WF.TR
Original (init)	$0.6154_{\pm 0.0000}$	$0.8952_{\pm 0.0000}$	$0.8706_{\pm 0.0000}$	$0.7970_{\pm 0.0000}$	$0.5175_{\pm 0.0000}$	$0.4137_{\pm 0.0000}$	$0.8170_{\pm 0.0000}$	$0.5274_{\pm 0.0000}$	$0.4356_{\pm 0.0000}$	$0.8279_{\pm 0.0000}$	$0.6198_{\pm 0.0000}$
Retrain (target)	$0.6172_{\pm 0.0000}$	$0.8720_{\pm 0.0000}$	$0.8790_{\pm 0.0000}$	$0.8080_{\pm 0.0000}$	$0.5151_{\pm 0.0000}$	$0.4121_{\pm 0.0000}$	$0.7620_{\pm 0.0000}$	$0.5215_{\pm 0.0000}$	$0.4380_{\pm 0.0000}$	$0.8519_{\pm 0.0000}$	$0.6212_{\pm 0.0000}$
GD-AdamW	$0.4877_{\pm 0.0015}$	$0.7153_{\pm 0.0437}$	$0.3565_{\pm 0.0063}$	$0.4076_{\pm 0.0033}$	$0.5254_{\pm 0.0094}$	$0.3659_{\pm 0.0041}$	$0.5998_{\pm 0.0306}$	$0.5006_{\pm 0.0095}$	$0.4002_{\pm 0.0079}$	$0.7780_{\pm 0.0187}$	$0.5844_{\pm 0.0028}$
GD-Sophia	$0.4659_{\pm 0.0040}$	$0.6854_{\pm 0.0088}$	$0.2970_{\pm 0.0175}$	$0.3873_{\pm 0.0014}$	$0.5209_{\pm 0.0069}$	$0.3658_{\pm 0.0035}$	$0.5436_{\pm 0.0255}$	$0.5008_{\pm 0.0066}$	$0.3975_{\pm 0.0096}$	$0.7732_{\pm 0.0143}$	$0.5830_{\pm 0.0071}$
SU-GD-Muon	$0.5194_{\pm 0.0028}$	$0.8232_{\pm 0.0391}$	$0.4968_{\pm 0.0312}$	$0.4528_{\pm 0.0001}$	$0.5318_{\pm 0.0051}$	$0.3699_{\pm 0.0057}$	$0.6196_{\pm 0.0175}$	$0.4965_{\pm 0.0056}$	$0.3987_{\pm 0.0108}$	$0.7887_{\pm 0.0164}$	$0.5705_{\pm 0.0135}$
SU-GD-KLShampoo	$0.5190_{\pm 0.0045}$	$0.7602_{\pm 0.0508}$	$0.4848_{\pm 0.0127}$	$0.4602_{\pm 0.0140}$	$0.5288_{\pm 0.0117}$	$0.3697_{\pm 0.0055}$	$0.6149_{\pm 0.0126}$	$0.4932_{\pm 0.0071}$	$0.4055_{\pm 0.0104}$	$0.8061_{\pm 0.0099}$	$0.5893_{\pm 0.0105}$
SU-GD-SOAP	$0.5304_{\pm 0.0027}$	$0.7714_{\pm 0.0529}$	$0.5375_{\pm 0.0121}$	$0.4768_{\pm 0.0065}$	$0.5317_{\pm 0.0103}$	$0.3750_{\pm 0.0038}$	$0.6359_{\pm 0.0249}$	$0.4973_{\pm 0.0053}$	$0.4061_{\pm 0.0127}$	$0.8056_{\pm 0.0209}$	$0.5922_{\pm 0.0137}$
NPO-AdamW	$0.6121_{\pm 0.0005}$	$0.9102_{\pm 0.0200}$	$0.8461_{\pm 0.0033}$	$0.7610_{\pm 0.0066}$	$0.5091_{\pm 0.0014}$	$0.4156_{\pm 0.0004}$	$0.8203_{\pm 0.0047}$	$0.5235_{\pm 0.0008}$	$0.4455_{\pm 0.0002}$	$0.8146_{\pm 0.0042}$	$0.6242_{\pm 0.0012}$
NPO-Sophia	$0.5781_{\pm 0.0029}$	$0.9075_{\pm 0.0123}$	$0.6820_{\pm 0.0162}$	$0.5276_{\pm 0.0160}$	$0.4848_{\pm 0.0023}$	$0.4148_{\pm 0.0017}$	$0.7920_{\pm 0.0041}$	$0.5223_{\pm 0.0040}$	$0.4514_{\pm 0.0009}$	$0.8236_{\pm 0.0099}$	$0.6263_{\pm 0.0022}$
SU-NPO-Muon	$0.5861_{\pm 0.0009}$	$0.9245_{\pm 0.0252}$	$0.7198_{\pm 0.0095}$	$0.5606_{\pm 0.0125}$	$0.4934_{\pm 0.0034}$	$0.4149_{\pm 0.0018}$	$0.8020_{\pm 0.0215}$	$0.5204_{\pm 0.0025}$	$0.4500_{\pm 0.0052}$	$0.8173_{\pm 0.0107}$	$0.6311_{\pm 0.0047}$
SU-NPO-KLShampoo	$0.5633_{\pm 0.0128}$	$0.8919_{\pm 0.0159}$	$0.6060_{\pm 0.0811}$	$0.5046_{\pm 0.0035}$	$0.4733_{\pm 0.0130}$	$0.4103_{\pm 0.0034}$	$0.7993_{\pm 0.0437}$	$0.5146_{\pm 0.0088}$	$0.4425_{\pm 0.0134}$	$0.8267_{\pm 0.0018}$	$0.6122_{\pm 0.0119}$
SU-NPO-SOAP	$0.5945_{\pm 0.0051}$	$0.9084_{\pm 0.0161}$	$0.7452_{\pm 0.0402}$	$0.6395_{\pm 0.0042}$	$0.4980_{\pm 0.0078}$	$0.4128_{\pm 0.0012}$	$0.8153_{\pm 0.0136}$	$0.5178_{\pm 0.0059}$	$0.4427_{\pm 0.0037}$	$0.8169_{\pm 0.0056}$	$0.6195_{\pm 0.0061}$
RMU-AdamW	$0.5422_{\pm 0.0076}$	$0.6302_{\pm 0.0555}$	$0.6358_{\pm 0.0119}$	$0.5128_{\pm 0.0039}$	$0.4838_{\pm 0.0019}$	$0.3925_{\pm 0.0005}$	$0.6919_{\pm 0.0235}$	$0.5072_{\pm 0.0008}$	$0.4244_{\pm 0.0019}$	$0.8108_{\pm 0.0081}$	$0.5884_{\pm 0.0015}$
