CROP: CURVATURE-RESTRICTED ORTHOGONAL PROJECTION FOR RETRAIN-FAITHFUL LLM UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning removes the influence of specific training data from a language model without retraining. Existing methods choose their update direction without reference to the joint forget and retain curvature and are graded by suppression, which rewards overshooting the retrained model. We present CROP (Curvature-Restricted Orthogonal Projection), built on the measured premise that memorization concentrates in parameter directions sharply curved under the forget loss and flat under the retain loss. CROP restricts forget ascent to this curvature-selected subspace, projects the step orthogonal to the retain gradient, anchors on retain descent, and selects its operating point by the distance between its forget profile and the retrained model's, or by a calibrated oracle-free rule. On TOFU, under a leakage-free protocol that selects checkpoints on validation authors and scores held-out ones, CROP leads RMU, the strongest baseline, on held-out forget quality on two of three seeds at a utility cost; as an analysis upper bound, its oracle-best checkpoint lies within KS D 0.045–0.060 of the retrained model on every seed, close to the 0.028–0.050 spread between independent retrains. On MUSE it lands closest to the retrained reference for two 7B families, and on WMDP, which has no retrain oracle, a supplementary suppression test shows it suppresses hazardous knowledge further than RMU while remaining functional. Under full fine-tuning, LoRA, and partial-forget relearning attacks it recovers no faster than the attacked retrain oracle, while RMU, NPO, PDU, and WGA rebound to near pre-unlearning knowledge.

1 INTRODUCTION

Large language models reproduce private records, copyrighted text, and hazardous knowledge from their training corpora on demand (Carlini et al., 2021; 2023). When such content must be removed, the gold standard is retraining on a corpus that excludes it, but sharding-based exact unlearning (Bourtoule et al., 2021) does not apply to monolithically pretrained LLMs and retraining per deletion request is prohibitive, which has produced approximate-unlearning methods that edit the trained model directly, graded on TOFU (Maini et al., 2024), MUSE (Shi et al., 2025), and WMDP (Li et al., 2024).

The existing methods, which we survey in Section 2, share two structural commitments. First, most choose their update direction without reference to the joint forget and retain curvature, whether shaping the forget loss over all trainable parameters (Jang et al., 2023; Liu et al., 2022; Zhang et al., 2024; Fan et al., 2025b; Wang et al., 2025; Yang et al., 2025; Dong et al., 2025; Entesari et al., 2025), applying the same pressure through reference models (Rafailov et al., 2023; Peng et al., 2026), or steering forget activations at chosen layers (Li et al., 2024); the second-order exceptions precondition a full-parameter step rather than selecting where it acts (Section 2). In these methods any direction that reduces the forget objective is used, including directions carrying shared general knowledge, which is why aggressive forgetting so often destroys utility. Second, they are implicitly graded by suppression, which rewards overshoot. A model retrained without the forget data still assigns it plausible likelihood, so driving forget probability far below this reference does not approach the oracle but becomes distinguishable from it in the opposite direction. We argue the right target is retrain-faithfulness, meaning proximity to the retrained model's forget profile, under which the oracle is by definition optimal and overshoot is penalized exactly like undershoot.

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

Our starting point is a geometric measurement rather than a new loss. Across the fine-tuned LLMs we study, memorization of the forget set concentrates in parameter directions along which the forget-loss Hessian is sharply curved while the retain-loss Hessian is flat (retain-to-forget curvature ratios of roughly 0.01 to 0.2, against ~ 1.2 for random directions; Appendix N). We operationalize this in CROP (Curvature-Restricted Orthogonal Projection, Section 3), which estimates the memorization subspace online with finite-difference Hessian-vector products and power iteration, restricts forget ascent to it, projects the step orthogonal to the retain gradient so forgetting cannot raise retain loss to first order, anchors on retain descent, and bounds flattening with a forget-loss gate. When the benchmark provides a retrain oracle, CROP also selects its own operating point by tracking the Kolmogorov–Smirnov (KS) distance between its forget truth-ratio profile and the oracle's; a calibrated oracle-free rule recovers the same operating point on Llama without any retrained model (Appendix D), which we treat as the practical default with oracle selection as the analysis upper bound.

The evidence spans three benchmarks, two model scales, three architecture families, and, on TOFU-1B, three random seeds under a leakage-free protocol that selects checkpoints on validation authors and reports only on held-out ones. Under this protocol CROP leads RMU, the strongest baseline, on held-out forget quality on two of three seeds and trails it on the third, at a utility cost (Table 2). As an analysis upper bound, CROP's oracle-best checkpoint lies inside the spread between independent retrains on two seeds and just above it on the third, whereas RMU's enters that spread on no seed. The ablations of Section 6 show that the measured geometry, not generic regularization, does the work, and that under full fine-tuning, LoRA, and partial-forget relearning attacks CROP recovers knowledge no faster than the attacked retrain oracle while RMU and NPO rebound to near pre-unlearning levels.

Our contributions are threefold: (i) we measure curvature separability of memorization, showing that sharp-for-forget, flat-for-retain directions exist across the fine-tuned LLMs we study and can be found online with gradient evaluations alone; (ii) we introduce CROP, which restricts forget ascent to this measured subspace, projects it orthogonal to the retain gradient, a first-order retain-safety property of the pseudo-gradient that we state as design motivation and measure under AdamW, and stabilizes it with a retain anchor and a bounded gate; and (iii) we show that operating-point selection is a necessary part of curvature-based unlearning, since trajectories revert with no training-side warning, and propose profile-tracked selection under a leakage-free protocol with an oracle-free variant, evaluated on TOFU and MUSE, with WMDP as a supplementary suppression test, across three model families with ablations, size scaling, and a relearning-attack suite controlled by the attacked retrain oracle.

2 Related Work

Three lines of work are most relevant to CROP. Exact unlearning retrains without the deleted data and remains prohibitive at LLM scale even with sharding (Bourtoule et al., 2021); approximate methods edit the trained model directly and include gradient ascent and difference (Jang et al., 2023; Liu et al., 2022), preference-based objectives (Rafailov et al., 2023; Maini et al., 2024; Zhang et al., 2024), reweighted, distilled, or constrained variants (Fan et al., 2025b; Wang et al., 2025; Yang et al., 2025; Dong et al., 2025; Entesari et al., 2025; Peng et al., 2026), and mid-layer activation steering (Li et al., 2024). A second line brings second-order structure to data removal (Guo et al., 2020; Golatkar et al., 2020; Koh & Liang, 2017; McKinney et al., 2025), localizes knowledge inside networks (Meng et al., 2022; 2023; Maini et al., 2023), and ties loss curvature to memorization (Keskar et al., 2017; Garg et al., 2024; Foret et al., 2021); CROP takes its coordinate system from this line. The nearest second-order method, K-FADE (McKinney et al., 2025), takes uphill Gauss–Newton steps on the forget set under a Kronecker-factored curvature approximation, so curvature preconditions an update spanning every parameter; CROP instead uses the joint forget and retain curvature to select a low-dimensional subspace, estimated online without a parametric Hessian approximation, and confines ascent to it.[1] Two further neighbors share one ingredient each. Sharpness-aware unlearning (Fan et al., 2025a) flattens the forget loss over all parameters so that a relearning attack cannot re-enter a sharp basin; CROP uses sharpness as a coordinate system instead, ascending only along sharp-for-forget, flat-for-retain directions. Projecting an update orthogonal to a protected gradient descends from continual and multi-task learning (Farajtabar et al., 2020; Yu et al., 2020); CROP

[1] K-FADE is not in our experimental comparison because no official implementation was available to us (Section 4.1).

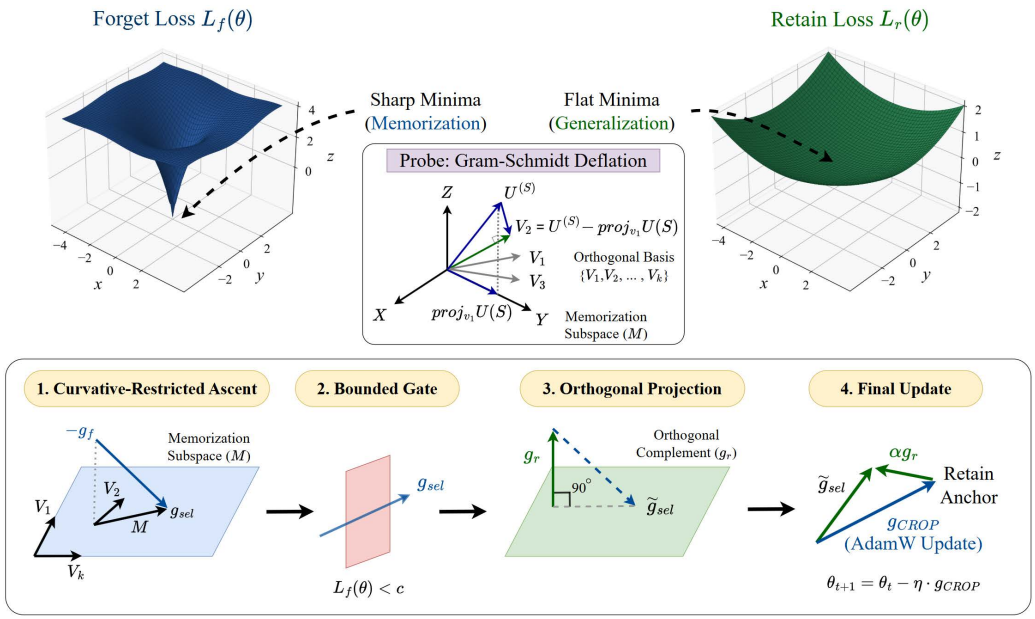


Figure 1: Overview of CROP. A curvature probe extracts top forget-Hessian eigendirections and filters them by retain flatness; ascent is restricted to this subspace, projected orthogonal to the retain gradient, gated, and anchored on retain descent; the delivered checkpoint is the one closest to the retrain oracle's forget profile.

applies it inside the curvature-selected subspace rather than to the full gradient, with the exact projector of Appendix A keeping the composition inside that subspace. Liang & Li (2026) ask at what training phase a model is best unlearned; CROP is complementary, asking which directions of parameter space to use at whatever phase the model is in. Finally, the benchmarks (Maini et al., 2024; Shi et al., 2025; Li et al., 2024), membership-inference audits (Shokri et al., 2017; Shi et al., 2024), and studies showing shallow unlearning is easily reverted (Hu et al., 2025; Zhang et al., 2025; Lucki et al., 2025) together favor retrain-faithfulness over maximal suppression, which CROP adopts as its training objective.

Each baseline family leaves a gap that CROP is built to close. Gradient ascent and difference follow the full forget gradient through directions shared with retained knowledge, unbounded, so they either barely move at a safe budget or degrade into unusable text. Preference and reweighting objectives (NPO, SimNPO, WGA, SatImp) bound the ascent but keep the full gradient as the direction and are tuned to suppression, so they overshoot the retrained model or pay with utility, and DPO's refusal target is an answer no retrained model produces. Constrained formulations (PDU, CA-NPO) regulate a scalar forget loss but not where the update acts, and representation steering (RMU) suppresses an activation rather than the memorization, which returns under a short relearning attack. All stop by budget or early stopping with no signal for whether the returned checkpoint is the closest to the retrained model; Section 3 addresses these gaps in the same order.

3 Methods

CROP (Curvature-Restricted Orthogonal Projection) rests on one measured premise: memorization of the forget set concentrates in parameter directions along which the forget loss is sharply curved while the retain loss is flat. It restricts forget ascent to this curvature-selected subspace, projects the step orthogonal to the retain gradient, stabilizes it with a retain anchor and a bounded gate, and, when a retrain oracle is available, selects its own operating point by tracking its distributional distance to the oracle (Figure 1).

3.1 Problem Setup and Notation

Let $f_{\boldsymbol{\theta}}$ denote an autoregressive language model with parameters $\boldsymbol{\theta} \in \mathbb{R}^d$, obtained by training on a corpus $\mathcal{D} = \mathcal{D}_f \cup \mathcal{D}_r$, where $\mathcal{D}_f$ is the forget set whose influence must be removed and $\mathcal{D}_r$ is the retain set whose knowledge must be preserved. For a dataset $\mathcal{S}$ of prompt–answer pairs we write the

token-averaged negative log-likelihood loss as

$$\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) = -\frac{1}{N_{\mathcal{S}}} \sum_{(x,y)\in\mathcal{S}} \sum_{u=1}^{|y|} \log p_{\boldsymbol{\theta}}(y_u \mid x, y_{<u}), \qquad N_{\mathcal{S}} = \sum_{(x,y)\in\mathcal{S}} |y|, \tag{1}$$

where the average runs over all answer tokens of $\mathcal{S}$, and abbreviate $\mathcal{L}_f \coloneqq \mathcal{L}_{\mathcal{D}_f}$ and $\mathcal{L}_r \coloneqq \mathcal{L}_{\mathcal{D}_r}$. We write the corresponding gradients as $\boldsymbol{g}_f \coloneqq \nabla_{\boldsymbol{\theta}}\mathcal{L}_f(\boldsymbol{\theta})$ and $\boldsymbol{g}_r \coloneqq \nabla_{\boldsymbol{\theta}}\mathcal{L}_r(\boldsymbol{\theta})$, and the Hessians as $\boldsymbol{H}_f \coloneqq \nabla^2_{\boldsymbol{\theta}}\mathcal{L}_f(\boldsymbol{\theta})$ and $\boldsymbol{H}_r \coloneqq \nabla^2_{\boldsymbol{\theta}}\mathcal{L}_r(\boldsymbol{\theta})$. Throughout, $\langle\cdot,\cdot\rangle$ is the Euclidean inner product on $\mathbb{R}^d$, $\|\cdot\|$ the induced norm, $\mathbb{I}[\cdot]$ the indicator function, and $\boldsymbol{I}_d$ the $d \times d$ identity. In practice all gradient and curvature quantities are stochastic estimates on the current minibatches; we suppress the minibatch index.

The retrain oracle $\boldsymbol{\theta}^*$ is a model trained from scratch on $\mathcal{D}_r$ alone; its cost motivates approximate unlearning, whose goal we take to be retrain-faithfulness: the unlearned model should match $\boldsymbol{\theta}^*$'s behavioral profile on the forget set rather than be maximally suppressive, since a model that has never seen $\mathcal{D}_f$ still assigns it plausibly-shaped likelihood. We quantify faithfulness by distributional distance to the oracle's profile, treating a non-rejected KS test as evidence of proximity, not of equivalence. We scope the definition to the forget profile deliberately: it is the one quantity for which the retrained model is optimal by construction, whereas utility has no target beyond the retrained model's own value, so we report it as a separate axis and count any shortfall as a cost, not as faithfulness.

3.2 Curvature Separability of Memorization

CROP starts from a geometric measurement. For a model fine-tuned on data containing $\mathcal{D}_f$, consider the top eigendirections of the forget-loss Hessian $\boldsymbol{H}_f$, which are the directions along which the optimizer carved the sharpest, most example-specific minima. For a unit vector $\boldsymbol{v} \in \mathbb{R}^d$, define the retain-to-forget curvature ratio

$$\rho(\boldsymbol{v}) \;\coloneqq\; \frac{\left|\boldsymbol{v}^\top \boldsymbol{H}_r\, \boldsymbol{v}\right|}{\boldsymbol{v}^\top \boldsymbol{H}_f\, \boldsymbol{v}}, \qquad \boldsymbol{v}^\top \boldsymbol{H}_f\, \boldsymbol{v} > 0. \tag{2}$$

Empirically, the top eigendirections of $\boldsymbol{H}_f$ exhibit $\rho(\boldsymbol{v}) \approx 0.01$–$0.2$ across the benchmarks we study, and the same separation holds at the 7B scale (median $\rho = 0.04$ within the trainable slab), where their retain curvature is near zero or negative. Memorization is therefore empirically curvature-separable: a low-dimensional subspace exists that is sharp for the forget loss yet flat for the retain loss. Its directions rotate across probe minibatches and deform under the unlearning updates while the sharp/flat property reappears on every draw, so the probe is refreshed online rather than computed once. Figure 3 shows the signature direction by direction; Appendix N verifies it against random controls and validates the estimator.

3.3 Curvature Probe for Estimating the Memorization Subspace

Materializing $\boldsymbol{H}_f \in \mathbb{R}^{d\times d}$ is infeasible, but Hessian–vector products (HVPs) are cheap. CROP estimates them by a forward finite-difference scheme requiring only gradient evaluations,

$$\boldsymbol{H}_{\mathcal{S}}\, \boldsymbol{v} \;\approx\; \widehat{\boldsymbol{H}}_{\mathcal{S}}\, \boldsymbol{v} \;\coloneqq\; \frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta} + \epsilon \boldsymbol{v}) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})}{\epsilon}, \qquad \epsilon = 10^{-3}, \tag{3}$$

which avoids double backpropagation and runs unmodified under memory-efficient attention kernels.