RMU-Sophia	$0.4925_{\pm 0.0000}$	$0.8604_{\pm 0.0000}$	$0.4359_{\pm 0.0000}$	$0.3889_{\pm 0.0000}$	$0.4046_{\pm 0.0000}$	$0.3995_{\pm 0.0000}$	$0.5662_{\pm 0.0000}$	$0.4982_{\pm 0.0000}$	$0.4145_{\pm 0.0000}$	$0.7866_{\pm 0.0000}$	$0.5419_{\pm 0.0000}$
SU-RMU-Muon	$0.5578_{\pm 0.0026}$	$0.6681_{\pm 0.0337}$	$0.6656_{\pm 0.0110}$	$0.5491_{\pm 0.0030}$	$0.4926_{\pm 0.0005}$	$0.4030_{\pm 0.0009}$	$0.7038_{\pm 0.0107}$	$0.5256_{\pm 0.0017}$	$0.4213_{\pm 0.0017}$	$0.8251_{\pm 0.0107}$	$0.6007_{\pm 0.0017}$
SU-RMU-KLShampoo	$0.5765_{\pm 0.0025}$	$0.5877_{\pm 0.0145}$	$0.7711_{\pm 0.0099}$	$0.6616_{\pm 0.0036}$	$0.5058_{\pm 0.0013}$	$0.4105_{\pm 0.0015}$	$0.7512_{\pm 0.0201}$	$0.5292_{\pm 0.0022}$	$0.4312_{\pm 0.0021}$	$0.8393_{\pm 0.0040}$	$0.6080_{\pm 0.0030}$
SU-RMU-SOAP	$0.5901_{\pm 0.0017}$	$0.6749_{\pm 0.0345}$	$0.7950_{\pm 0.0069}$	$0.7000_{\pm 0.0036}$	$0.5160_{\pm 0.0029}$	$0.4125_{\pm 0.0038}$	$0.7795_{\pm 0.0108}$	$0.5327_{\pm 0.0047}$	$0.4338_{\pm 0.0040}$	$0.8279_{\pm 0.0070}$	$0.6096_{\pm 0.0040}$
UNDIAL	$0.3722_{\pm 0.0053}$	$0.9216_{\pm 0.0236}$	$0.1967_{\pm 0.0046}$	$0.2742_{\pm 0.0109}$	$0.3003_{\pm 0.0031}$	$0.3471_{\pm 0.0047}$	$0.4247_{\pm 0.0518}$	$0.3883_{\pm 0.0086}$	$0.4022_{\pm 0.0036}$	$0.7200_{\pm 0.0250}$	$0.5030_{\pm 0.0057}$
SatImp	$0.6139_{\pm 0.0010}$	$0.8629_{\pm 0.0131}$	$0.8557_{\pm 0.0013}$	$0.7961_{\pm 0.0023}$	$0.5140_{\pm 0.0006}$	$0.4169_{\pm 0.0003}$	$0.8162_{\pm 0.0047}$	$0.5251_{\pm 0.0002}$	$0.4447_{\pm 0.0006}$	$0.8279_{\pm 0.0070}$	$0.6243_{\pm 0.0002}$
WGA	$0.6147_{\pm 0.0005}$	$0.8916_{\pm 0.0037}$	$0.8680_{\pm 0.0016}$	$0.7755_{\pm 0.0068}$	$0.5228_{\pm 0.0019}$	$0.4147_{\pm 0.0025}$	$0.7670_{\pm 0.0141}$	$0.5358_{\pm 0.0022}$	$0.4374_{\pm 0.0025}$	$0.8336_{\pm 0.0081}$	$0.6273_{\pm 0.0023}$
AltPO	$0.5947_{\pm 0.0003}$	$0.8943_{\pm 0.0026}$	$0.7652_{\pm 0.0011}$	$0.6960_{\pm 0.0007}$	$0.4672_{\pm 0.0006}$	$0.4138_{\pm 0.0002}$	$0.8495_{\pm 0.0000}$	$0.5145_{\pm 0.0012}$	$0.4416_{\pm 0.0004}$	$0.8382_{\pm 0.0037}$	$0.6058_{\pm 0.0009}$
SimNPO	$0.6140_{\pm 0.0007}$	$0.9155_{\pm 0.0135}$	$0.8639_{\pm 0.0016}$	$0.7549_{\pm 0.0096}$	$0.5204_{\pm 0.0017}$	$0.4112_{\pm 0.0021}$	$0.7803_{\pm 0.0047}$	$0.5322_{\pm 0.0021}$	$0.4354_{\pm 0.0013}$	$0.8441_{\pm 0.0132}$	$0.6266_{\pm 0.0010}$
PDU	$0.5975_{\pm 0.0018}$	$0.8261_{\pm 0.0258}$	$0.7320_{\pm 0.0028}$	$0.6906_{\pm 0.0018}$	$0.5119_{\pm 0.0011}$	$0.4347_{\pm 0.0009}$	$0.7587_{\pm 0.0131}$	$0.5402_{\pm 0.0015}$	$0.4882_{\pm 0.0009}$	$0.8348_{\pm 0.0040}$	$0.6702_{\pm 0.0011}$
IdkDPO	$0.5987_{\pm 0.0007}$	$0.9016_{\pm 0.0058}$	$0.8138_{\pm 0.0007}$	$0.7249_{\pm 0.0014}$	$0.4852_{\pm 0.0006}$	$0.4021_{\pm 0.0015}$	$0.8512_{\pm 0.0047}$	$0.5029_{\pm 0.0011}$	$0.4525_{\pm 0.0027}$	$0.8108_{\pm 0.0000}$	$0.6171_{\pm 0.0036}$
GradAscent	$0.6135_{\pm 0.0007}$	$0.8727_{\pm 0.0155}$	$0.8607_{\pm 0.0012}$	$0.7952_{\pm 0.0046}$	$0.5120_{\pm 0.0006}$	$0.4142_{\pm 0.0000}$	$0.8070_{\pm 0.0082}$	$0.5250_{\pm 0.0011}$	$0.4396_{\pm 0.0007}$	$0.8194_{\pm 0.0000}$	$0.6192_{\pm 0.0014}$
CEU	$0.6129_{\pm 0.0011}$	$0.8601_{\pm 0.0257}$	$0.8426_{\pm 0.0005}$	$0.8064_{\pm 0.0010}$	$0.5098_{\pm 0.0006}$	$0.4161_{\pm 0.0005}$	$0.8073_{\pm 0.0129}$	$0.5226_{\pm 0.0016}$	$0.4445_{\pm 0.0006}$	$0.8256_{\pm 0.0032}$	$0.6231_{\pm 0.0005}$

Table 26: Utility detail — Llama-3.2-3B-Instruct on Forget 10%.