Every T_{p} optimizer steps, CROP extracts the top k eigendirections of $\widehat{\boldsymbol{H}}_f$ by power iteration with Gram–Schmidt deflation. For direction index $i = 1, \ldots, k$, starting from $\boldsymbol{u}^{(0)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ normalized to unit length, we apply, for $s = 1, \ldots, P$, the update

$$\tilde{\boldsymbol{u}}^{(s)} \;=\; \widehat{\boldsymbol{H}}_f \left(\boldsymbol{u}^{(s-1)} - \sum_{j<i} \left\langle \boldsymbol{u}^{(s-1)}, \boldsymbol{v}_j \right\rangle \boldsymbol{v}_j \right), \qquad \boldsymbol{u}^{(s)} \;=\; \frac{\tilde{\boldsymbol{u}}^{(s)}}{\left\|\tilde{\boldsymbol{u}}^{(s)}\right\|}, \tag{4}$$

where $\{\boldsymbol{v}_j\}_{j<i}$ are all previously extracted directions, whether or not they pass the filter below; a final deflation and normalization of $\boldsymbol{u}^{(P)}$ yields $\boldsymbol{v}_i$. Each candidate is then scored by its Rayleigh quotients under both Hessians,

$$\lambda_i^f \;\coloneqq\; \boldsymbol{v}_i^\top \widehat{\boldsymbol{H}}_f\, \boldsymbol{v}_i, \qquad \lambda_i^r \;\coloneqq\; \boldsymbol{v}_i^\top \widehat{\boldsymbol{H}}_r\, \boldsymbol{v}_i, \tag{5}$$

and admitted to the memorization subspace only if it is sharp for the forget loss and sufficiently flat for the retain loss, yielding

$$\mathcal{V} \;=\; \Big\{\, \boldsymbol{v}_i \;:\; \lambda_i^f > 0 \;\text{ and }\; \big|\lambda_i^r\big| < \rho_{\max}\, \lambda_i^f \,\Big\}, \qquad \mathcal{M} \;=\; \operatorname{span}(\mathcal{V}), \tag{6}$$

with $\rho_{\max}$ the retain-flatness threshold (default 0.25). The filter bounds only the diagonal of $\boldsymbol{U}^\top \widehat{\boldsymbol{H}}_r \boldsymbol{U}$ for the kept basis $\boldsymbol{U}$, not the retain curvature of linear combinations; Appendix A measures the full matrix and the curvature ratio along the realized step, which is flat at the median (0.08) but exceeds $\rho_{\max}$ on about a third of ascent steps, so the flatness property holds per direction, not per step. Deep-network Hessians are indefinite, so the $\lambda_i^f > 0$ condition in equation 6 discards negative-curvature directions (Lanczos check in Appendix N). Since $\mathcal{V} \subseteq \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ with $|\mathcal{V}| = k' \le k$ is a subset of an orthonormal set, the orthogonal projector onto $\mathcal{M}$ is

$$\boldsymbol{\Pi}_{\mathcal{M}} \;=\; \sum_{\boldsymbol{v} \in \mathcal{V}} \boldsymbol{v}\boldsymbol{v}^\top. \tag{7}$$

At 7B scale, where k full-dimensional direction vectors do not fit in memory, the whole mechanism runs on a contiguous slab of layers $\boldsymbol{\theta}_{\mathcal{B}} \subset \boldsymbol{\theta}$ with all other parameters frozen (Appendix B, Tables 6 and 7).

3.4 The CROP Update

At every optimizer step, given the current forget and retain gradients $\boldsymbol{g}_f$ and $\boldsymbol{g}_r$ from equation 1, CROP constructs its update in four stages. Every quantity below is a pseudo-gradient handed to a descent optimizer, so the realized displacement points along its negative (exactly for plain gradient descent, up to diagonal preconditioning for AdamW, quantified below).

(i) Curvature-restricted ascent. The forget pseudo-gradient is the negated forget gradient projected onto the memorization subspace,

$$\boldsymbol{g}_{\text{sel}} \;=\; \boldsymbol{\Pi}_{\mathcal{M}}\,(-\boldsymbol{g}_f) \;=\; \sum_{\boldsymbol{v} \in \mathcal{V}} \langle -\boldsymbol{g}_f,\, \boldsymbol{v}\rangle\, \boldsymbol{v} \;=\; -\,\boldsymbol{\Pi}_{\mathcal{M}}\, \boldsymbol{g}_f. \tag{8}$$

(ii) Bounded-flattening gate. Unbounded ascent drives the model into degenerate outputs, the failure mode of plain gradient ascent, so CROP gates ascent on the current forget loss,

$$\boldsymbol{g}_{\text{sel}} \;\leftarrow\; \mathbb{I}[\mathcal{L}_f(\boldsymbol{\theta}) < c] \cdot \boldsymbol{g}_{\text{sel}}, \tag{9}$$

where c is the forget cap. Once the forget negative log-likelihood exceeds c ascent switches off, re-engaging if retain co-training re-lowers $\mathcal{L}_f$; the gate regulates the mean forget loss, not the shape of the forget profile, which is why delivery relies on the profile-tracked selection of Section 3.5.

(iii) Retain-orthogonal projection. The gated ascent step is projected onto the orthogonal complement of the retain gradient,

$$\tilde{\boldsymbol{g}}_{\text{sel}} \;=\; \left(\boldsymbol{I}_d - \frac{\boldsymbol{g}_r \boldsymbol{g}_r^\top}{\|\boldsymbol{g}_r\|^2}\right) \boldsymbol{g}_{\text{sel}} \;=\; \boldsymbol{g}_{\text{sel}} - \frac{\langle \boldsymbol{g}_{\text{sel}},\, \boldsymbol{g}_r\rangle}{\|\boldsymbol{g}_r\|^2}\, \boldsymbol{g}_r. \tag{10}$$

Because $\boldsymbol{\Pi}_{\mathcal{M}}$ and this projector do not commute, $\tilde{\boldsymbol{g}}_{\text{sel}}$ carries a component outside $\mathcal{M}$ of relative size at most $|\cos(\boldsymbol{g}_{\text{sel}}, \boldsymbol{g}_r)|$ (median 0.03 on TOFU-1B), and it ascends $\mathcal{L}_f$ to first order on every instrumented step. Proposition 2 (Appendix A) gives the exact projector onto $\mathcal{M} \cap \boldsymbol{g}_r^{\perp}$, which removes the leak and reaches oracle-best checkpoints more faithful than the composed form's on two seeds and less on one, within the run-to-run variability of a single seed (Table 4). Every reported experiment, ablation, and stress test uses the composed form, which we keep as the reported default rather than rerun the suite; the exact projector is provided as an option.

Proposition 1 (First-order retain safety, pre-optimizer)**.** *Let $\tilde{\boldsymbol{g}}_{\text{sel}}$ be given by equation 10 with $\boldsymbol{g}_r \neq \boldsymbol{0}$. Then $\langle \tilde{\boldsymbol{g}}_{\text{sel}}, \boldsymbol{g}_r\rangle = 0$, and for the displacement $\boldsymbol{\theta}' = \boldsymbol{\theta} - \eta\, \tilde{\boldsymbol{g}}_{\text{sel}}$ realized by a plain gradient step on the pseudo-gradient $\tilde{\boldsymbol{g}}_{\text{sel}}$ with learning rate $\eta > 0$,*

$$\mathcal{L}_r(\boldsymbol{\theta}') \;=\; \mathcal{L}_r(\boldsymbol{\theta}) \;-\; \eta\, \langle \boldsymbol{g}_r,\, \tilde{\boldsymbol{g}}_{\text{sel}}\rangle + O(\eta^2) \;=\; \mathcal{L}_r(\boldsymbol{\theta}) + O(\eta^2). \tag{11}$$

That is, the forget component of the update cannot change the retain loss to first order.

The proof, immediate from the projector, is in Appendix A. We state the proposition as the design motivation for step (iii) rather than as a guarantee on the realized update, since AdamW's preconditioning and the anchor of step (iv) both act after it; what survives them is measured below.

(iv) Retain anchor. Finally, an explicit retain-descent term is added, and the combined direction is handed to the optimizer (AdamW (Loshchilov & Hutter, 2019) in all experiments),

$$\boldsymbol{g}_{\mathrm{CROP}} \;=\; \tilde{\boldsymbol{g}}_{\mathrm{sel}} + \alpha\, \boldsymbol{g}_r, \qquad \boldsymbol{\theta}_{t+1} \;=\; \mathrm{AdamW}(\boldsymbol{\theta}_t,\, \boldsymbol{g}_{\mathrm{CROP}})\,, \tag{12}$$

where $\alpha > 0$ is the anchor strength and $\mathrm{AdamW}(\boldsymbol{\theta}_t, \boldsymbol{g})$ denotes one AdamW step from $\boldsymbol{\theta}_t$ on the pseudo-gradient $\boldsymbol{g}$. The anchor repairs the retain-loss drift that AdamW's preconditioning reintroduces: per-step instrumentation (Table 5, Appendix A) shows that without the anchor the realized update stays nearly orthogonal to $\boldsymbol{g}_r$ yet leaves a small first-order drift, while with it the realized step is deliberately not orthogonal to $\boldsymbol{g}_r$ and the retain loss descends. By Proposition 1 the anchor and forget components never oppose each other to first order; the ablations of Section 6 show that curvature selection and the anchor are essential and that the projection and gate materially improve faithfulness.

3.5 Operating-Point Selection via Oracle-Profile Tracking

Unlearning trajectories need not descend monotonically toward the oracle: the model can overshoot or partially re-memorize under retain co-training before approaching again, so neither a fixed budget nor an early stop is guaranteed to end at the best model visited (Figure 4, Appendix C). On benchmarks with a retrain oracle, CROP therefore treats operating-point selection as part of the method. Every T_{d} steps it computes the benchmark's truth ratio for each forget example and the two-sample Kolmogorov–Smirnov statistic $D(\boldsymbol{\theta})$ between this profile and the oracle's (equation 13, equation 14); both samples have fixed, equal size, so minimizing D online is equivalent to maximizing the reported forget-quality p-value. CROP snapshots the model at every improvement of D and delivers the global-minimum-D snapshot, the best model visited rather than the last. Because selecting and evaluating on the same forget examples is adaptive model selection, the reported protocol chooses the checkpoint on validation authors and scores it on held-out test authors, over five random splits and three random seeds (Section 5.1, Appendix C); oracle-best selection is reported as an upper bound, and Appendix D gives an oracle-free variant, evaluated on three random seeds of each family (Tables 9 and 10), whose selected checkpoint is a row of Table 1.

4 Experimental Setup

4.1 Benchmarks, Models, and Baselines

We evaluate CROP on three benchmarks spanning five target models. On TOFU (Maini et al., 2024) (forget10) we unlearn Llama-3.2-1B-Instruct (Grattafiori et al., 2024) and Qwen2.5-1.5B-Instruct (Yang et al., 2024), where a retain-gold model retrained without the forget set provides an exact oracle reference; on MUSE-News (Shi et al., 2025), Llama-2-7B (Touvron et al., 2023) and Qwen2.5-7B (Yang et al., 2024), again with a retrained oracle; and on WMDP (Li et al., 2024) we unlearn hazardous biosecurity and cybersecurity knowledge from Zephyr-7B-β (Tunstall et al., 2024). On TOFU-1B the reported protocol is leakage-free, consisting of full-length dense-checkpoint runs under three random seeds (0, 1, and 2) in which the checkpoint is chosen on ten validation authors and scored on the ten held-out test authors over five random splits (Appendix C); RMU is re-run under identical instrumentation, seeds, and selection access. On TOFU-1.5B CROP prepends a short unrestricted-ascent warm-up for the saturated target and runs a 48-epoch budget (Appendix F); on MUSE and WMDP, which provide no truth-ratio profile, every method runs a fixed budget with dense checkpointing and reports the checkpoint closest to the retrain reference under the benchmark's own metrics, so those rows are best-attainable operating points rather than leakage-free estimates. At 7B CROP trains only a contiguous layer slab (Appendix B), whereas the baselines run in their published form and, for most of them, update every parameter. We compare against nine baselines covering gradient-based objectives (GradAscent (Jang et al., 2023), NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025b), WGA (Wang et al., 2025), SatImp (Yang et al., 2025), PDU (Entesari et al., 2025), CA-NPO (Peng et al., 2026)), preference-based unlearning (DPO with IdK targets (Rafailov et al., 2023; Maini et al., 2024)), and representation steering (RMU (Li et al., 2024)), all trained and evaluated in the OpenUnlearning framework (Dorna et al., 2025) with the benchmarks' official metrics. We also add NPO+SAM (Fan et al., 2025a), sharpness-aware NPO built for relearning resilience, on every benchmark and in the relearning suite (NPO's recipe, $\rho = 0.01$). K-FADE (McKinney et al., 2025) is omitted for lack of an official implementation.

4.2 Evaluation Metrics

On TOFU, forget quality (Forget KS-p) is the p-value of a two-sample Kolmogorov–Smirnov test comparing the model's forget truth-ratio profile (equation 13) against the retain-gold model's; Model

Table 1: Results on TOFU-1B (Llama-3.2-1B-Instruct, forget10). Best in bold, second underlined. Prob, ROUGE, TR, and ES are judged by closeness to the retain-gold oracle (O). The CROP row is the checkpoint selected against the retain-gold profile on these same examples, an analysis upper bound (Table 2 gives the leakage-free estimate); CROP-OF is the same recipe's checkpoint chosen by the oracle-free band rule of Appendix D, without consulting the retrained model. KS D is the effect size against the retain-gold profile with a 95% example-bootstrap interval; † marks equality not rejected at $\alpha = 0.05$ ($D \leq 0.0925$ at n=400). Four independent retrains of the oracle differ pairwise by $D = 0.028$–0.050 (Appendix C).

Method	KS-p ↑	KS D [95% CI] ↓	MU ↑	Prob → O	ROUGE → O	TR → O	ES → O	PrivLeak → 0
Target	2×10^{-22}	0.352 [0.30, 0.42]	0.6003	0.8807	0.8194	0.4752	0.7005	−99.46
Retain-gold (O)	1.0000	0.000 (ref.)	0.5911	0.1161	0.3791	0.6275	0.0590	23.48
GradAscent (Jang et al., 2023)	1×10^{-20}	0.338 [0.28, 0.41]	0.5852	0.7882	0.6959	0.5162	0.5009	−99.19
NPO (Zhang et al., 2024)	6×10^{-4}	0.142 [0.12, 0.20]	0.2100	0.0654	0.1261	0.6925	<u>0.0575</u>	<u>−11.11</u>
DPO (IdK) (Maini et al., 2024)	6×10^{-14}	0.278 [0.23, 0.35]	0.5611	0.4994	0.0933	0.5264	0.2277	−95.38
RMU (Li et al., 2024)	$0.1111^\dagger$	0.085 [0.07, 0.15]	0.5288	0.0148	0.1641	0.6772	0.0356	55.39
WGA (Wang et al., 2025)	7×10^{-10}	0.233 [0.19, 0.30]	0.2220	0.0604	0.2535	0.6749	**0.0600**	−78.28
NPO+SAM (Fan et al., 2025a)	4×10^{-4}	0.145 [0.11, 0.22]	0.2888	0.2102	0.2207	0.6763	0.0822	−59.18
SimNPO (Fan et al., 2025b)	4×10^{-17}	0.307 [0.27, 0.38]	0.5827	0.3809	0.4642	0.4710	0.1801	−93.03
SatImp (Yang et al., 2025)	2×10^{-21}	0.345 [0.30, 0.41]	**0.5959**	0.8531	0.7683	0.4695	0.6127	−99.42
PDU (Entesari et al., 2025)	1×10^{-20}	0.338 [0.29, 0.40]	<u>0.5900</u>	0.0067	0.0684	0.8219	0.0390	57.10
CA-NPO (Peng et al., 2026)	1×10^{-3}	0.135 [0.08, 0.21]	0.5582	<u>0.1384</u>	**0.3053**	0.5778	0.0772	−36.66
CROP (Ours)	<u>$0.4158^\dagger$</u>	<u>0.063</u> [0.06, 0.14]	0.4585	0.1583	<u>0.2955</u>	<u>0.6669</u>	0.0630	**−10.57**
CROP-OF (Ours)	**$0.5812^\dagger$**	**0.055** [0.05, 0.13]	0.5101	**0.1159**	0.2812	**0.6280**	0.0551	13.18

Utility (MU) aggregates retain-side likelihood, ROUGE (Lin, 2004), and truth-ratio scores, and Forget Prob, Forget ROUGE, Extraction Strength (ES), and PrivLeak (min-$k\%$ membership inference (Shi et al., 2024)) are all judged by proximity to the oracle's values rather than by minimization. MUSE (VerbMem, KnowMem-f/-r, PrivLeak) and WMDP (multiple-choice accuracy on Bio, Cyber, and held-out Chem, with chance 0.25 as the ideal; MMLU (Hendrycks et al., 2021) and MT-Bench (Zheng et al., 2023) for utility) use the benchmarks' own metrics, defined in Appendix E; configurations, baseline parity, and per-step cost (1.8–$2.5\times$ a GradDiff step, set by the probe refresh interval) are in Appendix F (Tables 11, 12, 13; Figure 5). Because a large KS p-value only means the equality hypothesis is not rejected at the given sample size, never that the distributions are equal, we accompany p-values with the KS statistic D as the effect size, example-bootstrap confidence intervals, and a secondary Wasserstein distance between log truth-ratio profiles (Tables 1 and 2). All tables mark the method closest to the retrain reference in bold and the second closest underlined.

5 Results

5.1 Retrain-Faithful Forgetting at Both TOFU Scales

Tables 1 and 2 show the central result from two angles. Table 1 places CROP's oracle-selected operating point in the full field (KS-p 0.4158 at MU 0.4585; strongest baseline RMU at 0.1111), and the oracle-free row reaches KS-p 0.5812 at MU 0.5101 without the retrained model. Because that point is chosen against the oracle on the evaluation examples, Table 2 re-derives the comparison under the leakage-free protocol, three random seeds, and identical instrumentation, and it is the estimate we stand on. First, on held-out authors CROP leads on forget quality on seeds 1 and 2 but trails RMU on seed 0, for a mean KS-p of 0.37 ± 0.21 vs. 0.20 ± 0.07, a gap the three-seed variance does not make significant, at an exposed utility cost (MU 0.415 vs. 0.523). At matched utility, which exists only on seed 0, CROP is more faithful than RMU's best checkpoint (Table 8). Second, as the analysis upper bound, CROP's oracle-best checkpoint lies within KS D 0.045–0.060 of the retrained model on every seed, close to the 0.028–0.050 spread between independent retrains, whereas RMU's oracle-best enters that spread on no seed. On TOFU-1.5B (Table 14, Appendix G) CROP is the only entry crossing the threshold (KS-p 0.111 against 1×10^{-3} for the best baseline).

The per-metric pattern explains why: the retain-gold model still assigns forget answers non-trivial probability, CROP lands near its values where aggressive-looking baselines overshoot, and CROP's PrivLeak has the smallest magnitude on both tiers while most baselines sit near the -100 extreme (Appendix G).

5.2 Uniform Movement Toward the Retrained Model on MUSE

Tables 16 and 15 (Appendix H) report both 7B families in full. On Llama-2-7B only WGA and CROP move meaningfully toward the reference; WGA overshoots VerbMem while CROP lands closest on VerbMem (0.1769) and extraction strength (0.0285), and both pay in retain knowledge

Table 2: Leakage-free selection on TOFU-1B, with checkpoints chosen on 10 validation authors, scored on 10 held-out authors (mean±std over 5 splits, 3 random seeds: 0, 1, 2). Peak KS-p is the oracle-selected upper bound; KS D [95% CI] and W_1, the Wasserstein distance between log truth-ratio profiles, are at each seed's modal selected checkpoint (Appendix C).

Method	Seed	Held-out KS-p	Held-out KS D	MU	Peak KS-p	KS D [95% CI]	W_1
CROP	0	0.091 ± 0.093	0.132 ± 0.019	0.446	0.813	0.095 [0.065, 0.153]	0.164
CROP	1	0.419 ± 0.267	0.092 ± 0.022	0.410	0.700	0.050 [0.048, 0.128]	0.200
CROP	2	0.597 ± 0.133	0.077 ± 0.008	0.389	0.468	0.060 [0.050, 0.130]	0.090
CROP	mean	0.369 ± 0.210	0.100 ± 0.023	0.415 ± 0.023	0.660 ± 0.143	–	0.151
RMU	0	0.291 ± 0.201	0.103 ± 0.019	0.523	0.523	0.065 [0.055, 0.138]	0.083
RMU	1	0.136 ± 0.167	0.126 ± 0.022	0.518	0.065	0.108 [0.070, 0.173]	0.158
RMU	2	0.166 ± 0.129	0.117 ± 0.018	0.528	0.065	0.098 [0.068, 0.173]	0.166
RMU	mean	0.198 ± 0.067	0.115 ± 0.009	0.523 ± 0.004	0.218 ± 0.216	–	0.136

(KnowMem-r 0.38 against the retrained model's 0.56). On Qwen2.5-7B, CROP is closest on both forget-side axes simultaneously (VerbMem 0.1951 vs. reference 0.1918; KnowMem-f 0.2084 vs. 0.1881). The limitation is Qwen PrivLeak, near the -95 floor (Appendix H).

5.3 WMDP as a Supplementary Suppression Test

WMDP has no retrain oracle, so retrain-faithfulness cannot be assessed there and this section is a supplementary suppression test, not a generality check. The operational goal becomes driving hazardous-knowledge accuracy toward chance (0.25) while keeping the model functional, a dual requirement that eliminates most of the field: GradAscent and PDU reach chance accuracy but collapse to MT-Bench 1.0, while WGA, SimNPO, SatImp, and NPO+SAM keep utility but leave WMDP-Bio above 0.58 (Table 18, Appendix I). CROP reaches the deepest suppression among methods that remain functional (WMDP-Bio 0.31, WMDP-Cyber 0.25) while sustaining MT-Bench 4.28 and MMLU 0.45, ten points below RMU; RMU suppresses less and keeps more, so the two bracket the functional frontier. The held-out chemistry column exposes a limitation in target specificity: CROP reduces WMDP-Chem to 0.3725, the largest reduction among functional methods, with no chemistry corpus in the unlearning data, and the same effect spills into adjacent benign categories (Appendix I, Table 17).

Trade-offs and what the numbers mean. Read jointly (Appendix J), the tables place forgetting depth and utility on a hard Pareto surface whose reachable end CROP extends (Figure 2); deeper is not better, since suppression metrics reward overshoot the KS test detects, and CROP is best or second best on the primary forget-side axis of every benchmark and family.

6 Ablation Study

6.1 Component-wise Analysis

Table 3 (Appendix G) removes CROP's five components on TOFU-1B one at a time (see also Appendices L and M); the two curvature rows carry the central claim. Removing the curvature restriction entirely, which recovers the GradDiff objective (Liu et al., 2022), collapses the forget KS-p from 0.4158 to 1×10^{-17}; a random subspace of the same dimension, with every other component active, still collapses it to 9×10^{-6}, consistent with the direct measurements of Figure 3 and Appendix N. Restricting ascent to a low-dimensional subspace is therefore not what makes CROP work; ascending in the curvature-selected subspace is.

The remaining rows separate by failure axis: without the retain anchor the model is destroyed (utility 0.0000); without the orthogonal projection utility survives but forget quality drops five-fold; the flatness filter and the gate are balance components, since removing either still passes the threshold at a cost, and every knockout that leaves a functional model worsens privacy leakage (-10.6 for the full method against -29 to -79).

6.2 Stress Tests on Forget-Set-Size Scaling and Relearning

Figures 7 and 8 (Appendix K) vary the forget-set size across five splits (1%–10%), transferring every method's 10%-split hyperparameters unchanged (Zhang et al., 2024; Fan et al., 2025b; Wang et al., 2025). CROP crosses the $\alpha = 0.05$ threshold at all five sizes (KS-p 0.27–0.70) at utility 0.40–0.46, whereas RMU passes only at 10% and WGA at none, because CROP's operative quantities are measured online while the baselines' fixed coefficients encode the tuning scale.

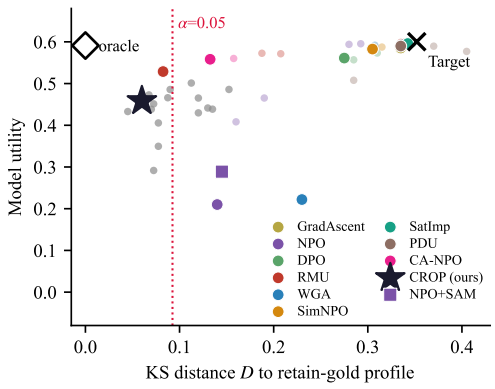


Figure 2: Forgetting-versus-utility frontier on TOFU-1B over every evaluated operating point (sweep arms faint, reported rows solid). Left of the dotted line the KS equality test is not rejected; only one RMU point reaches that region.

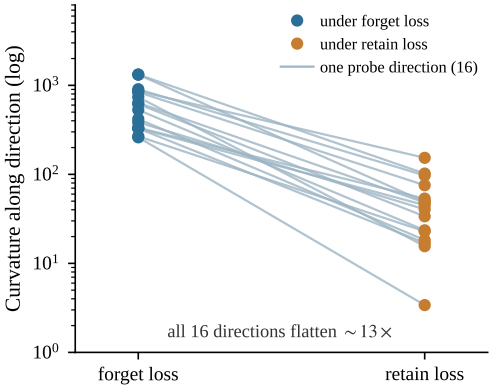


Figure 3: The separability signature. Each line is one probe direction, joining its forget-loss curvature to its retain-loss curvature (log axis). All 16 fall by about an order of magnitude (median $\rho = 0.077$).

Table 3: Component-wise ablation of CROP on TOFU-1B (forget10); one component removed per variant, all else fixed. Rel. KS-p is the ratio to the full method; "w/o curvature restriction" recovers GradDiff (Liu et al., 2022). † KS equality test against retain-gold not rejected at $\alpha = 0.05$. KS-p is a deterministic function of D at n=400, so identical values recur across tables.

Variant	Forget KS-p ↑	Rel. KS-p	MU ↑	ΔMU	PrivLeak
CROP (Full)	0.4158†	1.00×	0.4585	—	−10.57
w/o curvature selection	9×10^{-6}	2×10^{-5}×	0.4240	−0.035	−79.14
w/o curvature restriction	1×10^{-17}	3×10^{-17}×	0.4445	−0.014	−50.20
w/o retain-flatness filter	0.3671†	0.88×	0.4174	−0.041	−29.42
w/o bounded-flattening gate	0.2107†	0.51×	0.4257	−0.033	−32.30
w/o orthogonal projection	0.0783†	0.19×	0.4660	+0.007	−31.02
w/o retain anchor	0.0652†	0.16×	0.0000	−0.459	−2.95

The relearning suite (Figure 10; Table 19, Appendix K) attacks nine victims, with the attacked retrain oracle as control and the un-unlearned target as ceiling, so that removal means recovering no faster than an honest learner. Across three full fine-tuning rates CROP's recovery tracks the oracle's, while RMU, NPO, PDU, WGA, and NPO+SAM, the last built for this attack, rebound to 0.42–0.87, approaching pre-unlearning levels. Attacking half the forget authors restores RMU's never-attacked half (+0.749) while CROP's held-out authors move less than the oracle's own drift (+0.017 vs. +0.081), so the attacker gains from CROP only what it explicitly retrains.

7 Conclusion

By restricting ascent to a measured sharp-for-forget, flat-for-retain subspace, retain-orthogonal and retain-anchored, CROP leads RMU on held-out forget quality on two of three seeds under a leakage-free protocol, at a utility cost; as an analysis upper bound its oracle-best checkpoint lies within KS D 0.06 of the retrained model on every seed, close to the spread between independent retrains; and it relearns no faster than the attacked retrain oracle, unlike RMU, NPO, PDU, WGA, and NPO+SAM. Unlearning should be measured and optimized as distance to the retrained model, tracked oracle-free when needed (Appendix D).

Limitations. CROP pays a utility cost on TOFU-1B and in MUSE retain knowledge, costs 1.8–2.5× a first-order step, selects MUSE and WMDP checkpoints by benchmark metrics rather than leakage-free, undershoots oracle-free calibration on Qwen, spills WMDP suppression into adjacent categories, trains only a layer slab at 7B, and leaves a small off-subspace component that Appendix A's exact projector removes (Appendix O).

AI USE STATEMENT

In this work, we used generative AI tools (Claude, Anthropic) to assist with writing and editing the manuscript, with implementing evaluation and analysis scripts, and with orchestrating experiment runs under author-specified protocols. The method, the experimental design, and all scientific claims are the authors' own; every AI-assisted script, number, and passage of text was reviewed and verified by the authors. We did not use generative AI for research ideation or to generate experimental data. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments run in the open-source OpenUnlearning framework with the benchmarks' official evaluation suites. Appendix F gives the complete CROP configuration per benchmark (including the TOFU-1.5B warm-up and budget), baseline hyperparameter provenance, and compute costs; Appendix C the tracking statistic, the selection and validation/test split protocol, the retrained-oracle recipe, and the random seeds behind every multi-seed result (0, 1, 2 on TOFU-1B; 0, 3, 5 on TOFU-1.5B); Appendix D the oracle-free rule and its calibration; Appendix K the stress-test protocols; and Appendix N the curvature-verification protocols. The CROP trainer is a single self-contained file, and we will release code, configurations, and per-checkpoint evaluation logs.

ETHICS STATEMENT

This work studies machine unlearning, whose intended uses are removing private, copyrighted, or hazardous content from trained language models. All experiments use public benchmarks (TOFU, MUSE, and WMDP) with their released models and corpora; TOFU's forget set consists of fictitious authors, so no real individual's data is unlearned or exposed. The WMDP biosecurity forget corpus is distributed by its authors on request under their access terms, and we release no hazardous text, no unlearned-then-recovered outputs, and no model whose hazardous knowledge exceeds that of the public base model. Our relearning-attack suite is an evaluation of removal, not an attack tool: it fine-tunes on the benchmark's own forget data under a fixed small budget and is reported only through aggregate recovery statistics. We note two societal considerations. First, unlearning that merely suppresses knowledge can give a false sense of removal; our retrain-faithfulness criterion and relearning controls are meant to make that gap visible rather than hide it. Second, CROP's suppression on WMDP spills into adjacent benign categories, a utility cost that deployments must weigh explicitly. No human subjects or personal data were involved.

REFERENCES

Lucas Bourtoule et al. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Nicholas Carlini et al. Extracting training data from large language models. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pp. 2633–2650. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini et al. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Yijiang River Dong 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, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 8827–8840. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.444. URL https://doi.org/10.18653/v1/2025.naacl-long.444.

Vineeth Dorna et al. Openunlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Taha Entesari et al. Constrained entropic unlearning: A primal-dual framework for large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Chongyu Fan et al. Towards LLM unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025a. URL https://proceedings.mlr.press/v267/fan25e.html.

Chongyu Fan et al. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/02443e4e008231e0af0855f3cc70ed17-Abstract-Conference.html.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773. PMLR, 2020. URL http://proceedings.mlr.press/v108/farajtabar20a.html.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Isha Garg, Deepak Ravikumar, and Kaushik Roy. Memorization through the lens of curvature of loss function around samples. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15083–15101. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/garg24a.html.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9301–9309. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00932.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv:2407.21783*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842. PMLR, 2020. URL http://proceedings.mlr.press/v119/guo20c.html.

Dan Hendrycks et al. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Shengyuan Hu et al. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=fMNRYBvcQN.

Joel Jang et al. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 14389–14408. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.805. URL https://doi.org/10.18653/v1/2023.acl-long.805.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017.

Nathaniel Li et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

Yuanbang Liang and Yang Li. Grokked models are better unlearners. In *Forty-Second Annual Conference on Uncertainty in Artificial Intelligence*, 2026. URL https://openreview.net/forum?id=0BE0CvhYaN.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents, CoLLAs 2022, 22-24 August 2022, McGill University, Montréal, Québec, Canada*, volume 199 of *Proceedings of Machine Learning Research*, pp. 243–254. PMLR, 2022. URL https://proceedings.mlr.press/v199/liu22a.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jakub Lucki et al. An adversarial perspective on machine unlearning for AI safety. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=J5IRyTKZ9s.

Pratyush Maini et al. Can neural network memorization be localized? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23536–23557. PMLR, 2023. URL https://proceedings.mlr.press/v202/maini23a.html.

Pratyush Maini et al. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Lev E McKinney et al. Gauss-newton unlearning for the LLM era. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*, 2025. URL https://openreview.net/forum?id=VFfttnDvW6.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Kevin Meng et al. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Yuefeng Peng et al. Forget to know, remember to use: Context-aware unlearning for large language models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=x2eaMhqfcK.

Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Weijia Shi et al. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Weijia Shi et al. MUSE: machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 3–18. IEEE Computer Society, 2017. doi: 10.1109/SP.2017.41. URL https://doi.org/10.1109/SP.2017.41.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Lewis Tunstall et al. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aKkAwZB6JV.

Qizhou Wang et al. Rethinking LLM unlearning objectives: A gradient perspective and go beyond. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=huo8MqVH6t.

An Yang et al. Qwen2.5 technical report. *arXiv:2412.15115*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Puning Yang et al. Exploring criteria of loss reweighting to enhance LLM unlearning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/yang25ak.html.

Tianhe Yu et al. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

Zhiwei Zhang et al. Catastrophic failure of LLM unlearning via quantization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=lHSeDYamnz.

Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Appendix Table of Contents

A Proofs, Subspace Leakage, and Empirical Scope **14**

B Slab Selection at Billion-Parameter Scale **16**

C Operating-Point Selection Details **17**

D Oracle-Free Operating-Point Selection **19**

E Metric Definitions **22**

F Implementation and Computational Details **22**

G Additional TOFU Results **24**

H Full MUSE Results **25**

I WMDP Utility Breakdown and Spillover Analysis **26**

J Extended Cross-Benchmark Analysis **27**

K Stress-Test Protocols and Analysis **28**

L Distributional Faithfulness **29**

M Hyperparameter Sensitivity **30**

N Curvature-Premise Verification **31**

O Limitations **33**

A Proofs, Subspace Leakage, and Empirical Scope

Proof of Proposition 1.

Proof. Substituting equation 10, $\langle \tilde{\boldsymbol{g}}_{\text{sel}}, \boldsymbol{g}_r \rangle = \langle \boldsymbol{g}_{\text{sel}}, \boldsymbol{g}_r \rangle - \frac{\langle \boldsymbol{g}_{\text{sel}}, \boldsymbol{g}_r \rangle}{\|\boldsymbol{g}_r\|^2} \langle \boldsymbol{g}_r, \boldsymbol{g}_r \rangle = 0$. The second claim is a first-order Taylor expansion of $\mathcal{L}_r$ around $\boldsymbol{\theta}$; since the inner product vanishes, the sign of the displacement is immaterial. □

Proposition 2 (Exact retain-orthogonal projection within the memorization subspace)**.** *Let* $\boldsymbol{p} \coloneqq \boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{g}_r$. *If* $\boldsymbol{p} \neq \boldsymbol{0}$*, the matrix* $\boldsymbol{R} \coloneqq \boldsymbol{\Pi}_{\mathcal{M}} - \boldsymbol{p}\boldsymbol{p}^\top / \|\boldsymbol{p}\|^2$ *is the orthogonal projector onto* $\mathcal{M} \cap \boldsymbol{g}_r^\perp$*, and* $\boldsymbol{d} \coloneqq \boldsymbol{R}\boldsymbol{g}_f$ *satisfies* $\boldsymbol{d} \in \mathcal{M}$*,* $\langle \boldsymbol{d}, \boldsymbol{g}_r \rangle = 0$*, and* $\langle \boldsymbol{g}_f, \boldsymbol{d} \rangle = \|\boldsymbol{d}\|^2 \geq 0$*. If* $\boldsymbol{p} = \boldsymbol{0}$ *then* $\mathcal{M} \subseteq \boldsymbol{g}_r^\perp$ *and* $\boldsymbol{R} = \boldsymbol{\Pi}_{\mathcal{M}}$.

Proof. $\boldsymbol{\Pi}_{\mathcal{M}}$ is symmetric and idempotent, and $\boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{p} = \boldsymbol{p}$ because $\boldsymbol{p} \in \mathcal{M}$, so $\boldsymbol{R}$ is symmetric and $\boldsymbol{R}^2 = \boldsymbol{\Pi}_{\mathcal{M}} - 2\boldsymbol{p}\boldsymbol{p}^\top / \|\boldsymbol{p}\|^2 + \boldsymbol{p}\boldsymbol{p}^\top / \|\boldsymbol{p}\|^2 = \boldsymbol{R}$; hence $\boldsymbol{R}$ is an orthogonal projector. For any $\boldsymbol{x}$, $\boldsymbol{R}\boldsymbol{x} \in \mathcal{M}$, and $\langle \boldsymbol{R}\boldsymbol{x}, \boldsymbol{g}_r \rangle = \langle \boldsymbol{R}\boldsymbol{x}, \boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{g}_r \rangle = \langle \boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{x}, \boldsymbol{p} \rangle - \langle \boldsymbol{x}, \boldsymbol{p} \rangle = 0$, using $\langle \boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{x}, \boldsymbol{p} \rangle = \langle \boldsymbol{x}, \boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{p} \rangle = \langle \boldsymbol{x}, \boldsymbol{p} \rangle$; so the range of $\boldsymbol{R}$ lies in $\mathcal{M} \cap \boldsymbol{g}_r^\perp$. Conversely, if $\boldsymbol{m} \in \mathcal{M}$ and $\langle \boldsymbol{m}, \boldsymbol{g}_r \rangle = 0$ then $\langle \boldsymbol{m}, \boldsymbol{p} \rangle = \langle \boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{m}, \boldsymbol{g}_r \rangle = 0$ and $\boldsymbol{R}\boldsymbol{m} = \boldsymbol{m}$, so the range is exactly $\mathcal{M} \cap \boldsymbol{g}_r^\perp$. Finally $\langle \boldsymbol{g}_f, \boldsymbol{R}\boldsymbol{g}_f \rangle = \langle \boldsymbol{g}_f, \boldsymbol{R}^2 \boldsymbol{g}_f \rangle = \|\boldsymbol{R}\boldsymbol{g}_f\|^2$. The case $\boldsymbol{p} = \boldsymbol{0}$ is immediate. □

In the trainer the exact update is $\tilde{\boldsymbol{g}}_{\text{sel}} = \boldsymbol{g}_{\text{sel}} - \left(\langle \boldsymbol{g}_{\text{sel}}, \boldsymbol{g}_r \rangle / \|\boldsymbol{p}\|^2 \right) \boldsymbol{p}$, which differs from equation 10 only in subtracting along $\boldsymbol{p} = \boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{g}_r$ instead of along $\boldsymbol{g}_r$; a descent step on it displaces parameters along $-\boldsymbol{R}(-\boldsymbol{g}_f) = \boldsymbol{R}\boldsymbol{g}_f = \boldsymbol{d}$.