Method	Utility									
	Util	Fluency	R.Prob	R.ROUGE	R.TR	RA.Prob	RA.ROUGE	RA.TR	WF.Prob	WF.ROUGE	WF.TR
Original (init)	$0.6662_{\pm 0.0000}$	$0.8614_{\pm 0.0000}$	$0.9429_{\pm 0.0000}$	$0.9197_{\pm 0.0000}$	$0.5409_{\pm 0.0000}$	$0.4795_{\pm 0.0000}$	$0.8513_{\pm 0.0000}$	$0.6244_{\pm 0.0000}$	$0.4941_{\pm 0.0000}$	$0.8932_{\pm 0.0000}$	$0.6622_{\pm 0.0000}$
Retrain (target)	$0.6678_{\pm 0.0000}$	$0.8895_{\pm 0.0000}$	$0.9420_{\pm 0.0000}$	$0.9192_{\pm 0.0000}$	$0.5352_{\pm 0.0000}$	$0.4615_{\pm 0.0000}$	$0.8515_{\pm 0.0000}$	$0.5951_{\pm 0.0000}$	$0.4732_{\pm 0.0000}$	$0.8818_{\pm 0.0000}$	$0.6396_{\pm 0.0000}$
GD-AdamW	$0.6348_{\pm 0.0007}$	$0.8561_{\pm 0.0057}$	$0.8326_{\pm 0.0013}$	$0.6652_{\pm 0.0039}$	$0.5665_{\pm 0.0011}$	$0.4587_{\pm 0.0040}$	$0.7660_{\pm 0.0059}$	$0.6091_{\pm 0.0059}$	$0.4796_{\pm 0.0035}$	$0.8566_{\pm 0.0035}$	$0.6513_{\pm 0.0043}$
GD-Sophia	$0.4658_{\pm 0.0034}$	$0.3179_{\pm 0.0066}$	$0.2979_{\pm 0.0075}$	$0.3861_{\pm 0.0035}$	$0.6049_{\pm 0.0029}$	$0.4635_{\pm 0.0132}$	$0.5996_{\pm 0.0248}$	$0.6242_{\pm 0.0121}$	$0.4599_{\pm 0.0068}$	$0.8215_{\pm 0.0102}$	$0.6392_{\pm 0.0030}$
SU-GD-Muon	$0.5839_{\pm 0.0030}$	$0.7460_{\pm 0.0057}$	$0.6030_{\pm 0.0120}$	$0.5223_{\pm 0.0073}$	$0.5941_{\pm 0.0012}$	$0.4634_{\pm 0.0119}$	$0.6860_{\pm 0.0066}$	$0.6175_{\pm 0.0136}$	$0.4480_{\pm 0.0023}$	$0.8378_{\pm 0.0017}$	$0.6290_{\pm 0.0028}$
SU-GD-KLShampoo	$0.5781_{\pm 0.0037}$	$0.7363_{\pm 0.0107}$	$0.5820_{\pm 0.0106}$	$0.5125_{\pm 0.0022}$	$0.5949_{\pm 0.0015}$	$0.4635_{\pm 0.0128}$	$0.6738_{\pm 0.0212}$	$0.6199_{\pm 0.0148}$	$0.4454_{\pm 0.0016}$	$0.8364_{\pm 0.0173}$	$0.6240_{\pm 0.0043}$
SU-GD-SOAP	$0.4799_{\pm 0.0712}$	$0.3401_{\pm 0.1502}$	$0.4637_{\pm 0.0902}$	$0.4277_{\pm 0.0926}$	$0.6129_{\pm 0.0037}$	$0.4992_{\pm 0.0043}$	$0.5626_{\pm 0.0271}$	$0.6732_{\pm 0.0100}$	$0.4301_{\pm 0.0107}$	$0.7658_{\pm 0.0568}$	$0.6358_{\pm 0.0158}$
NPO-AdamW	$0.6608_{\pm 0.0010}$	$0.9003_{\pm 0.0037}$	$0.9193_{\pm 0.0048}$	$0.8330_{\pm 0.0183}$	$0.5671_{\pm 0.0054}$	$0.5042_{\pm 0.0091}$	$0.8466_{\pm 0.0166}$	$0.6572_{\pm 0.0066}$	$0.4792_{\pm 0.0058}$	$0.9004_{\pm 0.0072}$	$0.6424_{\pm 0.0062}$
NPO-Sophia	$0.6538_{\pm 0.0021}$	$0.9098_{\pm 0.0044}$	$0.8815_{\pm 0.0060}$	$0.7597_{\pm 0.0157}$	$0.5677_{\pm 0.0058}$	$0.4952_{\pm 0.0161}$	$0.8467_{\pm 0.0168}$	$0.6436_{\pm 0.0206}$	$0.4874_{\pm 0.0095}$	$0.8899_{\pm 0.0115}$	$0.6514_{\pm 0.0104}$
SU-NPO-Muon	$0.6528_{\pm 0.0011}$	$0.9096_{\pm 0.0067}$	$0.8740_{\pm 0.0055}$	$0.7498_{\pm 0.0161}$	$0.5665_{\pm 0.0036}$	$0.5069_{\pm 0.0146}$	$0.8474_{\pm 0.0140}$	$0.6588_{\pm 0.0215}$	$0.5144_{\pm 0.0257}$	$0.8896_{\pm 0.0117}$	$0.6879_{\pm 0.0307}$
SU-NPO-KLShampoo	$0.6544_{\pm 0.0018}$	$0.8948_{\pm 0.0143}$	$0.8833_{\pm 0.0069}$	$0.7632_{\pm 0.0187}$	$0.5637_{\pm 0.0086}$	$0.5057_{\pm 0.0109}$	$0.8638_{\pm 0.0113}$	$0.6542_{\pm 0.0138}$	$0.5067_{\pm 0.0225}$	$0.8996_{\pm 0.0188}$	$0.6759_{\pm 0.0318}$
SU-NPO-SOAP	$0.6609_{\pm 0.0016}$	$0.8680_{\pm 0.0184}$	$0.9270_{\pm 0.0082}$	$0.8621_{\pm 0.0236}$	$0.5677_{\pm 0.0046}$	$0.5179_{\pm 0.0151}$	$0.8592_{\pm 0.0123}$	$0.6677_{\pm 0.0256}$	$0.5270_{\pm 0.0019}$	$0.8592_{\pm 0.0140}$	$0.6967_{\pm 0.0077}$
RMU-AdamW	$0.6322_{\pm 0.0045}$	$0.5573_{\pm 0.0351}$	$0.9236_{\pm 0.0006}$	$0.8719_{\pm 0.0029}$	$0.5384_{\pm 0.0003}$	$0.4900_{\pm 0.0005}$	$0.7997_{\pm 0.0062}$	$0.6322_{\pm 0.0022}$	$0.5050_{\pm 0.0012}$	$0.8732_{\pm 0.0000}$	$0.6736_{\pm 0.0016}$