Table 4: Geometry of the realized forget step on TOFU-1B (shipped recipe, seeds 0–2), comparing the composed projection used in all reported results with the exact projector of Proposition 2. Step rows summarize the ascent-active steps of three full-length runs (101 composed, 95 exact) and probe rows every curvature probe. Leakage is the fraction of the projected step's norm outside $\mathcal{M}$ (the exact projector's value is the bf16 rounding floor); "ascends" means the first-order change $-\langle \boldsymbol{g}_f, \tilde{\boldsymbol{g}}_{\text{sel}} \rangle$ is positive; $\rho(\boldsymbol{d})$ is the retain-to-forget curvature ratio along the realized step; the worst in-subspace retain curvature is the largest eigenvalue in magnitude of $\boldsymbol{U}^\top \widehat{\boldsymbol{H}}_r \boldsymbol{U}$ over the kept basis $\boldsymbol{U}$, relative to the largest eigenvalue of $\boldsymbol{U}^\top \widehat{\boldsymbol{H}}_f \boldsymbol{U}$. Both projectors ascend the forget loss on every step; the composed one leaks a few percent of the step outside $\mathcal{M}$, and the exact one is more faithful on seeds 1 and 2, less on seed 0, and more stable, with many more passing checkpoints.

	Composed (shipped)	Exact (Prop. 2)
Per ascent step; median (90th percentile)		
Fraction of the step outside $\mathcal{M}$	0.029 (0.100)	0.008 (0.012)
Steps that ascend $\mathcal{L}_f$ to first order	100%	100%
$\rho(\boldsymbol{d})$ along the realized step	0.084 (0.781)	0.045 (0.255)
Steps with $\rho(\boldsymbol{d}) < \rho_{\max} = 0.25$	68%	89%
$\lvert \cos(\boldsymbol{g}_f, \boldsymbol{g}_r) \rvert$; median (max)	0.055 (0.231)	0.036 (0.250)
Per probe; median (90th percentile)		
Worst in-$\mathcal{M}$ retain / top forget curvature	0.005 (0.051)	0.017 (0.071)
Probes with a negative eigenvalue of $\boldsymbol{U}^\top \widehat{\boldsymbol{H}}_f \boldsymbol{U}$	83%	88%
Oracle-best checkpoint: KS D / KS-p / MU (passing checkpoints of 40)		
Seed 0	0.045 / 0.813 / 0.501 (28)	0.058 / 0.523 / 0.446 (11)
Seed 1	0.050 / 0.700 / 0.413 (5)	0.037 / 0.942 / 0.451 (25)
Seed 2	0.060 / 0.468 / 0.389 (1)	0.055 / 0.581 / 0.465 (21)

Subspace leakage of the composed projection. The shipped update applies equation 10 to $\boldsymbol{g}_{\text{sel}} \in \mathcal{M}$, and the two projectors do not commute, so the realized forget pseudo-gradient is $\tilde{\boldsymbol{g}}_{\text{sel}} = \boldsymbol{g}_{\text{sel}} - c\,\boldsymbol{g}_r$ with $c = \langle \boldsymbol{g}_{\text{sel}}, \boldsymbol{g}_r \rangle / \|\boldsymbol{g}_r\|^2$, whose component outside $\mathcal{M}$ is $-c\,(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\mathcal{M}})\boldsymbol{g}_r$, of relative norm at most $\lvert \cos(\boldsymbol{g}_{\text{sel}}, \boldsymbol{g}_r) \rvert$. A descent step on it changes the forget loss to first order by $\eta\,(\|\boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{g}_f\|^2 - \langle \boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{g}_f, \boldsymbol{g}_r \rangle \langle \boldsymbol{g}_f, \boldsymbol{g}_r \rangle / \|\boldsymbol{g}_r\|^2)$, which is positive unless the forget and retain gradients are strongly aligned (in two dimensions, $\mathcal{M} = \text{span}(\boldsymbol{e}_1)$, $\boldsymbol{g}_f = (1, 2)^\top$, $\boldsymbol{g}_r = (1, 1)^\top$ gives a descent direction), so the composed step is not guaranteed to be a first-order ascent direction of $\mathcal{L}_f$. Table 4 measures both effects on every eighth ascent-active step of a full-length TOFU-1B run under the shipped recipe: the off-subspace fraction $\|(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\mathcal{M}})\tilde{\boldsymbol{g}}_{\text{sel}}\| / \|\tilde{\boldsymbol{g}}_{\text{sel}}\|$, the first-order forget change $-\langle \boldsymbol{g}_f, \tilde{\boldsymbol{g}}_{\text{sel}} \rangle$ normalized by $\|\boldsymbol{\Pi}_{\mathcal{M}} \boldsymbol{g}_f\|^2$ (equal to 1 for pure in-subspace ascent), and the cosine between the forget and retain gradients that governs both. The same table reports the cross-term check of the flatness filter, namely the largest eigenvalue in magnitude of $\boldsymbol{U}^\top \widehat{\boldsymbol{H}}_r \boldsymbol{U}$ over the kept basis $\boldsymbol{U}$ at every probe against the smallest eigenvalue of $\boldsymbol{U}^\top \widehat{\boldsymbol{H}}_f \boldsymbol{U}$, together with the retain-to-forget curvature ratio along the ascent direction actually taken, and the operating point reached by the exact projector of Proposition 2 under the identical recipe and seeds. The exact projector reaches oracle-best checkpoints at $D = 0.058$, 0.037, and 0.055 on seeds 0–2 (KS-p 0.52, 0.94, 0.58) with 11, 25, and 21 passing checkpoints, against 0.045, 0.050, and 0.060 with 28, 5, and 1 for the composed form, and its final checkpoint passes on seeds 1 and 2 where the composed form's reverts; a second run of the composed form on seed 0 reached $D = 0.055$ with only 2 passing checkpoints, so the difference between the two projectors lies within run-to-run variability of the same seed. Along the realized direction the retain-to-forget curvature ratio is 0.08 at the median for the composed form (0.04 for the exact one) but exceeds $\rho_{\max}$ on 32% (11%) of ascent steps, mostly late in training when the forget curvature along the step has decayed, so the per-direction filter does not translate into a per-step flatness guarantee; the worst-case retain curvature inside the subspace stays below 0.05 of the top forget curvature at the 90th percentile, but the kept basis contains a negative forget-curvature combination on most probes, so it is not a clean eigenbasis. All reported results use the composed form; the exact projector is provided as an option and is a natural default for future work.

Empirical scope under AdamW. Proposition 1 concerns the pseudo-gradient handed to the optimizer; AdamW's moment estimates, preconditioning, and decoupled weight decay all lie outside

Table 5: Per-step instrumentation of retain-orthogonality under AdamW on TOFU-1B (median over ascent-active steps; p90 = 90th percentile of $|\cdot|$). The α=0 run isolates the projected forget path through the optimizer; the α=1 column is the full method, where the anchor deliberately adds retain descent. Proposition 1 holds exactly at the pseudo-gradient level; without the anchor, AdamW's preconditioning perturbs the realized update's alignment with $\boldsymbol{g}_r$ by less than 0.01, and the resulting retain-loss drift is predicted by the first-order Taylor term of the realized displacement.

	$\alpha = 0$ (forget path only)	$\alpha = 1$ (full method)
$\cos(\tilde{\boldsymbol{g}}_{\text{sel}}, \boldsymbol{g}_r)$ after projection	-0.0000 (p90 0.0009)	-0.0000 (p90 0.0002)
$\cos(\Delta\boldsymbol{\theta}, \boldsymbol{g}_r)$, realized AdamW step	$+0.0045$ (p90 0.0091)	-0.132 (p90 0.151)
corr(predicted, actual $\Delta\mathcal{L}_r$)	0.984	0.869
median actual $\Delta\mathcal{L}_r$ per step	$+5.1 \times 10^{-3}$	-0.447

its scope and are covered only by the measurements below. To measure what survives, we instrument every training step of two full-length TOFU-1B runs, namely the shipped recipe ($\alpha = 1$) and an anchor-free run ($\alpha = 0$) that isolates the projected forget path. At each instrumented step we record the cosine between the post-projection pseudo-gradient and $\boldsymbol{g}_r$, the cosine between the realized parameter update $\Delta\boldsymbol{\theta}$ (measured across the optimizer step) and $\boldsymbol{g}_r$, and a first-order Taylor check comparing the predicted retain-loss change $\langle \boldsymbol{g}_r, \Delta\boldsymbol{\theta} \rangle$ against the actual change on the same retain minibatch. Table 5 summarizes the outcome. Projection orthogonality is numerically exact at the pseudo-gradient level; the realized α=0 AdamW step preserves it to $|\cos(\Delta\boldsymbol{\theta}, \boldsymbol{g}_r)| < 0.01$ at the 90th percentile with a residual retain-loss drift of $+5 \times 10^{-3}$ per step that the first-order Taylor term of the realized displacement predicts at correlation 0.984, so the drift is a first-order effect of AdamW's preconditioning rather than a second-order term; in the full method (α=1) the realized step is not orthogonal to $\boldsymbol{g}_r$ by design (median cosine -0.132), because the anchor adds retain descent, and the net retain-loss change is -0.45 per step. The w/o-anchor ablation of Table 3 (utility 0.0000) shows the cumulative effect of leaving this drift unrepaired over a full run.

B Slab Selection at Billion-Parameter Scale

The curvature probe of Section 3.3 stores k direction vectors of the same dimensionality as the trainable parameters, which is infeasible for full 7B-parameter models. Following the mid-layer convention of representation-steering methods, widened from single projection matrices to full transformer blocks, CROP confines the trainable set to a contiguous slab $\mathcal{B}$ of consecutive layers with parameters $\boldsymbol{\theta}_{\mathcal{B}} \subset \boldsymbol{\theta}$, and the entire mechanism runs within $\mathbb{R}^{|\boldsymbol{\theta}_{\mathcal{B}}|}$; all other parameters are frozen. Concretely, on Llama-based 7B models we use three consecutive early-middle transformer blocks ($\sim$0.7B parameters, $\sim 10\%$ of the model); on Qwen2.5-7B we widen the slab to layers 2–9 ($\sim$1.9B parameters, $\sim 25\%$ of the model). The slab is fixed by this convention before any oracle is consulted and plays no part in checkpoint selection; the position sweep below verifies the choice after the fact. The restriction also gives CROP a smaller update budget than the baselines, which run in their published form and, for most of them, update every parameter. Curvature separability survives the restriction. At probe refreshes during MUSE unlearning, 11–12 of 12 probed directions pass the retain-flatness filter, with forget-side curvature around $+30$ against retain-side curvature near zero or negative (Appendix N). The slab is chosen wide rather than deep. In our 7B comparisons, covering more layers with fewer probe directions separated curvature more cleanly than concentrating a richer probe in a narrower slab, while remaining large enough to act as a sufficient actuator for flattening memorization.

Slab-position sensitivity. To verify that the slab's depth placement, rather than the mere existence of a restricted trainable subset, drives the 7B results, we repeat each model's full recipe with the slab moved through the network at its shipped width. On Llama-2-7B we train four 3-layer slabs (607M parameters each) placed early (layers 1–3), early-middle (5–7, the main-result configuration), middle (15–17), and late (27–29), each under the identical α-descent schedule and with the operating checkpoint selected by the same closest-to-Retrain rule in every arm. Table 6 shows that the early-middle placement is the only one that approaches the Retrain oracle across the metric suite. The early slab reaches moderate forget-side values but collapses retained knowledge (KnowMem-r 0.02 against the oracle's 0.56), while the middle and late slabs leave verbatim memorization 1.4–$1.6\times$ above the oracle level with PrivLeak far from the oracle (-86.1 and -63.0 vs. -4.7). Repeating the sweep on Qwen2.5-7B at its shipped 8-layer width (Table 7) gives the complementary picture.

Table 6: Slab-position sensitivity on MUSE-News (Llama-2-7B). Each row moves the 3-layer (607M-parameter) trainable slab to a different depth and reruns the full α-descent recipe; the operating checkpoint is selected by the same closest-to-Retrain rule in every row. The early-middle slab (layers 5–7) is the main-result configuration of Table 15. Best (closest to Retrain) in bold, second underlined.

Slab	Layers	VerbMem ↓	KnowMem-f ↓	KnowMem-r ↑	PrivLeak → Oracle	ES ↓
Retrain (Oracle)	full	0.2016	0.3279	0.5602	−4.72	0.0244
early	1–3	<u>0.2744</u>	<u>0.3877</u>	0.0216	<u>−38.94</u>	<u>0.0633</u>
early-middle	5–7	**0.1769**	0.4178	<u>0.3820</u>	**−23.78**	**0.0285**
middle	15–17	0.2907	**0.3772**	0.3463	−86.06	0.0715
late	27–29	0.3212	0.4479	**0.3946**	−63.01	0.0806

Table 7: Slab-position sensitivity on MUSE-News (Qwen2.5-7B). Each row moves the 8-layer (1.86B-parameter) trainable slab, the width of the main-result configuration, to a different depth of the 28-layer stack and reruns the full recipe (main leg + fine-walk); the operating checkpoint is selected by the same closest-to-Retrain rule in every row. The early slab (layers 2–9) is the main-result configuration of Table 16 and already abuts the bottom of the stack, so no meaningfully earlier placement exists at this width. Best (closest to Retrain) in bold, second underlined.

Slab	Layers	VerbMem ↓	KnowMem-f ↓	KnowMem-r ↑	PrivLeak → Oracle	ES ↓
Retrain (Oracle)	full	0.1918	0.1881	0.3111	0	0.0250
early	2–9	0.1951	0.2084	0.2532	−94.93	**0.0274**
early-middle	6–13	0.2031	**0.1850**	<u>0.2920</u>	**−93.18**	<u>0.0284</u>
middle	10–17	<u>0.1938</u>	<u>0.2068</u>	**0.3126**	<u>−93.24</u>	0.0318
late	19–26	**0.1926**	0.2080	0.2516	−98.28	0.0303

Every placement reaches Retrain-level forget metrics (composite distance-to-Retrain 0.014–0.024, against 0.115–0.240 on Llama-2), with deeper slabs matching or slightly exceeding the shipped 2–9 configuration; the 2–9 slab already abuts the bottom of the 28-layer stack, so no meaningfully earlier placement exists at this width. Position sensitivity is therefore architecture-dependent, decisive on Llama-2 and benign on Qwen2.5 at the shipped width, and in both models the shipped configuration sits at or near the best observed operating point, so no placement choice is decisive for the cross-benchmark conclusions.

C Operating-Point Selection Details

This appendix gives the formal definitions behind the oracle-profile tracking of Section 3.5.

Truth-ratio profile. For each forget example i with question x_i, a paraphrased correct answer $\tilde{y}_i$, and m_i perturbed (incorrect) answers $\{\hat{y}_i^{\,j}\}_{j=1}^{m_i}$, let $\bar{\ell}(y \mid x)$ denote the per-token average negative log-likelihood of answer y given question x under the current model. The truth ratio of example i is

$$\mathrm{TR}_i(\boldsymbol{\theta}) \;=\; \frac{\exp\Big(-\frac{1}{m_i}\sum_{j=1}^{m_i}\bar{\ell}\Big(\hat{y}_i^{\,j} \mid x_i\Big)\Big)}{\exp\big(-\bar{\ell}(\tilde{y}_i \mid x_i)\big)}, \tag{13}$$

i.e., the ratio of the model's mean perturbed-answer likelihood to its paraphrased-answer likelihood, computed with the benchmark's exact evaluation formula. The collection $\{\mathrm{TR}_i(\boldsymbol{\theta})\}_{i=1}^{n}$ over all n forget examples forms the model's truth-ratio profile.

Kolmogorov–Smirnov tracking. Let $F_{\boldsymbol{\theta}}$ and F^* denote the empirical cumulative distribution functions of the current model's and the retrain oracle's truth-ratio profiles, respectively. Every T_{d} steps, CROP computes the two-sample Kolmogorov–Smirnov statistic

$$D(\boldsymbol{\theta}) \;=\; \sup_{x\in\mathbb{R}} \left|F_{\boldsymbol{\theta}}(x) - F^*(x)\right|. \tag{14}$$

Both samples have fixed, equal size n, so the map from the internal statistic D to the benchmark's official KS p-value is exact and monotone, and minimizing D online is equivalent to maximizing

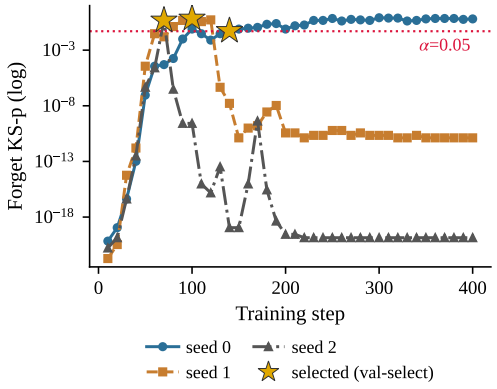


Figure 4: Official forget KS-p along the full-length trajectory for three random seeds (0, 1, 2; identical recipe). One seed converges into a passing plateau, one reverts by ten orders of magnitude after a passing pocket, one has a single passing checkpoint; utility (not shown) rises smoothly on all three, so the reversions are invisible from the training side. Stars mark the checkpoint chosen by the leakage-free selection rule (Section 3.5), which lands in the passing region on every seed.