RMU-Sophia	$0.5945_{\pm 0.0145}$	$0.3536_{\pm 0.0563}$	$0.9252_{\pm 0.0007}$	$0.8786_{\pm 0.0022}$	$0.5349_{\pm 0.0005}$	$0.4858_{\pm 0.0004}$	$0.8404_{\pm 0.0096}$	$0.6248_{\pm 0.0009}$	$0.5091_{\pm 0.0004}$	$0.8871_{\pm 0.0024}$	$0.6706_{\pm 0.0011}$
SU-RMU-Muon	$0.6390_{\pm 0.0021}$	$0.6113_{\pm 0.0216}$	$0.9154_{\pm 0.0009}$	$0.8639_{\pm 0.0026}$	$0.5363_{\pm 0.0006}$	$0.4917_{\pm 0.0012}$	$0.8163_{\pm 0.0041}$	$0.6374_{\pm 0.0031}$	$0.5013_{\pm 0.0003}$	$0.8732_{\pm 0.0000}$	$0.6728_{\pm 0.0022}$
SU-RMU-KLShampoo	$0.6369_{\pm 0.0026}$	$0.5937_{\pm 0.0237}$	$0.9172_{\pm 0.0005}$	$0.8653_{\pm 0.0008}$	$0.5368_{\pm 0.0009}$	$0.4917_{\pm 0.0013}$	$0.8102_{\pm 0.0093}$	$0.6364_{\pm 0.0027}$	$0.5033_{\pm 0.0005}$	$0.8761_{\pm 0.0040}$	$0.6743_{\pm 0.0019}$
SU-RMU-SOAP	$0.6484_{\pm 0.0002}$	$0.6617_{\pm 0.0092}$	$0.9322_{\pm 0.0003}$	$0.8978_{\pm 0.0002}$	$0.5359_{\pm 0.0008}$	$0.4989_{\pm 0.0035}$	$0.8390_{\pm 0.0152}$	$0.6445_{\pm 0.0065}$	$0.5164_{\pm 0.0020}$	$0.8989_{\pm 0.0000}$	$0.6858_{\pm 0.0006}$
UNDIAL	$0.3973_{\pm 0.0023}$	$0.7927_{\pm 0.0049}$	$0.2154_{\pm 0.0009}$	$0.3245_{\pm 0.0021}$	$0.3289_{\pm 0.0001}$	$0.3873_{\pm 0.0048}$	$0.3573_{\pm 0.0157}$	$0.4585_{\pm 0.0069}$	$0.4242_{\pm 0.0024}$	$0.6959_{\pm 0.0034}$	$0.5607_{\pm 0.0023}$
SatImp	$0.5948_{\pm 0.0082}$	$0.9081_{\pm 0.0055}$	$0.4543_{\pm 0.0300}$	$0.6478_{\pm 0.0155}$	$0.4807_{\pm 0.0067}$	$0.5691_{\pm 0.0083}$	$0.8416_{\pm 0.0190}$	$0.7271_{\pm 0.0099}$	$0.5815_{\pm 0.0085}$	$0.8824_{\pm 0.0088}$	$0.7692_{\pm 0.0100}$
WGA	$0.6384_{\pm 0.0102}$	$0.7550_{\pm 0.0370}$	$0.9143_{\pm 0.0188}$	$0.8863_{\pm 0.0165}$	$0.4968_{\pm 0.0265}$	$0.4447_{\pm 0.0049}$	$0.7858_{\pm 0.0168}$	$0.5725_{\pm 0.0071}$	$0.5306_{\pm 0.0165}$	$0.8668_{\pm 0.0076}$	$0.7073_{\pm 0.0228}$
AltPO	$0.6136_{\pm 0.0020}$	$0.8817_{\pm 0.0004}$	$0.7577_{\pm 0.0023}$	$0.5967_{\pm 0.0100}$	$0.4340_{\pm 0.0015}$	$0.5376_{\pm 0.0077}$	$0.8420_{\pm 0.0000}$	$0.6859_{\pm 0.0101}$	$0.4968_{\pm 0.0019}$	$0.8699_{\pm 0.0101}$	$0.6641_{\pm 0.0044}$
SimNPO	$0.6617_{\pm 0.0009}$	$0.8916_{\pm 0.0128}$	$0.9487_{\pm 0.0002}$	$0.9066_{\pm 0.0080}$	$0.5548_{\pm 0.0009}$	$0.4663_{\pm 0.0054}$	$0.7967_{\pm 0.0127}$	$0.6138_{\pm 0.0067}$	$0.4949_{\pm 0.0011}$	$0.8561_{\pm 0.0000}$	$0.6505_{\pm 0.0029}$
PDU	$0.6236_{\pm 0.0115}$	$0.7168_{\pm 0.1551}$	$0.7605_{\pm 0.0085}$	$0.7061_{\pm 0.0202}$	$0.5298_{\pm 0.0054}$	$0.5228_{\pm 0.0084}$	$0.7902_{\pm 0.0215}$	$0.6573_{\pm 0.0199}$	$0.5916_{\pm 0.0014}$	$0.8695_{\pm 0.0060}$	$0.7628_{\pm 0.0035}$
IdkDPO	$0.6440_{\pm 0.0079}$	$0.9151_{\pm 0.0065}$	$0.9429_{\pm 0.0004}$	$0.8855_{\pm 0.0086}$	$0.5169_{\pm 0.0023}$	$0.4202_{\pm 0.0132}$	$0.8433_{\pm 0.0183}$	$0.5414_{\pm 0.0199}$	$0.4761_{\pm 0.0089}$	$0.8666_{\pm 0.0094}$	$0.6385_{\pm 0.0080}$
GradAscent	$0.6092_{\pm 0.0012}$	$0.9038_{\pm 0.0014}$	$0.6236_{\pm 0.0024}$	$0.5849_{\pm 0.0074}$	$0.4686_{\pm 0.0014}$	$0.4623_{\pm 0.0040}$	$0.8957_{\pm 0.0103}$	$0.5995_{\pm 0.0057}$	$0.4969_{\pm 0.0069}$	$0.9123_{\pm 0.0099}$	$0.6696_{\pm 0.0092}$
CEU	$0.0000_{\pm 0.0000}$	$0.0630_{\pm 0.0068}$	$0.0000_{\pm 0.0000}$	$0.0036_{\pm 0.0000}$	$0.1751_{\pm 0.0018}$	$0.3297_{\pm 0.0023}$	$0.0000_{\pm 0.0000}$	$0.3486_{\pm 0.0055}$	$0.3645_{\pm 0.0030}$	$0.0057_{\pm 0.0040}$	$0.4135_{\pm 0.0043}$

Table 27: Utility detail — Llama-3.2-3B-Instruct on Forget 5%.

Method	Utility									
	Util	Fluency	R.Prob	R.ROUGE	R.TR	RA.Prob	RA.ROUGE	RA.TR	WF.Prob	WF.ROUGE	WF.TR
Original (init)	$0.6740_{\pm 0.0000}$	$0.8535_{\pm 0.0000}$	$0.9430_{\pm 0.0000}$	$0.9151_{\pm 0.0000}$	$0.5404_{\pm 0.0000}$	$0.4789_{\pm 0.0000}$	$0.8513_{\pm 0.0000}$	$0.6244_{\pm 0.0000}$	$0.4934_{\pm 0.0000}$	$0.8778_{\pm 0.0000}$	$0.6642_{\pm 0.0000}$
Retrain (target)	$0.6753_{\pm 0.0000}$	$0.8665_{\pm 0.0000}$	$0.9420_{\pm 0.0000}$	$0.9117_{\pm 0.0000}$	$0.5343_{\pm 0.0000}$	$0.4769_{\pm 0.0000}$	$0.8413_{\pm 0.0000}$	$0.6281_{\pm 0.0000}$	$0.4823_{\pm 0.0000}$	$0.8846_{\pm 0.0000}$	$0.6494_{\pm 0.0000}$
GD-AdamW	$0.3570_{\pm 0.0393}$	$0.1814_{\pm 0.0168}$	$0.1758_{\pm 0.0554}$	$0.2574_{\pm 0.0278}$	$0.6039_{\pm 0.0086}$	$0.4976_{\pm 0.0072}$	$0.5159_{\pm 0.0584}$	$0.6484_{\pm 0.0119}$	$0.4707_{\pm 0.0056}$	$0.7230_{\pm 0.0326}$	$0.6750_{\pm 0.0021}$
GD-Sophia	$0.4228_{\pm 0.0651}$	$0.2826_{\pm 0.0621}$	$0.2390_{\pm 0.0855}$	$0.3247_{\pm 0.0617}$	$0.6149_{\pm 0.0080}$	$0.4898_{\pm 0.0118}$	$0.5760_{\pm 0.0496}$	$0.6389_{\pm 0.0123}$	$0.4644_{\pm 0.0087}$	$0.7771_{\pm 0.0501}$	$0.6515_{\pm 0.0139}$
SU-GD-Muon	$0.3902_{\pm 0.0575}$	$0.2234_{\pm 0.0605}$	$0.2183_{\pm 0.0754}$	$0.2766_{\pm 0.0531}$	$0.6102_{\pm 0.0026}$	$0.4977_{\pm 0.0046}$	$0.5446_{\pm 0.0400}$	$0.6394_{\pm 0.0078}$	$0.4717_{\pm 0.0060}$	$0.6963_{\pm 0.0778}$	$0.6614_{\pm 0.0044}$
SU-GD-KLShampoo	$0.3922_{\pm 0.0707}$	$0.2326_{\pm 0.0942}$	$0.2279_{\pm 0.0853}$	$0.2820_{\pm 0.0570}$	$0.6086_{\pm 0.0026}$	$0.4956_{\pm 0.0036}$	$0.5547_{\pm 0.0428}$	$0.6407_{\pm 0.0114}$	$0.4664_{\pm 0.0080}$	$0.6906_{\pm 0.1033}$	$0.6630_{\pm 0.0070}$
SU-GD-SOAP	$0.5455_{\pm 0.0220}$	$0.5033_{\pm 0.0889}$	$0.4977_{\pm 0.0452}$	$0.4587_{\pm 0.0325}$	$0.6079_{\pm 0.0016}$	$0.5181_{\pm 0.0084}$	$0.6501_{\pm 0.0243}$	$0.6683_{\pm 0.0047}$	$0.4503_{\pm 0.0042}$	$0.8187_{\pm 0.0141}$	$0.6520_{\pm 0.0072}$
NPO-AdamW	$0.6025_{\pm 0.0045}$	$0.9200_{\pm 0.0139}$	$0.5071_{\pm 0.0091}$	$0.5308_{\pm 0.0381}$	$0.5126_{\pm 0.0021}$	$0.4685_{\pm 0.0036}$	$0.8910_{\pm 0.0043}$	$0.6155_{\pm 0.0055}$	$0.4833_{\pm 0.0021}$	$0.9008_{\pm 0.0071}$	$0.6604_{\pm 0.0019}$
NPO-Sophia	$0.6473_{\pm 0.0044}$	$0.9217_{\pm 0.0037}$	$0.8250_{\pm 0.0065}$	$0.6709_{\pm 0.0057}$	$0.5372_{\pm 0.0058}$	$0.4794_{\pm 0.0234}$	$0.8289_{\pm 0.0074}$	$0.6212_{\pm 0.0279}$	$0.4798_{\pm 0.0147}$	$0.8937_{\pm 0.0057}$	$0.6479_{\pm 0.0146}$
SU-NPO-Muon	$0.6130_{\pm 0.0027}$	$0.9366_{\pm 0.0087}$	$0.6884_{\pm 0.0109}$	$0.5421_{\pm 0.0128}$	$0.5109_{\pm 0.0030}$	$0.4375_{\pm 0.0058}$	$0.8932_{\pm 0.0030}$	$0.5653_{\pm 0.0080}$	$0.4933_{\pm 0.0078}$	$0.8929_{\pm 0.0087}$	$0.6604_{\pm 0.0128}$
SU-NPO-KLShampoo	$0.6082_{\pm 0.0034}$	$0.9333_{\pm 0.0099}$	$0.6956_{\pm 0.0154}$	$0.5549_{\pm 0.0144}$	$0.5097_{\pm 0.0034}$	$0.4240_{\pm 0.0049}$	$0.8891_{\pm 0.0097}$	$0.5463_{\pm 0.0062}$	$0.4822_{\pm 0.0117}$	$0.9020_{\pm 0.0053}$	$0.6471_{\pm 0.0155}$
SU-NPO-SOAP	$0.6509_{\pm 0.0038}$	$0.9172_{\pm 0.0052}$	$0.8007_{\pm 0.0192}$	$0.6538_{\pm 0.0250}$	$0.5504_{\pm 0.0080}$	$0.5022_{\pm 0.0064}$	$0.8978_{\pm 0.0156}$	$0.6561_{\pm 0.0125}$	$0.5543_{\pm 0.0147}$	$0.8963_{\pm 0.0162}$	$0.7383_{\pm 0.0168}$
RMU-AdamW	$0.6444_{\pm 0.0048}$	$0.5632_{\pm 0.0359}$	$0.9295_{\pm 0.0006}$	$0.8872_{\pm 0.0031}$	$0.5373_{\pm 0.0007}$	$0.4930_{\pm 0.0029}$	$0.8293_{\pm 0.0019}$	$0.6368_{\pm 0.0023}$	$0.5081_{\pm 0.0017}$	$0.8658_{\pm 0.0073}$	$0.6762_{\pm 0.0023}$
RMU-Sophia	$0.6474_{\pm 0.0008}$	$0.7178_{\pm 0.0007}$	$0.8611_{\pm 0.0051}$	$0.7578_{\pm 0.0093}$	$0.5303_{\pm 0.0005}$	$0.4978_{\pm 0.0015}$	$0.8280_{\pm 0.0248}$	$0.6457_{\pm 0.0020}$	$0.5183_{\pm 0.0017}$	$0.8521_{\pm 0.0040}$	$0.6848_{\pm 0.0021}$
SU-RMU-Muon	$0.6486_{\pm 0.0040}$	$0.6371_{\pm 0.0409}$	$0.9098_{\pm 0.0005}$	$0.8514_{\pm 0.0013}$	$0.5308_{\pm 0.0004}$	$0.4929_{\pm 0.0022}$	$0.8097_{\pm 0.0047}$	$0.6421_{\pm 0.0029}$	$0.5018_{\pm 0.0013}$	$0.8704_{\pm 0.0040}$	$0.6769_{\pm 0.0003}$