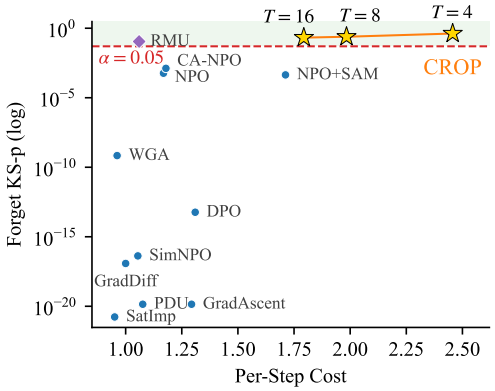


Figure 5: Measured per-step cost against official forget quality on TOFU-1B. Baselines (blue circles) cluster at 0.95–1.31× GradDiff's per-step cost, NPO+SAM's second forget pass raises it to 1.71×, and only RMU (purple diamond) crosses the $\alpha = 0.05$ threshold (dashed red). CROP (gold stars) traces the probe refresh dial $T \in \{16, 8, 4\}$, remaining above the threshold at every setting down to 1.79×, within 1.05× of the most expensive baseline.

the reported forget-quality p-value. CROP maintains the running best $D_{\text{best}} = \min_{t' \le t} D(\boldsymbol{\theta}_{t'})$ and snapshots the model at every improvement; training halts when $D(\boldsymbol{\theta}_t) \le \tau$ for a target threshold τ, or at budget exhaustion. The deliverable is the global-minimum-D snapshot. CROP does not stop at the first local minimum of D, since a shallow early minimum need not be the global one.

Leakage-free validation/test protocol. Figure 4 shows the official forget KS-p along the three full-length seed trajectories on which selection operates. Selecting a checkpoint by the same statistic on the same examples later used for evaluation constitutes adaptive model selection, so the reported TOFU-1B protocol separates the two. The forget10 split contains 20 authors with 20 questions each, contiguous in the benchmark ordering (example i belongs to author $\lfloor i/20 \rfloor$). For each of five random partitions, ten authors form the validation set and ten the test set; the checkpoint is chosen by the KS distance between the model's and the oracle's truth-ratio profiles restricted to validation authors (n=200), and all reported statistics are computed on the held-out test authors only. The procedure is repeated with training seeds 0, 1, and 2 (seed 0 is the framework default used by the originally published run), and applied with identical checkpoint density and selection access to RMU. Two honesty notes. First, on seed 0 the validation statistic, with half the sample size, selects checkpoints from an earlier, shallower pocket on 2 of 5 splits whose test-side p falls below 0.05 (0.022); the per-seed means in Table 2 include these splits. Second, full-length training with the success-stop disabled explores past the originally published stopping point, which is why the peak column of Table 2 exceeds the 0.4158 of Table 1. The table also reports a secondary Wasserstein distance W_1 between log truth-ratio profiles at each seed's modal selected checkpoint. We include it because it does not uniformly favor either method, showing near-parity between CROP and RMU at the delivered points (0.151 vs. 0.136 on average) even where the KS test separates them, so the KS verdicts above should be read as the primary but not the only distributional evidence.

Retrain-to-retrain band. A KS p-value states how separable a model is from one retrained oracle, but two retrained oracles are not identical either. To calibrate D we retrained the retain-gold model three more times under the recipe that reproduces the published oracle (learning rate 5×10^{-5}, five epochs, seeds 0–2; the framework's default fine-tuning recipe, at 10^{-5}, under-trains the model to utility 0.43 and is rejected against the published oracle at $p<10^{-3}$). The four retrains differ pairwise by $D = 0.028$–0.050 (mean 0.036) with $p = 0.70$–0.998, and they agree on utility (0.576–0.591) and forget probability (0.096–0.116). This band is the practical floor for retrain-faithfulness on

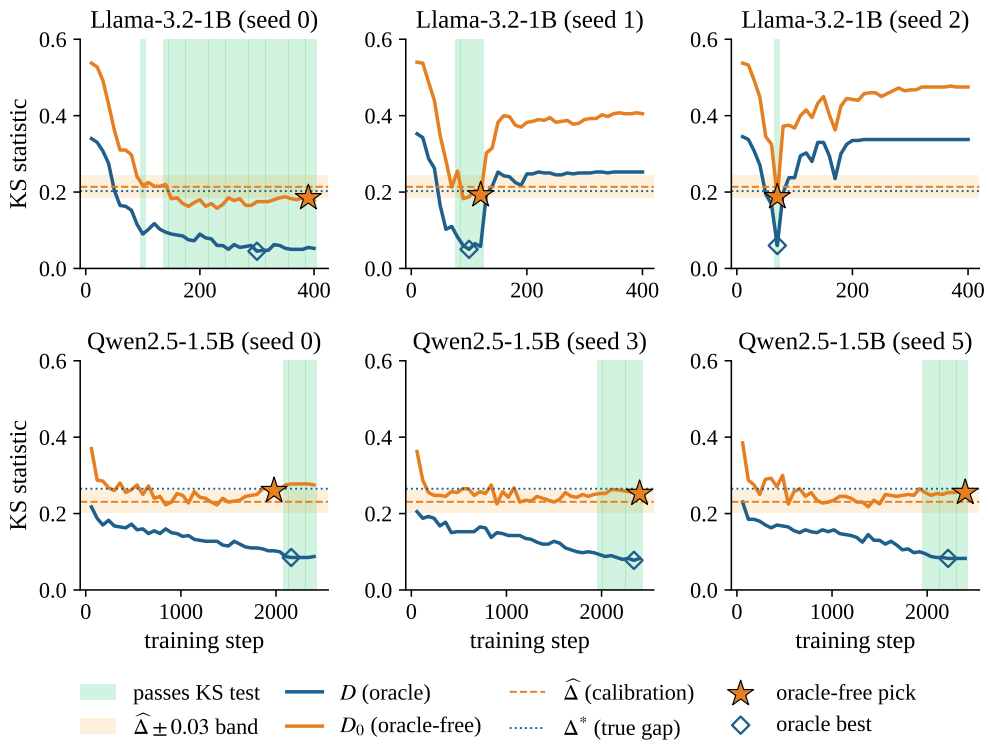


Figure 6: Oracle-free operating-point selection on three randomly drawn seed trajectories per family (top, TOFU-1B on Llama-3.2-1B, seeds 0, 1, 2; bottom, TOFU-1.5B on Qwen2.5-1.5B, seeds 0, 3, 5), each run under the identical recipe. The oracle distance D (blue) and the oracle-free distance D_0 (orange) along each dense-checkpoint trajectory, with checkpoints passing the benchmark KS test shaded green. The calibration level $\widehat{\Delta}$ (dashed orange), computed from the finetuning shift on public world-facts data, lands 0.011 above the true base-to-retrain gap Δ^* (dotted blue) on Llama and 0.034 below it on Qwen; the shaded orange band is $\widehat{\Delta} \pm 0.03$. The star marks the oracle-free pick, the latest checkpoint inside the band, and the hollow diamond the oracle's best-D checkpoint. On Llama seeds 1 and 2 the trajectory re-memorizes after a brief faithful window and the final checkpoint fails by ten or more orders of magnitude, yet the pick lands inside the window, on seed 2 at the single passing checkpoint; on Qwen the pick passes on seeds 3 and 5 and lands one checkpoint before the passing window on seed 0, where D_0 drifts out of the undershooting band as the trajectory enters it.

TOFU-1B. CROP's oracle-best checkpoint lies inside it on seeds 0 and 1 ($D = 0.045, 0.050$) and 0.010 above its upper edge on seed 2 (0.060; Table 2), and its published row sits just above it (0.063). Under the same oracle-best selection, RMU's checkpoint does not enter the band on any seed (0.058, 0.092, 0.092; Table 8), its modal leakage-free selections sit farther still (0.065, 0.108, 0.098; Table 2), and no other baseline comes within twice the band's width.

Utility-matched comparison. Because the leakage-free rows of Table 2 compare CROP and RMU at different utility levels, Table 8 asks what each method achieves at the other's utility. RMU's instrumented trajectories never drop below MU 0.48 on any seed, so no RMU checkpoint exists at CROP's published utility. In the other direction, CROP's trajectory on seed 0 passes through MU 0.50 with $D = 0.045$ (KS-p 0.813), more faithful than RMU's best checkpoint at MU 0.52 ($D = 0.058$, KS-p 0.523); on seeds 1 and 2 CROP's faithful window lies at MU 0.39–0.41 and its checkpoints above RMU's floor are early ones that have not yet forgotten. The two methods therefore occupy overlapping utility ranges on one seed, where CROP is more faithful at equal utility, and disjoint ranges on the other two, where the comparison is a genuine trade-off rather than a dominance.

D Oracle-Free Operating-Point Selection

The selection of Appendix C consults the retrain oracle's truth-ratio profile, which exists on TOFU and MUSE but not in deployments where retraining is infeasible. This appendix shows that the same

Table 8: Utility-matched comparison on TOFU-1B. For each of the three instrumented seeds (0, 1, 2), the most faithful checkpoint of each method, RMU's lowest utility along its trajectory, and CROP's most faithful checkpoint among those at or above that utility. RMU never drops below MU 0.48, so a matched comparison exists only where CROP's faithful window reaches that level (seed 0); on seeds 1 and 2 CROP's passing checkpoints sit at MU 0.39–0.41 and the two methods occupy disjoint utility ranges.

	RMU, most faithful			RMU	CROP at MU $\geq$ RMU's floor		
Seed	KS D	KS-p	MU	lowest MU	KS D	KS-p	MU
0	0.058	0.523	0.524	0.481	**0.045**	**0.813**	0.501
1	0.092	0.065	0.532	0.484	0.245	$< 10^{-3}$	0.485
2	0.092	0.065	0.531	0.483	0.195	$< 10^{-3}$	0.495

operating point can be recovered without any retrained model, using only the unlearning checkpoints, the publicly available base model, and public question-answer data outside the unlearning corpus.

Oracle-free distance. Let $\boldsymbol{\theta}_0$ denote the base pretrained model before any TOFU finetuning. Since the TOFU authors are fictitious, $\boldsymbol{\theta}_0$ has never seen them, and its truth-ratio profile on the forget set is a no-knowledge reference obtainable off the shelf. We define $D_0(\boldsymbol{\theta})$ as the two-sample KS statistic of equation 14 computed between the profiles of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Across 74 evaluated models from our experiments, spanning multiple unlearning methods, training stages, and both model families, D_0 tracks the oracle distance D with Spearman correlation 0.92 on Llama-3.2-1B and 0.95 on Qwen2.5-1.5B.

Calibration. The correct target is not $D_0 = 0$. Retain-only finetuning shifts truth-ratio profiles even on data the model has never seen, so the retrain oracle itself sits at a positive distance $\Delta^* = D_0(\text{retrain})$ from the base model, measured at 0.2025 on Llama-3.2-1B and 0.2650 on Qwen2.5-1.5B. This gap is precisely a finetuning-induced profile shift on unseen data, and such shifts are observable without the oracle. We therefore estimate the gap as $\widehat{\Delta}$, the KS statistic between the target's and the base's truth-ratio profiles on world facts. The estimate lands 0.011 above the true gap on Llama (0.2137 against 0.2025) and 0.034 below it on Qwen (0.2308 against 0.2650), where it undershoots by slightly more than the band half-width used below. The oracle-free rule selects the checkpoint whose D_0 is closest to $\widehat{\Delta}$.

Protocol. For each family we run the main-table recipe as an instrumentation run with early stopping disabled and dense checkpointing, then evaluate every checkpoint with the full benchmark suite, giving ground-truth forget quality alongside D_0 for the entire trajectory (40 checkpoints on Llama-3.2-1B over 8 epochs, 40 on Qwen2.5-1.5B over the 48-epoch budget of Table 11). These runs instrument selection and do not replace the main-table results. Figure 6 overlays D, D_0, the calibration level $\widehat{\Delta}$, and the checkpoints passing the benchmark's KS test for the three random seeds of each family.

Result. Table 9 compares the policies. On Llama-3.2-1B the oracle-free rule, which takes the latest checkpoint within ± 0.03 of $\widehat{\Delta}$, selects step 390 (KS-p 0.581, MU 0.510), within $\Delta D = 0.010$ of the oracle's own pick. On Qwen2.5-1.5B the calibration undershoots the true gap by 0.034, so D_0 drifts out of the band just as the trajectory enters the passing region, and the rule selects step 1980 (KS-p 0.030, MU 0.369), one checkpoint before the first passing one; the oracle's pick is step 2160 (KS-p 0.111, MU 0.406). Over the three Qwen seeds of Table 10 the rule's pick passes on two, against three for oracle-best selection, so on this family the rule is limited by its calibration rather than by the trajectory. A simpler variant that takes the single checkpoint closest to $\widehat{\Delta}$ is less reliable. On Llama it stops at step 120, where D_0 first crosses $\widehat{\Delta}$ inside a shallow early pocket, and that checkpoint fails the benchmark's test (KS-p 0.008); on Qwen it stops at step 900 (KS-p 2×10^{-4}), where D_0 first touches $\widehat{\Delta}$. Taking the final checkpoint without any selection illustrates why a selection signal is needed at all; it passes on Llama seed 0 and on the three Qwen seeds but fails by ten or more orders of magnitude on Llama seeds 1 and 2 (Table 10). The residual error of the oracle-free rule is bounded by the noise floor of the benchmark's own test statistic rather than by the calibration; its out-of-sample validation is reported below.

Table 9: Selection policies on the seed-0 Llama and the Qwen trajectories of Figure 6. The oracle-free rule needs no retrained model; it lands within $\Delta D = 0.010$ of the oracle's pick on Llama, and on Qwen, where the calibration undershoots the true gap by 0.034, it stops one checkpoint before the passing window ($\Delta D = 0.018$). Taking the final checkpoint without any selection is unreliable across seeds (Table 10).

	TOFU-1B (Llama-3.2-1B)				TOFU-1.5B (Qwen2.5-1.5B)			
Policy	step	D	KS-p	MU	step	D	KS-p	MU
Oracle-free rule (ours)	390	0.055	**0.581**	0.510	1980	0.102	0.030	0.369
Final checkpoint (no selection)	400	0.052	**0.641**	0.510	2400	0.088	**0.094**	0.391
Oracle-best checkpoint (needs retrain)	300	0.045	**0.813**	0.501	2160	0.085	**0.111**	0.406

Table 10: Oracle-free selection on three randomly drawn seed trajectories per family (seed values 0, 1, 2 on TOFU-1B and 0, 3, 5 on TOFU-1.5B), each run under the identical recipe. Each entry is the official KS-p of the checkpoint the rule selects; bold marks a passing checkpoint (KS-$p > 0.05$). The oracle-free rule takes the latest checkpoint within ± 0.03 of $\widehat{\Delta}$, a half-width that covers the calibration error (0.011 on Llama) with margin for the n=400 KS lattice. On TOFU-1B it passes on all three seeds where the variant that takes the single checkpoint closest to $\widehat{\Delta}$ passes on one and the final checkpoint on one. On TOFU-1.5B (48-epoch warm-up recipe) the calibration undershoots the true gap by 0.034, slightly more than the band half-width, so the rule passes on two of the three seeds where the oracle-best checkpoint passes on all three. The last row of each block is the best checkpoint an oracle with access to the retrained model would pick.

	TOFU-1B (Llama-3.2-1B)			TOFU-1.5B (Qwen2.5-1.5B)		
Selection rule	Seed 0	Seed 1	Seed 2	Seed 0	Seed 3	Seed 5
Oracle-free rule (ours)	**0.581**	**0.523**	**0.468**	0.030	**0.131**	**0.131**
Closest-to-$\widehat{\Delta}$ variant	0.008	0.016	**0.468**	10^{-4}	10^{-4}	10^{-4}
Final checkpoint (no selection)	**0.641**	10^{-11}	10^{-20}	**0.094**	**0.131**	**0.131**
Oracle-best checkpoint (needs retrain)	**0.813**	**0.700**	**0.468**	**0.111**	**0.181**	**0.131**

Behavior across seeds. Table 10 applies the rule to three randomly drawn seed trajectories per family. On Llama the closest-to-$\widehat{\Delta}$ variant recovers the operating point only on seed 2, whose D_0 sharply separates the faithful pocket from the rest of the trajectory; on seeds 0 and 1 it prefers a crossing point whose $|D_0 - \widehat{\Delta}|$ is smaller by less than the rule's own error budget, since the calibration estimate is only accurate to 0.011 on Llama and the n=400 KS statistic moves on a 0.0025 lattice. The band rule recovers a near-oracle operating point on all three (KS-p 0.581, 0.523, 0.468, against oracle-best 0.813, 0.700, 0.468). On Qwen the band rule passes on seeds 3 and 5, where the final checkpoint stays inside the band, and stops one checkpoint short on seed 0. The seed-0 pick of each family is reported as the oracle-free row of Tables 1 and 14.

Sensitivity. The real-authors split can serve as an alternative calibration source, estimating the gap at 0.170 on Llama and 0.120 on Qwen. It is less stable than world facts, and averaging the two sources undershoots the Qwen gap enough to select an early checkpoint deep in the failing region, while on Llama the averaged estimate recovers the oracle's exact pick. Broad factual data whose distribution is farthest from the unlearning corpus gives the most reliable shift reference, and we use world facts alone for the rule above.

De-saturation warm-up on TOFU-1.5B. The Qwen2.5-1.5B target memorizes the forget set to a forget NLL of 0.02 (forget probability 0.987), and cross-entropy curvature scales with $p(1-p)$, so its forget Hessian is nearly flat where training starts: the first probe finds forget curvature 79 against 754 on Llama-3.2-1B, and the curvature-restricted ascent starves. On this tier CROP therefore begins with unrestricted forget ascent, still projected orthogonal to the retain gradient and retain-anchored but without the curvature filter, until the forget NLL reaches 0.2 (reached in about twenty steps), then switches permanently to the mechanism of Section 3. The warm-up is off on every other benchmark. With the 48-epoch budget the ascent gate closes near step 1200 and the remaining retain-only phase

Table 11: CROP configuration by benchmark. FD = finite difference; "slab" restricts all mechanism state to the listed consecutive layers.