SU-RMU-KLShampoo	$0.6458_{\pm 0.0037}$	$0.5761_{\pm 0.0285}$	$0.9183_{\pm 0.0006}$	$0.8753_{\pm 0.0050}$	$0.5320_{\pm 0.0007}$	$0.4907_{\pm 0.0011}$	$0.8457_{\pm 0.0066}$	$0.6413_{\pm 0.0013}$	$0.5090_{\pm 0.0029}$	$0.8740_{\pm 0.0036}$	$0.6724_{\pm 0.0032}$
SU-RMU-SOAP	$0.6673_{\pm 0.0016}$	$0.7988_{\pm 0.0116}$	$0.9264_{\pm 0.0009}$	$0.8858_{\pm 0.0047}$	$0.5358_{\pm 0.0008}$	$0.4852_{\pm 0.0027}$	$0.8363_{\pm 0.0216}$	$0.6265_{\pm 0.0045}$	$0.5086_{\pm 0.0021}$	$0.8770_{\pm 0.0036}$	$0.6800_{\pm 0.0018}$
UNDIAL	$0.4165_{\pm 0.0111}$	$0.7880_{\pm 0.0134}$	$0.2328_{\pm 0.0078}$	$0.3218_{\pm 0.0052}$	$0.3215_{\pm 0.0052}$	$0.4071_{\pm 0.0098}$	$0.4057_{\pm 0.0555}$	$0.4836_{\pm 0.0113}$	$0.4469_{\pm 0.0018}$	$0.7776_{\pm 0.0197}$	$0.5954_{\pm 0.0060}$
SatImp	$0.6159_{\pm 0.0021}$	$0.8948_{\pm 0.0029}$	$0.5721_{\pm 0.0057}$	$0.5872_{\pm 0.0148}$	$0.5028_{\pm 0.0010}$	$0.4758_{\pm 0.0006}$	$0

Table 28: Utility detail — Llama-3.2-3B-Instruct on Forget 1%.

Method	Utility										
	Util	Fluency	R.Prob	R.ROUGE	R.TR	RA.Prob	RA.ROUGE	RA.TR	WF.Prob	WF.ROUGE	WF.TR
Original (init)	$0.6785_{\pm 0.0000}$	$0.8938_{\pm 0.0000}$	$0.9430_{\pm 0.0000}$	$0.9151_{\pm 0.0000}$	$0.5404_{\pm 0.0000}$	$0.4789_{\pm 0.0000}$	$0.8513_{\pm 0.0000}$	$0.6244_{\pm 0.0000}$	$0.4934_{\pm 0.0000}$	$0.8778_{\pm 0.0000}$	$0.6642_{\pm 0.0000}$
Retrain (target)	$0.6787_{\pm 0.0000}$	$0.8833_{\pm 0.0000}$	$0.9425_{\pm 0.0000}$	$0.9156_{\pm 0.0000}$	$0.5381_{\pm 0.0000}$	$0.4763_{\pm 0.0000}$	$0.8780_{\pm 0.0000}$	$0.6208_{\pm 0.0000}$	$0.4883_{\pm 0.0000}$	$0.8761_{\pm 0.0000}$	$0.6522_{\pm 0.0000}$
GD-AdamW	$0.4686_{\pm 0.0266}$	$0.3560_{\pm 0.0777}$	$0.2725_{\pm 0.0245}$	$0.3884_{\pm 0.0235}$	$0.5910_{\pm 0.0060}$	$0.4820_{\pm 0.0156}$	$0.6356_{\pm 0.0116}$	$0.6314_{\pm 0.0178}$	$0.4552_{\pm 0.0123}$	$0.8205_{\pm 0.0317}$	$0.6430_{\pm 0.0187}$
GD-Sophia	$0.4403_{\pm 0.0037}$	$0.3416_{\pm 0.0488}$	$0.2312_{\pm 0.0120}$	$0.3672_{\pm 0.0063}$	$0.5875_{\pm 0.0034}$	$0.4375_{\pm 0.0103}$	$0.5919_{\pm 0.0296}$	$0.5963_{\pm 0.0165}$	$0.4327_{\pm 0.0114}$	$0.8262_{\pm 0.0205}$	$0.6191_{\pm 0.0191}$
SU-GD-Muon	$0.6050_{\pm 0.0150}$	$0.7800_{\pm 0.0562}$	$0.6216_{\pm 0.0761}$	$0.5389_{\pm 0.0426}$	$0.5774_{\pm 0.0057}$	$0.4797_{\pm 0.0125}$	$0.7490_{\pm 0.0148}$	$0.6280_{\pm 0.0148}$	$0.4770_{\pm 0.0020}$	$0.8452_{\pm 0.0076}$	$0.6593_{\pm 0.0044}$
SU-GD-KLShampoo	$0.5900_{\pm 0.0135}$	$0.7124_{\pm 0.0637}$	$0.5650_{\pm 0.0664}$	$0.5168_{\pm 0.0244}$	$0.5824_{\pm 0.0044}$	$0.4824_{\pm 0.0209}$	$0.7293_{\pm 0.0315}$	$0.6298_{\pm 0.0199}$	$0.4812_{\pm 0.0098}$	$0.8312_{\pm 0.0061}$	$0.6622_{\pm 0.0111}$
SU-GD-SOAP	$0.5357_{\pm 0.0731}$	$0.5388_{\pm 0.2351}$	$0.4790_{\pm 0.1289}$	$0.4502_{\pm 0.0983}$	$0.5909_{\pm 0.0033}$	$0.5000_{\pm 0.0366}$	$0.6759_{\pm 0.0525}$	$0.6511_{\pm 0.0326}$	$0.4826_{\pm 0.0190}$	$0.7999_{\pm 0.0571}$	$0.6641_{\pm 0.0159}$
NPO-AdamW	$0.6730_{\pm 0.0004}$	$0.8858_{\pm 0.0121}$	$0.9101_{\pm 0.0026}$	$0.8582_{\pm 0.0061}$	$0.5410_{\pm 0.0004}$	$0.4778_{\pm 0.0010}$	$0.8580_{\pm 0.0041}$	$0.6228_{\pm 0.0012}$	$0.5167_{\pm 0.0025}$	$0.8721_{\pm 0.0121}$	$0.6911_{\pm 0.0032}$
NPO-Sophia	$0.6195_{\pm 0.0038}$	$0.9230_{\pm 0.0044}$	$0.6814_{\pm 0.0199}$	$0.5184_{\pm 0.0153}$	$0.5248_{\pm 0.0032}$	$0.4556_{\pm 0.0034}$	$0.8977_{\pm 0.0058}$	$0.5898_{\pm 0.0065}$	$0.5375_{\pm 0.0087}$	$0.8937_{\pm 0.0099}$	$0.7181_{\pm 0.0087}$
SU-NPO-Muon	$0.6642_{\pm 0.0003}$	$0.9102_{\pm 0.0158}$	$0.8516_{\pm 0.0038}$	$0.7511_{\pm 0.0078}$	$0.5411_{\pm 0.0005}$	$0.4802_{\pm 0.0011}$	$0.8630_{\pm 0.0047}$	$0.6257_{\pm 0.0020}$	$0.5349_{\pm 0.0036}$	$0.8806_{\pm 0.0081}$	$0.7076_{\pm 0.0036}$
SU-NPO-KLShampoo	$0.6544_{\pm 0.0045}$	$0.9135_{\pm 0.0063}$	$0.8103_{\pm 0.0156}$	$0.6955_{\pm 0.0087}$	$0.5347_{\pm 0.0039}$	$0.4750_{\pm 0.0131}$	$0.8828_{\pm 0.0137}$	$0.6204_{\pm 0.0169}$	$0.5320_{\pm 0.0131}$	$0.8825_{\pm 0.0203}$	$0.7088_{\pm 0.0160}$
SU-NPO-SOAP	$0.6569_{\pm 0.0054}$	$0.9016_{\pm 0.0161}$	$0.8199_{\pm 0.0246}$	$0.7262_{\pm 0.0226}$	$0.5328_{\pm 0.0051}$	$0.4784_{\pm 0.0146}$	$0.8693_{\pm 0.0134}$	$0.6235_{\pm 0.0163}$	$0.5289_{\pm 0.0154}$	$0.8730_{\pm 0.0013}$	$0.7061_{\pm 0.0182}$