	TOFU-1B	TOFU-1.5B	MUSE (7B)	WMDP (7B)
Learning rate	7×10^{-5}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Batch size	8	8	2	2
Budget	8 ep	48 ep	10 ep / leg	10 ep
Probe size k	8	12	12	12
Power iterations	4	4	4	4
Refresh interval	4	4	8	8
Flatness ratio ρ	0.25	0.25	0.25	0.25
Retain anchor α	1.0	2.0	$1.0 \rightarrow 0.5$	1.0
FD step ϵ	10^{-3}	10^{-3}	10^{-3}	10^{-3}
Forget cap	3.0	3.0	3.0	6.0
De-saturation warm-up	–	to NLL 0.2	–	–
Trainable set	full	full	slab	slab
Selection	best-D	best-D	ckpt sweep	ckpt sweep

widens the truth-ratio profile toward the oracle's; the same recipe at 36 epochs ends one lattice step outside the equivalence band ($D = 0.0975$), and a retain-only continuation at a constant learning rate does not close the gap, so the longer schedule, not the learning rate, is what carries the Qwen trajectory into the band. The best checkpoint of each of the three seeds in Table 10 lands inside the band.

E Metric Definitions

On TOFU, forget quality (Forget KS-p) is the p-value of a two-sample Kolmogorov–Smirnov test comparing the model's forget truth-ratio profile (equation 13) against the retain-gold model's, so larger values mean the two distributions are harder to separate. Model Utility (MU) aggregates likelihood, ROUGE (Lin, 2004), and truth-ratio (TR) scores over retain-side authors, real-world facts, and world knowledge. Forget Prob and Forget ROUGE report the mean answer likelihood and the ROUGE-L recall on forget questions, and both are judged by proximity to the oracle's values rather than by minimization. Extraction Strength (ES) measures the fraction of a forget answer that greedy decoding recovers from the shortest sufficient prefix, an adversarial completion probe. PrivLeak compares min-$k\%$ (Shi et al., 2024) membership-inference AUC between the unlearned and retrained models; a value near zero indicates that the membership signal of the forget set has been neutralized, while large negative or positive values reveal under- or over-exposure. On MUSE, VerbMem scores verbatim continuation of forget documents, KnowMem-f and KnowMem-r score question answering on forget and retain knowledge respectively, and PrivLeak is defined as above. On WMDP, forgetting is measured by multiple-choice accuracy on the hazardous Bio, Cyber, and held-out Chem splits, with random-guessing accuracy 0.25 as the ideal, and utility is measured by MMLU (Hendrycks et al., 2021) and MT-Bench (Zheng et al., 2023).

F Implementation and Computational Details

CROP configurations. Table 11 lists the full CROP configuration per benchmark. Two knobs adapt across settings, both for measured reasons. The deeper memorization of the TOFU-1.5B tier requires the stronger dose, and the 7B benchmarks use the mid-layer slab of Appendix B with per-leg retain-anchor reduction, since no truth-ratio profile exists to stop against. Everything else is shared. On benchmarks without an oracle profile, checkpoints are saved every half epoch and the deliverable is the checkpoint closest to the retrain reference under the benchmark's own metrics (normalized mean gap), verified by an independent re-evaluation. This rule consults the evaluation metrics themselves, since MUSE and WMDP provide no truth-ratio profile on which a validation/test split could be formed, and it is applied symmetrically to every method, so the MUSE and WMDP rows are best-attainable operating points under a shared selection rule rather than leakage-free estimates.

Baseline configurations and provenance. All baselines run in the OpenUnlearning framework (Dorna et al., 2025). Wherever the original authors publish benchmark-specific hyperparameters we use them (NPO β=0.1; SimNPO's per-benchmark configs, e.g. MUSE-News β=0.7, γ=1, λ=0.1 vs. TOFU β=4.5; PDU's published per-benchmark ϵ; RMU's author recipe with steering coefficient 6.5, α=1; WGA and SatImp from their authors' recommended sweeps, best value reported).

CA-NPO denotes the context-aware NPO instantiation of Peng et al. (2026), whose modular context term applies to several base objectives; we adopt its NPO variant with KL context distillation on the grounded contextual-use stream, computed in the distillation direction $\mathrm{KL}(p_{\mathrm{orig}} \,\|\, p_w)$. A controlled rerun with the paper's direction $\mathrm{KL}(p_w \,\|\, p_{\mathrm{orig}})$ leaves every reported metric within run-to-run variation. Where no benchmark-specific setting exists, we use the framework's published defaults. One provenance caveat deserves mention. A re-run of NPO under current framework defaults lands at a different operating point than the reported TOFU-1B row (MU 0.396 at KS-p 2.6×10^{-5}, versus the reported 0.210 at 6×10^{-4}); both points fail forget quality by orders of magnitude, so no conclusion changes, but we report the discrepancy and release both configurations. NPO+SAM (Fan et al., 2025a) runs under NPO's defaults with the authors' official SAM step ported into the framework (radius $\rho = 0.01$, their MUSE default). On WMDP it trains on a seeded random document subset of the joint corpus (40M characters per side, about ten million tokens, eight times what the 160-step budget consumes, at the corpus's natural 97:3 bio-to-cyber mixture), since the framework tokenizes a corpus as a single string and the full joint corpus exceeds available memory. At the authors' WMDP radius $\rho = 0.1$ the perturbation saturates the NPO loss at the SAM point on every step, so the forget gradient vanishes and the run leaves WMDP-Bio at the target's 0.645 (MMLU 0.575, MT-Bench 4.64); we therefore report $\rho = 0.01$. Methods that overshoot the retrained reference at the fixed 10-epoch MUSE budget are additionally given per-epoch checkpoint selection under the MUSE stopping rule (the checkpoint closest to the retrain reference), which uniformly improves them; the reported baseline rows are therefore each method's best defensible operating point, not its worst.

Selection parity on TOFU-1B. CROP's delivered checkpoint is chosen from a dense trajectory, so a fair question is whether the baselines would pass if given the same access. Table 12 answers it. We re-trained every TOFU-1B baseline from scratch under its published recipe with a checkpoint saved every epoch, scored all checkpoints against the retain-gold profile, and report the checkpoint with the smallest KS D beside the published row. The re-trained final models reproduce the published rows, which verifies the recipes; PDU's row, for instance, is matched to four decimals by the paper's initial dual variable of 100 with $\epsilon = 0.3$. Oracle-best selection moves no baseline into the passing region: the best checkpoint of any method other than RMU reaches at most KS-p 0.010 (PDU at epoch 5), and RMU's best (0.131 at epoch 7) is within noise of its row; NPO+SAM's faithfulness plateaus from epoch 6 onward ($D = 0.145$–0.147) while its utility recovers, so its best checkpoint is no more faithful than its final one, which re-trains to KS-p 3.3×10^{-4} and MU 0.317 against the published 4.4×10^{-4} and 0.289, within single-seed run-to-run variation. The published rows are therefore each baseline's best operating point under the same selection access CROP receives, and the main-table comparison stands. WGA deserves one note. Its authors' TOFU recipe, a weighting temperature of 7 for five epochs at batch 16 with a KL regularizer, was tuned on 7B models and does not unlearn Llama-3.2-1B at that budget (forget probability 0.86 against the target's 0.88, KS-p 8×10^{-22}, MU 0.602); the reported row, reproduced exactly by retain weight 0.1 for five epochs, is the setting under which WGA forgets at this scale.

Computational cost. CROP's overhead over first-order methods is the curvature probe. Each refresh performs $k(n_{\mathrm{power}}+2)$ finite-difference gradient evaluations, i.e. $\sim$48 extra evaluations every 4 steps at the TOFU configuration, and the profile rounds add one scoring pass over the forget and perturbed sets every T_{d} steps. Measured on a single 96 GB workstation GPU, a complete TOFU-1B run (training, profile tracking, and best-D snapshots) takes well under an hour; TOFU-1.5B at the extended budget takes $\sim$3.5 hours; a MUSE-7B leg with the layer slab takes $\sim$3 hours plus $\sim$2 minutes per checkpoint evaluation in the selection sweep. The slab restriction keeps the probe's direction vectors at $\sim$1.4 GB each in half precision, so the full mechanism fits alongside the 7B model and optimizer on one GPU.

The efficiency dial. The gradient-evaluation count overstates the realized cost, because probe evaluations run on small probe batches and cost far less than full optimizer steps. Table 13 reports per-step cost measured under an identical 200-step protocol for all nine baselines, NPO+SAM, and CROP at three probe refresh intervals, with the steady-state rate taken over the same step window on the same GPU. Costs are expressed relative to GradDiff, which Section 6.1 identifies as CROP without the curvature restriction, so the ratio isolates exactly what the probe adds. Measured this way, the shipped configuration costs $2.46\times$ GradDiff, and the refresh interval acts as a dial. Halving the probe frequency (T=8) brings the cost to $1.98\times$ at KS-p 0.2441 with utility unchanged, and

Table 12: Selection parity for the TOFU-1B baselines. Each method is re-trained from scratch under its published recipe with a checkpoint saved every epoch (20 for CA-NPO, 10 otherwise), every checkpoint is scored against the retain-gold profile, and the checkpoint with the smallest KS D is reported beside the published row. The re-trained final models reproduce the published rows, and oracle-best selection lifts no baseline into the passing region; RMU's best checkpoint is within noise of its row. The WGA row uses the retain weight 0.1 and five epochs that reproduce its published row; the authors' own recipe, tuned for 7B models, does not unlearn at this scale (see text).

	published row		oracle-best checkpoint			
Method	KS-p	MU	epoch	KS D	KS-p	MU
GradAscent	1.4×10^{-20}	0.585	10	0.338	1.4×10^{-20}	0.585
NPO	5.8×10^{-4}	0.210	7	0.138	1.0×10^{-3}	0.207
DPO (IdK)	5.8×10^{-14}	0.561	10	0.278	5.8×10^{-14}	0.561
RMU	0.111	0.529	7	0.083	0.131	0.501
WGA	6.9×10^{-10}	0.222	2	0.208	6.0×10^{-8}	0.101
NPO+SAM	4.4×10^{-4}	0.289	6	0.145	4.4×10^{-4}	0.251
SimNPO	4.2×10^{-17}	0.583	8	0.307	4.2×10^{-17}	0.582
SatImp	1.7×10^{-21}	0.596	5	0.340	6.9×10^{-21}	0.595
PDU	1.4×10^{-20}	0.590	5	0.115	1.0×10^{-2}	0.579
CA-NPO	1.3×10^{-3}	0.558	7	0.128	3.0×10^{-3}	0.356
CROP (published row)	0.416	0.459	–	0.063	0.416	0.459

Table 13: Measured per-step cost and forget quality on TOFU-1B. All timings use an identical 200-step protocol on the same GPU (batch size 8, no checkpoint saving), with the steady-state rate taken over the same step window for every method; baseline KS-p and MU are the Table 1 values, except GradDiff, which is the w/o-curvature-restriction variant of Table 3 and serves as the cost unit since it shares CROP's computational graph without the probe. T is CROP's probe refresh interval; the dial rows report their own delivered best-D evaluations. † KS-p > 0.05.

Method	s/step	×GradDiff	Forget KS-p	MU
GradDiff	1.92	1.00	1.2×10^{-17}	0.445
GradAscent	2.48	1.29	1.4×10^{-20}	0.585
NPO	2.24	1.17	5.8×10^{-4}	0.210
DPO	2.51	1.31	5.8×10^{-14}	0.561
RMU	2.03	1.06	$0.1111^\dagger$	0.529
WGA	1.84	0.96	6.9×10^{-10}	0.222
NPO+SAM	3.28	1.71	4.4×10^{-4}	0.289
SimNPO	2.02	1.05	4.2×10^{-17}	0.583
SatImp	1.82	0.95	1.7×10^{-21}	0.596
PDU	2.06	1.08	1.4×10^{-20}	0.590
CA-NPO	2.26	1.18	1.3×10^{-3}	0.558
CROP (T=16)	3.44	1.79	$0.2107^\dagger$	0.431
CROP (T=8)	3.80	1.98	$0.2441^\dagger$	0.459
CROP (T=4)	4.71	2.46	$\mathbf{0.4158}^\dagger$	0.459

T=16 reaches $1.79\times$, only $1.05\times$ NPO+SAM, the most expensive baseline at $1.71\times$ because its SAM point costs a second forget pass, at KS-p 0.2107. Every dial setting stays above the $\alpha = 0.05$ non-rejection threshold that ten of the eleven baselines never reach. The cheaper probes also stop sooner. Under the standard success-stop rule the T=8 and T=16 runs deliver their checkpoints after 265 and 223 steps, 19.8 and 14.9 minutes of total training, the same wall-clock class as the baselines' full training runs at the benchmark's standard budget. Figure 5 plots the measured frontier.

G Additional TOFU Results

Table 14 reports the complete TOFU-1.5B (Qwen2.5-1.5B-Instruct) results discussed in Section 5.1. The pattern of the 1B tier repeats at this scale. With the de-saturation warm-up and the 48-epoch budget of Appendix F, CROP attains KS-p 0.111 ($D = 0.085$), the only entry for which the KS equality test is not rejected at $\alpha = 0.05$, against 1×10^{-3} for the best baseline, while landing on

Table 14: Results on TOFU-1.5B (Qwen2.5-1.5B-Instruct, forget10). Best in bold, second underlined. Prob, ROUGE, TR, and ES are judged by closeness to the retain-gold oracle (O); CROP-OF is CROP under oracle-free selection. KS D is the effect size against the retain-gold profile with a 95% example-bootstrap interval; † marks equality not rejected at $\alpha = 0.05$ ($D \leq 0.0925$ at n=400). Every row is trained from the same target and scored against the same retain-gold model; CROP uses the de-saturation warm-up and the 48-epoch budget of Appendix F.

Method	KS-p ↑	KS D [95% CI] ↓	MU ↑	Prob → O	ROUGE → O	TR → O	ES → O	PrivLeak → 0
Target	2×10^{-18}	0.320 [0.29, 0.38]	0.4882	0.9866	0.9916	0.4036	0.9960	−99.95
Retain-gold (O)	1.0000	0.000 (ref.)	0.4848	0.0538	0.3534	0.5480	0.1099	16.92
GradAscent (Jang et al., 2023)	2×10^{-18}	0.320 [0.29, 0.38]	**0.4892**	0.9771	0.9645	0.4013	0.9687	−99.84
NPO (Zhang et al., 2024)	1×10^{-13}	0.275 [0.24, 0.34]	0.3900	0.0856	0.3468	0.4289	0.1293	−84.90
DPO (IdK) (Maini et al., 2024)	4×10^{-17}	0.307 [0.28, 0.36]	0.4552	0.9509	0.7785	0.4307	0.9529	−99.72
RMU (Li et al., 2024)	1×10^{-3}	0.135 [0.09, 0.20]	0.2577	0.0485	0.2948	0.5077	0.0871	−42.18
WGA (Wang et al., 2025)	1×10^{-14}	0.285 [0.26, 0.34]	0.4291	0.2083	**0.3560**	0.4250	0.1334	−89.30
NPO+SAM (Fan et al., 2025a)	1×10^{-13}	0.275 [0.22, 0.34]	0.4209	0.2502	0.5110	**0.5753**	0.0300	−83.90
SimNPO (Fan et al., 2025b)	6×10^{-18}	0.315 [0.28, 0.38]	0.4749	0.5142	0.5086	0.3910	0.2950	−97.79
SatImp (Yang et al., 2025)	3×10^{-18}	0.318 [0.28, 0.38]	0.4847	0.7959	0.6140	0.4022	0.4145	−99.04
PDU (Entesari et al., 2025)	3×10^{-9}	0.225 [0.21, 0.28]	0.1860	0.0489	0.1201	0.7256	**0.1034**	−63.28
CA-NPO (Peng et al., 2026)	1×10^{-16}	0.302 [0.27, 0.36]	0.4736	0.1834	0.4115	0.4115	0.1791	−90.01
CROP (Ours)	**0.1111**†	**0.085** [0.07, 0.13]	0.4065	**0.0514**	0.3222	0.5934	0.1017	26.59
CROP-OF (Ours)	0.0299	0.102 [0.08, 0.14]	0.3690	0.0594	0.3208	0.6048	0.1023	**26.58**

Table 15: Results on MUSE-News (Llama-2-7B). Best in bold, second underlined (closest to Retrain). Dashes mark methods MUSE cannot instantiate; DPO needs preference pairs with IdK targets, unavailable from raw news corpora.

Method	VerbMem ↓	KnowMem-f ↓	KnowMem-r ↑	PrivLeak → Oracle	ES ↓
Target (no unlearning)	0.5789	0.6443	0.5552	−99.81	0.2954
Retrain (Oracle)	0.2016	0.3279	0.5602	−4.72	0.0244
GradAscent (Jang et al., 2023)	0.5512	0.6578	0.5439	−99.79	0.2902
NPO (Zhang et al., 2024)	0.5024	0.6417	0.5324	−99.64	0.2120
DPO (IdK) (Maini et al., 2024)	–	–	–	–	–
RMU (Li et al., 2024)	0.3821	0.5222	0.4866	−99.81	0.1189
WGA (Wang et al., 2025)	0.1377	**0.3513**	0.3856	−**16.06**	0.0296
NPO+SAM (Fan et al., 2025a)	0.4111	0.6360	**0.5370**	−99.60	0.1390
SimNPO (Fan et al., 2025b)	0.3834	0.6049	0.5061	−99.85	0.1267
SatImp (Yang et al., 2025)	0.5060	0.6291	0.5096	−99.87	0.2125
PDU (Entesari et al., 2025)	0.3828	0.5412	0.5026	−99.73	0.1151
CA-NPO (Peng et al., 2026)	0.5139	0.6099	0.5332	−99.68	0.2194
CROP (Ours)	**0.1769**	0.4178	0.3820	−23.78	**0.0285**

the oracle's forget probability (0.051 against 0.054) and extraction strength (0.102 against 0.110); its PrivLeak (26.6) has the smallest magnitude among methods that forget.

Across both TOFU tables the baselines fall into three regimes. A first group barely moves. SatImp preserves utility near the target level but leaves forget probability above 0.80, and GradAscent at this budget changes almost nothing. A second group forgets by destroying the model. NPO on TOFU-1B drops model utility to 0.2100, RMU on TOFU-1.5B collapses to 0.2577, and PDU on TOFU-1.5B falls to 0.1860, each paying roughly half the model to buy a KS-p that remains many orders of magnitude below the $\alpha = 0.05$ non-rejection threshold. A third group, containing RMU at 1B and CROP, occupies the genuine trade-off region. Within that region CROP is the only method that converts its utility budget into a forget profile the benchmark's test does not reject.

H Full MUSE Results

Tables 15 and 16 report the complete MUSE-News results summarized in Section 5.2. On Llama-2-7B (Table 15) the field splits sharply. Seven of the eight baselines with reported results leave PrivLeak at the −99 floor and VerbMem above 0.38, meaning they remove neither the verbatim text nor the membership signal. Only WGA and CROP move meaningfully toward the retrained reference, and the decomposition between them is informative. WGA achieves the best KnowMem-f but overshoots on VerbMem, landing below the reference, so its verbatim forgetting is again distinguishable from the retrained model, only in the opposite direction. CROP moves both axes jointly and lands closest on VerbMem (0.1769) and extraction strength (0.0285) while placing second on

Table 16: Results on MUSE-News (Qwen2.5-7B). Best in bold, second underlined (closest to Retrain). Dashes mark methods MUSE cannot instantiate; DPO needs preference pairs with IdK targets, unavailable from raw news corpora.

Method	VerbMem ↓	KnowMem-f ↓	KnowMem-r ↑	PrivLeak → Oracle	ES ↓
Target (no unlearning)	0.5964	0.2751	0.2944	−100	0.5512
Retrain (Oracle)	0.1918	0.1881	0.3111	0	0.0250
GradAscent (Jang et al., 2023)	0.1610	0.2665	0.3257	−100	0.0363
NPO (Zhang et al., 2024)	0.4042	0.2471	0.3168	−100	0.3109
DPO (IdK) (Maini et al., 2024)	–	–	–	–	–
RMU (Li et al., 2024)	<u>0.1876</u>	<u>0.2163</u>	0.3101	−99.97	**0.0273**
WGA (Wang et al., 2025)	0.2421	0.2594	<u>0.3533</u>	−100	0.0729
NPO+SAM (Fan et al., 2025a)	0.3969	0.2505	0.3394	−100	0.2683
SimNPO (Fan et al., 2025b)	0.1755	0.3424	**0.3599**	−100	0.0379
SatImp (Yang et al., 2025)	0.5105	0.2585	0.3231	−100	0.4362
PDU (Entesari et al., 2025)	0.0922	0.2670	0.3201	**26.35**	0.0349
CA-NPO (Peng et al., 2026)	0.4706	0.2956	0.3286	−100	0.4114
CROP (Ours)	**0.1951**	**0.2084**	0.2532	<u>−94.93</u>	<u>0.0274</u>

KnowMem-f and PrivLeak. Both methods pay in retain knowledge, however: KnowMem-r falls to 0.3820 for CROP and 0.3856 for WGA against the retrained model's 0.5602, whereas the baselines that barely forget keep it above 0.48. Uniformity across metrics, rather than a single headline number, is what approximating the retrained model actually requires.

On Qwen2.5-7B (Table 16) CROP attains the closest value to the retrained model on both forget-side axes simultaneously, VerbMem 0.1951 against a reference of 0.1918 and KnowMem-f 0.2084 against 0.1881, with RMU second on both. The reported checkpoint is one point on a local frontier rather than a unique optimum. Neighboring checkpoints from the same training trajectory, omitted from the table, trade the axes against one another; one drawn fifty steps later gives up one point of VerbMem for a better KnowMem-f (0.1975) and a five-point recovery of retain knowledge (0.3031 against the reported 0.2532), and across dense checkpoint sweeps we consistently observe verbatim and knowledge-level forgetting fluctuating in opposite directions. Operating-point selection is therefore not a formality but a genuine part of the method, which is precisely why CROP tracks its distance to the oracle online whenever the benchmark permits it (Section 3.5). NPO+SAM (Fan et al., 2025a), run under NPO's MUSE recipe with the authors' radius $\rho = 0.01$ and selected by the same closest-to-retrain rule, behaves like its base objective on both families: VerbMem 0.41 and KnowMem-f 0.64 on Llama-2-7B (epoch 10, the authors' choice for News, gives 0.48 and 0.65) and 0.40 and 0.25 on Qwen2.5-7B, with PrivLeak at the floor on both, so the sharpness term moves neither the verbatim nor the knowledge-level operating point of NPO.

The tables also document an honest limitation. On Qwen2.5-7B CROP's PrivLeak remains near the -95 floor while PDU reaches the positive side of zero. CROP shapes loss curvature, not the min-k likelihood profile that the PrivLeak attack measures, so methods with explicit probability shaping retain an advantage on this one axis.

I WMDP Utility Breakdown and Spillover Analysis

Table 17 reports aggregate MMLU and the four MMLU subcategories topically adjacent to the unlearned WMDP domains, complementing the forget-side results of Table 18. Among methods that meaningfully unlearn, RMU is best on aggregate MMLU (0.5576), college biology, college computer science, and MT-Bench, with CROP second on each, so the two methods bracket the functional frontier, with RMU suppressing less and keeping more while CROP suppresses more and pays roughly ten points of MMLU for the additional depth. NPO+SAM sits with NPO here as on every other benchmark (WMDP-Bio 0.623 against NPO's 0.617, MMLU 0.570, MT-Bench 4.26), so the sharpness term adds no suppression.

The cost of CROP's limited target specificity appears in the adjacent benign categories. Virology drops to 0.2048 and college biology to 0.5069. The spillover is not uniform, however. On computer security, the benign category adjacent to the cyber forget domain, CROP retains the most knowledge of any method that meaningfully unlearns (0.34 against RMU's 0.27), though every such method loses roughly half the target's 0.68. We report this as a limitation of target specificity rather than

Table 17: Utility breakdown on WMDP (Zephyr-7B-β). Aggregate MMLU and the four MMLU subcategories topically adjacent to the unlearned domains (College Biology, Virology, College Computer Science, and Computer Security). Best in bold, second underlined; marks are assigned among methods that meaningfully unlearn (WMDP-Bio reduced by at least 10 points), as utility is vacuous for methods that leave hazardous knowledge intact. Forget-side results are reported in Table 18. Dashes mark methods WMDP cannot instantiate; DPO needs preference pairs and CA-NPO a grounded contextual-use stream, neither available from raw hazard corpora.

Method	MMLU ↑	College Bio ↑	Virology ↑	College CS ↑	Comp. Security ↑
Target (no unlearning)	0.5890	0.6597	0.5181	0.4900	0.6800
Retrain (Oracle)	—	—	—	—	—
GradAscent (Jang et al., 2023)	0.2465	0.2222	**0.3193**	0.1500	0.2400
NPO (Zhang et al., 2024)	0.5588	0.6528	0.4819	0.4100	0.6200
DPO (IdK) (Maini et al., 2024)	—	—	—	—	—
RMU (Li et al., 2024)	**0.5576**	**0.6181**	<u>0.3133</u>	**0.5000**	<u>0.2700</u>
WGA (Wang et al., 2025)	0.5708	0.6667	0.4819	0.4200	0.5800
NPO+SAM (Fan et al., 2025a)	0.5699	0.6181	0.4639	0.5100	0.7200
SimNPO (Fan et al., 2025b)	0.5645	0.6736	0.4819	0.4100	0.6600
SatImp (Yang et al., 2025)	0.5813	0.6944	0.4940	0.4600	0.7100
PDU (Entesari et al., 2025)	0.2346	0.2708	0.2048	0.2800	<u>0.2700</u>
CA-NPO (Peng et al., 2026)	—	—	—	—	—
CROP (Ours)	<u>0.4515</u>	<u>0.5069</u>	0.2048	<u>0.3700</u>	**0.3400**

Table 18: Results on WMDP (Zephyr-7B-β, joint bio+cyber corpus); WMDP-Chem is a held-out domain measuring cross-domain suppression. Best in bold, second underlined, assigned among non-degenerate methods only (for the hazard columns, MMLU within 15 points of target; for MT-Bench, WMDP-Bio reduced by $\geq$10 points). Full utility breakdown in Table 17. Dashes mark methods WMDP cannot instantiate; DPO needs preference pairs and CA-NPO a grounded contextual-use stream, neither available from raw hazard corpora.

Method	WMDP-Bio ↓	WMDP-Cyber ↓	WMDP-Chem ↓	MT-Bench ↑
Target (no unlearning)	0.6449	0.4227	0.4779	5.6380
Retrain (Oracle)	—	—	—	—
GradAscent (Jang et al., 2023)	0.2474	0.2431	0.2623	1.0000
NPO (Zhang et al., 2024)	0.6174	0.3754	0.4632	2.7190
DPO (IdK) (Maini et al., 2024)	—	—	—	—
RMU (Li et al., 2024)	<u>0.3362</u>	<u>0.2693</u>	<u>0.4314</u>	**5.4500**
WGA (Wang et al., 2025)	0.5884	0.3538	0.4706	3.8310
NPO+SAM (Fan et al., 2025a)	0.6229	0.4051	0.4534	4.2625
SimNPO (Fan et al., 2025b)	0.6410	0.4187	0.4436	4.2330
SatImp (Yang et al., 2025)	0.6614	0.4182	0.4657	5.0060
PDU (Entesari et al., 2025)	0.2529	0.2587	0.2426	1.0000
CA-NPO (Peng et al., 2026)	—	—	—	—
CROP (Ours)	**0.3087**	**0.2506**	**0.3725**	<u>4.2810</u>

as a tunable forgetting radius: nothing in the method localizes suppression to the hazardous domain beyond the forget corpus itself, and narrowing it is future work.

J Extended Cross-Benchmark Analysis

Forgetting depth and utility lie on a hard Pareto surface, and CROP extends its reachable end. Across every benchmark the methods trace a frontier in the plane spanned by forgetting depth and model utility. On TOFU-1B, CROP holds utility 0.4585 against RMU's 0.5288, and our hyperparameter sweeps confirm the coupling is structural rather than a tuning artifact. Raising the retain-anchor strength from 1 to 2 collapses KS-p from 0.4158 to 0.0652 while utility stays flat between 0.44 and 0.49, so no free utility exists beyond the operating point we report. The correct reading is not that CROP dominates every baseline everywhere. It is that CROP reaches the deepest point of the frontier on every seed (peak KS-p 0.47–0.81 across the three instrumented trajectories, against RMU's 0.065–0.52 under identical treatment), while RMU occupies a shallower

point with roughly 0.10 more utility and a faithful operating point that exists on only one seed in three. A practitioner who needs a reliably attainable faithful model has one option in these tables; a practitioner who needs mild suppression at maximal utility may still prefer RMU, and the tables let them see that choice explicitly.

Deeper is not better, and distance to retrain is the right objective. The tables repeatedly show methods being rewarded by naive suppression metrics for overshooting the oracle, PDU's forget probability of 0.0067 on TOFU-1B and its VerbMem of 0.0922 on MUSE-Qwen being the clearest cases. Under closest-to-retrain marking these entries lose to methods that land at the reference, and under the KS test the overshoot is detected as distinguishability. This reframing explains an apparent paradox in Table 1, namely that CROP wins forget quality decisively while several baselines post lower raw forget scores. Retrain faithfulness, not maximal erasure, is what CROP optimizes, and it is the only quantity for which the retrained model is by definition optimal.

Consistency across metrics and scales is itself a result. Under the closest-to-retrain criterion, CROP's two rows record thirteen field-best entries across the four oracle-equipped tables (five on TOFU-1B, four on TOFU-1.5B, two on each MUSE model; eight of them belong to the oracle-selected row alone), plus four on WMDP under its suppression criterion, which has no retrain reference, and it is the only method that is best or second best on the primary forget-side axis of every benchmark and every architecture family tested (Llama, Qwen, and Mistral-derived Zephyr). Baselines are strong in one regime and absent in another. WGA is competitive on MUSE-Llama2 but fails TOFU forgetting by nearly nine orders of magnitude in KS-p and barely moves WMDP-Bio. RMU is the strongest TOFU-1B baseline yet collapses utility at 1.5B and stalls on MUSE-Llama2. The stability of CROP across settings suggests that curvature-separable memorization extends beyond any single benchmark, though we claim it only for the models, scales, and benchmarks tested here.

The residual weaknesses are localized and measurable. Two costs recur. First, the utility ceiling on TOFU-1B and the MMLU reduction on WMDP are the price of operating at the deep end of the frontier (Appendix O). Second, min-k privacy leakage on MUSE-Qwen resists curvature shaping (Appendix H). Neither weakness is hidden by the aggregate numbers, and both indicate concrete directions for future work, coupling the curvature subspace with explicit likelihood shaping being the most natural.

K Stress-Test Protocols and Analysis

This appendix expands the stress tests summarized in Section 6.2.

Forget-set-size scaling. Following the protocol of the NPO, SimNPO, and WGA papers (Zhang et al., 2024; Fan et al., 2025b; Wang et al., 2025), we vary the forget-set size across five splits, namely the standard TOFU splits (1%, 5%, 10%) plus 3% and 7% splits constructed as deterministic prefix-author subsets of forget10 (retain oracles retrained on the corresponding complements under the recipe that reproduces the published oracles, learning rate 5×10^{-5} for five epochs, see the retrain-to-retrain band in Appendix C; perturbed sets are the row-aligned slices of forget10's), transferring every method's 10%-split hyperparameters unchanged; the success-stop criterion is held fixed across splits by scaling the stop threshold to the same significance level at each sample size. Figure 8 shows the result. CROP crosses the $\alpha = 0.05$ threshold at all five sizes (KS-p 0.2657, 0.697, 0.2705, 0.681, 0.4158 at 1/3/5/7/10%) at a stable utility of 0.40–0.46; the 7% split required a doubled budget, its first trajectory minimum being shallow (the multi-phase behavior of Section 3.5). The strongest baselines do not survive the transfer. RMU reaches the threshold only at 10% (1.4×10^{-2}, 1.2×10^{-4}, 4.9×10^{-10}, 2.0×10^{-6} at 1/3/5/7%), and WGA at none (6.8×10^{-3} down to 6.9×10^{-10}), both varying non-monotonically across splits. We attribute the difference to the fact that CROP's operative quantities are measured online, so the curvature probe and the oracle distance adapt to whatever forget set is given, whereas the baselines' fixed loss-shaping coefficients implicitly encode the forget-set scale they were tuned on. Per-split retuning would likely improve the baselines; CROP required none. Figures 7 and 8 plot KS D and KS-p against the forget-set size for CROP, RMU, and WGA.

Robustness to relearning. A model that merely suppresses knowledge should surrender it when an attacker fine-tunes on the forgotten data. Extending the relearning protocol of the SimNPO paper (Fan et al., 2025b), the suite of Table 19 attacks nine victims, the un-unlearned target model as a

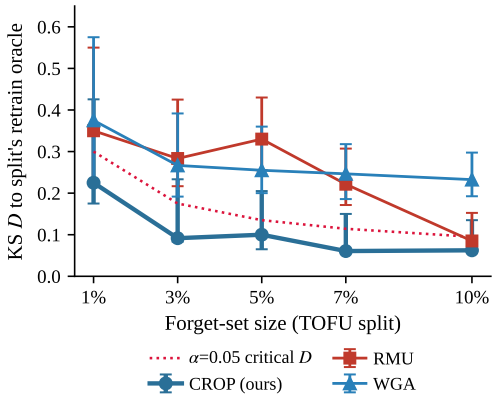


Figure 7: Forget-set-size scaling on TOFU-1B as an effect size. KS D between each method's forget profile and the split's own retrain oracle, with 95% example-bootstrap intervals at every split. The dotted line is the largest D not rejected at $\alpha = 0.05$ for each split's sample size. CROP is the only method at or below it at every size.

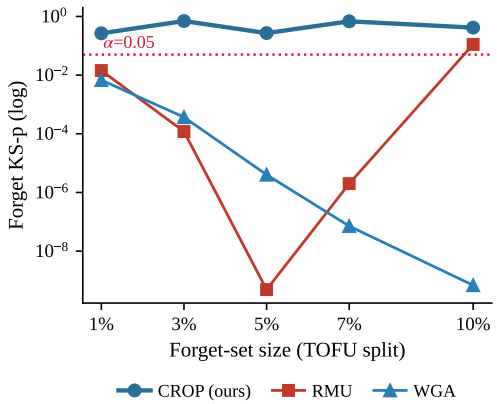


Figure 8: The same scaling experiment as Figure 7 in the benchmark's native p-value form, with 10%-split configurations transferred unchanged to all splits. CROP is the only method above the $\alpha = 0.05$ threshold (dotted) at every split.

ceiling, the retrain oracle, CROP at its published and full-budget operating points, RMU, NPO, PDU, WGA (the last two retrained under the recipes that reproduce their Table 1 rows exactly, Table 12), and NPO+SAM (Fan et al., 2025a), the sharpness-aware NPO variant designed for relearning resilience, trained under NPO's recipe with the official SAM step ported into the framework and its authors' radius $\rho = 0.01$ (Appendix F), under three attack families with a shared 50-step budget (batch 8, evaluated at steps 1, 5, 10, and 50, plus every ten steps at the reference rate). The three families are (i) full fine-tuning on the complete forget10 split at learning rates 5×10^{-6}, 10^{-5}, and 2×10^{-5}; (ii) a LoRA attack training a rank-8 adapter (α_{LoRA}=16, dropout 0, on the attention projections q, k, v, o) at lr 10^{-4}, with the adapter merged into the victim before each evaluation; and (iii) a partial-forget attack fine-tuning on ten of the twenty forget authors only (200 of 400 examples; two epochs, matching the 50-step budget), with recovery measured separately on the attacked and the held-out authors from the per-example evaluation records. Attacking the retrain oracle calibrates every number. An honest learner relearns, so removal means recovering at the oracle's rate rather than failing to recover. The results and their reading are given in Section 6.2, and Figure 10 plots the recovery curves at the reference rate. Three mechanism-level observations bear repeating. PDU, the strongest suppressor in Table 1, rebounds from a forget probability of 0.007 to 0.63–0.87 and its never-attacked authors recover by +0.700, so its near-zero forget scores are a mask rather than a removal; WGA behaves the same way (0.06 → 0.66–0.85, held-out +0.695), and its utility climbs from 0.22 to 0.59 under attack, so the damage its row shows is itself reversible. RMU is more vulnerable to the rank-8 adapter than to full fine-tuning (0.737 vs. 0.71 at step 10-equivalent budgets), consistent with its forgetting living in a low-dimensional activation mask, and its held-out authors recover almost completely when only the other half is attacked (+0.749), while CROP's move less than the attacked oracle's (+0.017 vs. +0.081). NPO+SAM, the one victim built for this attack, relearns marginally slower than plain NPO at the lowest rate (0.42 vs. 0.49 at step 50) and otherwise follows the same trajectory (0.64 vs. 0.67 at the reference rate, 0.79 at 2×10^{-5}, held-out +0.353 vs. +0.382), so flattening the forget loss around the unlearned point delays re-entry by a few steps without changing where the attacker ends up, whereas choosing which directions the update may use does. Under all evaluated protocols both methods lose statistical faithfulness as the attack shifts the profile; the security-relevant axis, knowledge recovered per step of attacker access, is where they separate.