RMU-AdamW	$0.6674_{\pm 0.0014}$	$0.8472_{\pm 0.0201}$	$0.9071_{\pm 0.0010}$	$0.8471_{\pm 0.0037}$	$0.5294_{\pm 0.0004}$	$0.4877_{\pm 0.0010}$	$0.8363_{\pm 0.0082}$	$0.6227_{\pm 0.0006}$	$0.4983_{\pm 0.0012}$	$0.8770_{\pm 0.0027}$	$0.6712_{\pm 0.0009}$
RMU-Sophia	$0.6590_{\pm 0.0025}$	$0.7600_{\pm 0.0454}$	$0.8977_{\pm 0.0012}$	$0.8123_{\pm 0.0078}$	$0.5248_{\pm 0.0016}$	$0.4917_{\pm 0.0012}$	$0.8503_{\pm 0.0024}$	$0.6318_{\pm 0.0023}$	$0.4915_{\pm 0.0013}$	$0.9169_{\pm 0.0013}$	$0.6677_{\pm 0.0020}$
SU-RMU-Muon	$0.6605_{\pm 0.0006}$	$0.8151_{\pm 0.0033}$	$0.8881_{\pm 0.0026}$	$0.7990_{\pm 0.0057}$	$0.5246_{\pm 0.0010}$	$0.4903_{\pm 0.0008}$	$0.8343_{\pm 0.0086}$	$0.6286_{\pm 0.0014}$	$0.5000_{\pm 0.0009}$	$0.8799_{\pm 0.0027}$	$0.6707_{\pm 0.0019}$
SU-RMU-KLShampoo	$0.6641_{\pm 0.0002}$	$0.7807_{\pm 0.0115}$	$0.9123_{\pm 0.0022}$	$0.8526_{\pm 0.0020}$	$0.5283_{\pm 0.0011}$	$0.4864_{\pm 0.0006}$	$0.8430_{\pm 0.0047}$	$0.6260_{\pm 0.0010}$	$0.4955_{\pm 0.0014}$	$0.8875_{\pm 0.0040}$	$0.6683_{\pm 0.0020}$
SU-RMU-SOAP	$0.6637_{\pm 0.0009}$	$0.8140_{\pm 0.0371}$	$0.9036_{\pm 0.0079}$	$0.8324_{\pm 0.0143}$	$0.5347_{\pm 0.0021}$	$0.4851_{\pm 0.0001}$	$0.8230_{\pm 0.0170}$	$0.6204_{\pm 0.0032}$	$0.5024_{\pm 0.0033}$	$0.8780_{\pm 0.0105}$	$0.6697_{\pm 0.0034}$
UNDIAL	$0.4386_{\pm 0.0041}$	$0.7696_{\pm 0.0112}$	$0.2586_{\pm 0.0016}$	$0.3266_{\pm 0.0023}$	$0.3087_{\pm 0.0022}$	$0.4163_{\pm 0.0022}$	$0.5908_{\pm 0.0345}$	$0.5041_{\pm 0.0025}$	$0.4441_{\pm 0.0075}$	$0.8075_{\pm 0.0094}$	$0.5695_{\pm 0.0076}$
SatImp	$0.6750_{\pm 0.0009}$	$0.8473_{\pm 0.0071}$	$0.9256_{\pm 0.0014}$	$0.9029_{\pm 0.0036}$	$0.5381_{\pm 0.0003}$	$0.4823_{\pm 0.0006}$	$0.8613_{\pm 0.0047}$	$0.6270_{\pm 0.0014}$	$0.5121_{\pm 0.0009}$	$0.8939_{\pm 0.0071}$	$0.6884_{\pm 0.0009}$
WGA	$0.6729_{\pm 0.0028}$	$0.8546_{\pm 0.0323}$	$0.9382_{\pm 0.0019}$	$0.8958_{\pm 0.0066}$	$0.5465_{\pm 0.0008}$	$0.4755_{\pm 0.0013}$	$0.8313_{\pm 0.0103}$	$0.6178_{\pm 0.0034}$	$0.5010_{\pm 0.0049}$	$0.8732_{\pm 0.0040}$	$0.6665_{\pm 0.0054}$
AltPO	$0.6611_{\pm 0.0006}$	$0.9298_{\pm 0.0102}$	$0.8478_{\pm 0.0012}$	$0.8126_{\pm 0.0021}$	$0.4918_{\pm 0.0003}$	$0.4735_{\pm 0.0008}$	$0.8840_{\pm 0.0067}$	$0.6124_{\pm 0.0007}$	$0.4976_{\pm 0.0005}$	$0.8920_{\pm 0.0040}$	$0.6750_{\pm 0.0004}$
SimNPO	$0.6718_{\pm 0.0009}$	$0.8725_{\pm 0.0030}$	$0.9328_{\pm 0.0014}$	$0.8731_{\pm 0.0095}$	$0.5449_{\pm 0.0006}$	$0.4743_{\pm 0.0005}$	$0.8308_{\pm 0.0057}$	$0.6163_{\pm 0.0022}$	$0.5001_{\pm 0.0035}$	$0.8675_{\pm 0.0070}$	$0.6668_{\pm 0.0045}$
PDU	$0.0984_{\pm 0.0107}$	$0.0328_{\pm 0.0071}$	$0.0230_{\pm 0.0030}$	$0.1226_{\pm 0.0039}$	$0.4985_{\pm 0.0030}$	$0.5347_{\pm 0.0025}$	$0.1608_{\pm 0.0090}$	$0.6713_{\pm 0.0038}$	$0.6194_{\pm 0.0031}$	$0.3481_{\pm 0.0494}$	$0.7974_{\pm 0.0023}$
IdkDPO	$0.6655_{\pm 0.0014}$	$0.9273_{\pm 0.0195}$	$0.8808_{\pm 0.0017}$	$0.8209_{\pm 0.0077}$	$0.5085_{\pm 0.0007}$	$0.4747_{\pm 0.0034}$	$0.8737_{\pm 0.0024}$	$0.6068_{\pm 0.0042}$	$0.5219_{\pm 0.0026}$	$0.8787_{\pm 0.0036}$	$0.6818_{\pm 0.0030}$
GradAscent	$0.6787_{\pm 0.0005}$	$0.9058_{\pm 0.0094}$	$0.9308_{\pm 0.0002}$	$0.9076_{\pm 0.0006}$	$0.5372_{\pm 0.0003}$	$0.4791_{\pm 0.0008}$	$0.8613_{\pm 0.0047}$	$0.6249_{\pm 0.0014}$	$0.5026_{\pm 0.0007}$	$0.8882_{\pm 0.0094}$	$0.6782_{\pm 0.0012}$
CEU	$0.6744_{\pm 0.0007}$	$0.8542_{\pm 0.0141}$	$0.9145_{\pm 0.0002}$	$0.9079_{\pm 0.0025}$	$0.5391_{\pm 0.0002}$	$0.4912_{\pm 0.0003}$	$0.8530_{\pm 0.0000}$	$0.6370_{\pm 0.0007}$	$0.5237_{\pm 0.0002}$	$0.8797_{\pm 0.0059}$	$0.6984_{\pm 0.0008}$

H Limitations and Broader Impacts