L Distributional Faithfulness

Here we open up the headline metric itself. The forget-quality KS test compares the distribution of per-example truth ratios against the retrained model, so Figure 11 plots the empirical CDFs directly.

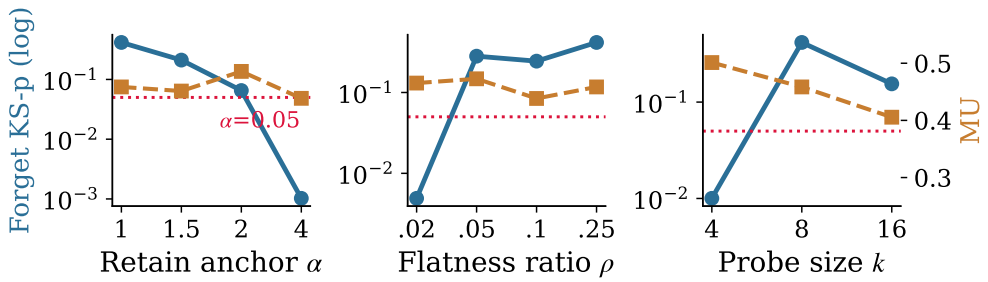


Figure 9: Sensitivity of CROP to its three hyperparameters on TOFU-1B, from sixteen officially evaluated sweep arms. Solid lines show forget KS-p (left axis, log scale), dashed lines show model utility (right axis), and the dotted line marks the $\alpha = 0.05$ threshold. The k=4 point was run at ρ=0.05. The shipped defaults (α=1, ρ=0.25, k=8) sit inside a broad stable basin, and each knob has one identifiable cliff.

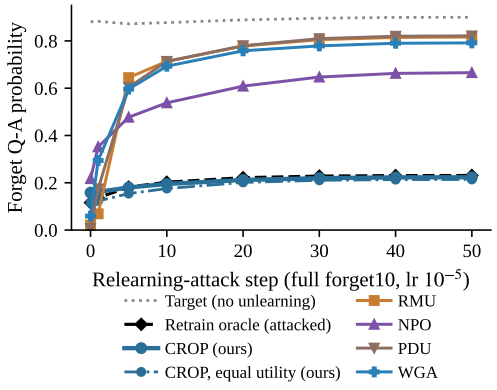


Figure 10: Relearning attack at lr 10^{-5} on the full forget set (one epoch, 50 steps) for the victims of Table 19 other than NPO+SAM, whose curve ($0.28 \rightarrow 0.64$) runs just below NPO's. The attacked retain oracle (dashed) rises from 0.12 to 0.23 and both CROP operating points track it, whereas RMU, NPO, PDU, and WGA rebound toward the target ceiling (dotted) within ten steps.

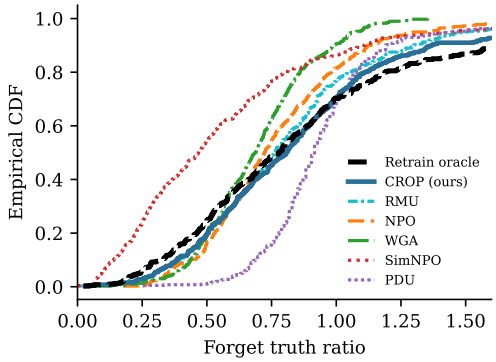


Figure 11: Empirical CDFs of per-example forget truth ratios on TOFU-1B. The KS statistic underlying forget quality is the maximal vertical gap to the retrain oracle (dashed). CROP's curve tracks the oracle across the entire distribution; each baseline deviates in a characteristic direction.

CROP's curve lies on top of the oracle's along its whole length, which is what a KS-p of 0.4158 means geometrically, since no vertical slice separates the two distributions. The baselines' failure modes become diagnosable rather than scalar. SimNPO's mass sits far left of the oracle (truth ratios too low on many examples, reflecting suppression without calibration); PDU's sits far right (its aggressive probability erasure distorts the profile past the oracle); NPO and WGA match the lower half but diverge in the upper half, under-forgetting their hardest examples. The KS test rewards landing on the reference distribution, and penalizes overshoot and undershoot alike.

M Hyperparameter Sensitivity

Figure 9 sweeps the three hyperparameters that govern the mechanism, the retain-anchor weight α, the retain-flatness ratio ρ, and the probe size k, with all other factors fixed. Two properties matter. First, the operating basin is broad. The threshold is crossed at α=1 across $\rho \in \{0.05, 0.10, 0.25\}$, so the recipe is not a knife-edge configuration. Second, each knob has exactly one cliff, and each cliff is diagnostic of the mechanism rather than arbitrary. Raising α from 1 to 2 collapses KS-p from 0.4158 to 0.0652 while utility stays flat (0.4585 vs. 0.4857), so the anchor is not a knob that trades

Table 19: Relearning-attack suite on TOFU-1B, reporting forget-answer probability at attack steps $1 \rightarrow 50$ (batch 8). The attacked retrain oracle is the reference for how fast a model that never saw the data relearns it; resistance means staying at or below this rate. Full FT sweeps the attack learning rate; LoRA attacks train a rank-8 adapter (α=16, attention projections, lr 10^{-4}); the partial-forget attack fine-tunes on authors 0–9 only (same 50-step budget) and reports the forget-probability change on the ten held-out authors (Δ held-out). The equal-utility CROP row is the final checkpoint of the fixed-budget run, whose utility (0.51) matches RMU's (0.53), so that recovery is also compared at equal utility. The target row attacks the model that never unlearned and marks the ceiling every victim recovers toward. NPO+SAM is the sharpness-aware NPO variant built for relearning resilience, trained under NPO's recipe with $\rho = 0.01$.

Victim	Full FT, lr 5×10^{-6}	Full FT, lr 10^{-5}	Full FT, lr 2×10^{-5}	LoRA, lr 10^{-4}	Δ held-out
Target (no unlearning, ceiling)	0.883 → 0.896	0.883 → 0.900	0.747 → 0.870	0.881 → 0.892	−0.009
Retrain oracle	0.122 → 0.173	0.137 → 0.231	0.165 → 0.283	0.117 → 0.205	+0.081
RMU	0.029 → 0.663	0.069 → 0.817	0.399 → 0.859	0.022 → 0.737	+0.749
NPO	0.292 → 0.488	0.352 → 0.666	0.414 → 0.794	0.228 → 0.546	+0.382
PDU	0.035 → 0.626	0.174 → 0.821	0.431 → 0.869	0.009 → 0.699	+0.700
WGA	0.169 → 0.655	0.294 → 0.792	0.438 → 0.848	0.070 → 0.671	+0.695
NPO+SAM	0.230 → 0.417	0.280 → 0.644	0.370 → 0.787	0.213 → 0.499	+0.353
CROP (Ours)	0.159 → 0.179	0.158 → 0.223	0.168 → 0.321	0.158 → 0.187	**+0.017**
CROP, equal utility (Ours)	0.118 → 0.155	0.123 → 0.214	0.137 → 0.325	0.116 → 0.176	+0.038

forgetting against utility, and over-weighting it simply drowns the projected ascent signal without buying anything back. Tightening ρ to 0.02 starves the memorization subspace (KS-p 0.0049) because almost no probed direction passes so strict a flatness test; the intermediate ρ=0.05 scores well (0.2812) but is seed-fragile, collapsing to 1.3×10^{-4} under a seed change, whereas the default ρ=0.25 crosses the threshold on every full-length seed trajectory (peak KS-p 0.47–0.81 over seeds 0–2, Table 2). The probe size shows the expected capacity knee. The k=4 setting cannot forget at all (KS-p 0.01 despite the highest utility in the sweep, 0.5012), while k=16 forgets readily but pays five points of utility (0.4056), consistent with a larger ascent subspace clipping more shared-knowledge directions. Learning rate behaves as a dose, with a doubling to 10^{-4} collapsing utility to 0.2916 at reduced forget quality. In short, the shipped defaults are not tuned to a lucky point; they sit in the interior of a measured basin whose boundaries are all explainable by the mechanism of Section 3.4.

N Curvature-Premise Verification

This appendix verifies the empirical premise of Section 3.2 and the estimators behind it, on the TOFU-1B target model with training-identical settings (bf16, batch 8, the finite-difference scheme of equation 3); estimator reference checks use an fp32 model with exact attention. Table 20 summarizes; Figure 3 plots the per-direction spectrum and Figure 12 the random-direction control.

The signature and its controls. Sixteen deflated power-iteration directions of $\widehat{\boldsymbol{H}}_f$ carry forget curvature λ^f between 261 and 1322 (median 573), all positive, with median $\rho = 0.077$; all sixteen pass the $\rho_{\max} = 0.25$ filter. Sixty-four random unit directions carry essentially no curvature under either loss (λ^f median -0.003) and show no separation (ρ median 1.24), and every basis vector of a random 8-dimensional orthonormal subspace is equally flat. This is in direct correspondence with the random-subspace knockout of Table 3.

Stability. Probes on three disjoint minibatches and a reseeded repeat each keep 8 of 8 directions with the same sharp-for-forget, flat-for-retain signature, but

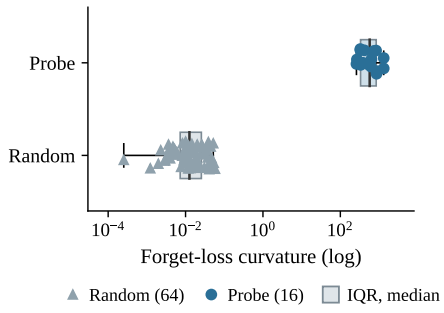


Figure 12: The random-direction control. Forget curvature $|\lambda^f|$ along each of the 16 probe directions (blue, top) versus 64 random unit directions (gray, bottom), shown as points with box summaries on a log axis. Random directions carry essentially no forget curvature, about four orders of magnitude less than probe directions, so the signature of Figure 3 is a property of the extracted directions, not of low-dimensional restriction as such.

Table 20: Curvature-premise verification. The upper block is the TOFU-1B target (Llama-3.2-1B-Instruct, full model); the lower block repeats the premise checks at 7B on the MUSE-News target (Llama-2-7B, the shipped three-layer slab of 607M parameters) and adds the training-level control. Each row asks one question about the probe, states the evidence, and gives the answer. λ^f is the forget-loss curvature along a direction, λ^r the retain-loss curvature, and $\rho = |\lambda^r|/\lambda^f$; a useful direction has large λ^f and small ρ.

Question	Evidence	Answer
TOFU-1B: Llama-3.2-1B-Instruct, full model		
Are the 16 probe directions sharp for the forget loss?	λ^f from 261 to 1322 (median 573), all positive	Yes
Are they flat for the retain loss?	ρ median 0.077; all 16 pass the $\rho_{\max}$=0.25 filter	Yes
Do random directions show the same signature?	64 random unit directions: λ^f median -0.003, ρ median 1.24	No
Does a random 8-dimensional subspace show it?	Every basis vector has $\lambda^f \approx 0$	No
Is the signature stable across minibatches and seeds?	8 of 8 directions keep it on every draw; the subspace rotates (principal cosine 0.06–0.07 across minibatches, 0.85 across seeds on one minibatch)	Signature yes; identity no
Is the finite-difference Hessian-vector product accurate?	Cosine to exact autograd 0.9998 at ϵ=10^{-4}, 0.98 at the shipped 10^{-3}; breaks down at 10^{-2}	Yes, at the shipped step
Does power iteration find the sharpest positive direction?	Lanczos largest-algebraic Ritz value $+14796$ (most negative -1115); power iteration converges to 13507	Yes
MUSE-News at 7B: Llama-2-7B, three-layer slab		
Are the 8 probe directions sharp for the forget loss and flat for the retain loss?	λ^f from 1.6 to 11.3 (median 3.2), all positive; $\lvert\lambda^r\rvert \le 1.4$; ρ median 0.042; 8 of 8 pass the filter	Yes
Do random directions or a random 8-dimensional subspace show the signature?	64 random unit directions: $\lvert\lambda^f\rvert \le 0.003$, ρ median 1.03; every subspace basis vector has $\lvert\lambda^f\rvert \le 0.001$	No
Is the signature stable across minibatches and seeds?	4 to 8 of 8 directions kept on every draw; principal cosine 0.09–0.12 across minibatches, 0.21 across seeds on one minibatch	Signature yes; identity no
Does a random subspace of the same dimension unlearn?	Full three-leg recipe at the shipped slab with random directions in place of the probe. Best checkpoint: VerbMem 0.32, KnowMem-f 0.44, PrivLeak -99.9, distance to retrain 0.23. The curvature probe reaches 0.18, 0.42, -23.8, and 0.03.	No

the extracted subspaces rotate, with principal cosines
between subspaces from different minibatches having
median 0.06–0.07 (for the same minibatch under a different seed, the top cosine is 0.85). The invariant is the existence and signature of the sharp subspace, not its identity, which is the empirical basis for the deforming-family framing of Section 3.2 and for refreshing the probe online rather than computing it once.

Estimator checks. Against exact double-backpropagation HVPs (fp32, exact attention), the finite-difference estimator at $\epsilon = 10^{-4}$ agrees to cosine 0.9998 on probe directions; at the shipped $\epsilon = 10^{-3}$ the cosine is 0.98 with Rayleigh-quotient errors of tens of percent that largely cancel in the ratio ρ; at $\epsilon = 10^{-2}$ the approximation fails on sharp directions. Finally, because deep-network Hessians are indefinite, power iteration could in principle converge to a large-magnitude negative eigenvalue. A 20-step Lanczos run with full reorthogonalization shows that on these forget Hessians the largest-magnitude Ritz value is the largest-algebraic one ($+14796$, against a most-negative value of -1115), and power iteration lands accordingly (13507); the $\lambda^f > 0$ condition of equation 6 guards any residual case.

Generality at 7B. The lower block of Table 20 repeats the spectrum, the random controls, and the stability check on the Llama-2-7B MUSE-News target within the shipped three-layer slab, with the

same bf16 finite-difference estimator and the training batch size, and adds the training-level control. Eight deflated power-iteration directions carry forget curvature between 1.6 and 11.3 (median 3.2), all positive, with retain curvature at most 1.4 and median $\rho = 0.042$; all eight pass the $\rho_{\max} = 0.25$ filter. The absolute curvatures are two orders of magnitude below the 1B values because this target is far from saturated (forget NLL 0.5) and the slab is a tenth of the model, but the filter acts on the ratio, and sixty-four random unit directions carry forget curvature of at most 0.003 in magnitude with ρ median 1.03, so the extracted directions exceed random ones by more than three orders of magnitude here as well. Across three minibatch draws and a reseeded repeat, four to eight of eight directions keep the signature while the subspaces from different minibatches are nearly orthogonal (principal cosine 0.09–0.12); a reseeded probe on the same minibatch overlaps its original only weakly (top cosine 0.21, against 0.85 at 1B), so at this scale the sharp subspace is a noisier object even though its signature is not. The training-level control settles whether that curvature does the forgetting. Replacing the probe with random orthonormal directions of the same dimension and rerunning the full three-leg Llama-2-7B recipe at the shipped slab, with checkpoint selection by the same closest-to-retrain rule, leaves the best checkpoint at VerbMem 0.322 and KnowMem-f 0.438 with PrivLeak still at the target's -99.9 floor, a distance to the retrained reference of 0.231 against 0.029 for the curvature probe at the same slab, which is closer to the un-unlearned target (0.295) than to the reference. The fp32 estimator checks of the upper block were not repeated at 7B.

O LIMITATIONS

CROP operates at the deep end of the forgetting–utility frontier. On TOFU-1B its published checkpoint has MU 0.46 against the oracle's 0.59 (held-out mean 0.415), and it overlaps RMU's utility range on only one of three seeds (Appendix C), so recovering utility without leaving the retrain-equivalence band is the main open problem. Our faithfulness criterion is defined on the forget profile (Section 3.1), so this retain-side gap is reported as a cost rather than folded into the faithfulness score, and a reader who weights retain fidelity equally should read CROP's TOFU-1B and MUSE-Llama rows as trading utility for forget-profile proximity. The analysis upper bound consults the retrained model; the oracle-free rule removes that dependence but relies on a calibration whose error is 0.011 on Llama and 0.034 on Qwen, where the undershoot places the pick one checkpoint short of the passing window on one of three seeds (Table 10). On WMDP the suppression extends to held-out chemistry and adjacent benign categories (Appendix I), and at 7B the probe's memory footprint confines CROP to a layer slab whose position matters on Llama-2 (Appendix B). Finally, the shipped update composes two projectors, which leaks a small component outside the selected subspace, and the per-direction flatness filter does not bound every step's retain curvature; Appendix A measures both and gives the exact projector that removes the first. Three further caveats are stated where they arise: on MUSE-Llama the delivered checkpoint gives up retain knowledge (KnowMem-r 0.38 against the retrained model's 0.56; Appendix H), the MUSE and WMDP rows are selected by the benchmark metrics rather than under a leakage-free split (Appendix F), and the curvature probe costs 1.8–$2.5\times$ a first-order step (Table 13).