Our evaluation extends TOFU and WMDP under the OpenUnlearning framework. Both benchmarks rely on fixed-format prompts or 4-way multiple-choice questions, which may not fully capture the model's residual knowledge under more diverse query distributions. The default forget-side aggregation combines paraphrase probability—the model's likelihood on paraphrased forget prompts—with the forget truth ratio, which compares likelihoods between paraphrased correct answers and perturbed incorrect ones. Both rely on paraphrased forget prompts and are therefore strongly correlated. As discussed in App. D, we replace paraphrase probability with the forget probability evaluated on the original (non-paraphrased) prompts, so that the paraphrase signal enters the aggregate only through the truth ratio. This reduces direct correlations between the aggregated components, but indirect correlations between submetrics remain, and refining the evaluation protocol toward greater metric independence is left to future work. To support such investigations, our detailed sub-metric tables in Tabs. 17–22 report both our protocol's components and the OpenUnlearning defaults, enabling reanalysis under alternative aggregations.

Our work improves machine unlearning for LLMs, which contributes to data privacy compliance (e.g., the right to be forgotten) and safer deployment by enabling more thorough removal of private, copyrighted, or hazardous knowledge. A potential negative impact is that more effective unlearning could be misused to selectively erase desirable knowledge or audit traces from a model; however, this risk is shared with all unlearning research and is not specific to our method.

Spectral updates incur higher time and memory overhead than entry-wise updates, and their behavior at frontier scale remains untested. The choice of ℓ_∞ layer aggregation in Theorem 2 is theoretically motivated, but its empirical advantage over ℓ_1 or ℓ_2 remains open. Finally, since spectral updates also benefit retain-only fine-tuning, separating an unlearning-specific effect from a general optimization gain is left to future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions claimed in the abstract and introduction are substantiated in the main body: the proposed spectral unlearning method is presented in Sec. 3, its theoretical analysis is provided in Sec. 4, and empirical results supporting the claims are reported in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to App. H.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our theoretical results assume that each weight matrix operates on RMS-normalized inputs, which is the standard setup in pre-norm transformer architectures. The full set of assumptions is stated in Sec. 4.1. We note that not all weight matrices in an LLM strictly satisfy the RMS-input assumption. We discuss these practical considerations in App. B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all hyperparameters required to reproduce our experiments in App. C. In addition, we report the complete raw sub-metric results across all models in App. G.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use the public TOFU [5], RWKU [25] and WMDP [6] benchmarks under the OpenUnlearning framework [27]. We provide our code with a detailed `README.md` that specifies the environment setup and the commands required to reproduce all experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Sec. 5 and App. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All results are averaged over three random seeds.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to App. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work is conducted according to the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Please refer to App. H.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our work does not release any new pre-trained models or datasets. We use publicly available benchmarks and publicly available pre-trained models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The models and the datasets are properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: the code is released along with clear documentation and instructions for reproduction.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve any human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We do not use LLMs as a non-standard tool in our method development.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.