DARC: DENOISED AUTHORIZED REPARAMETERIZED CODING FOR LLM UNLEARNING VIA FIRST-ORDER NOISE CANCELLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM unlearning faces a disclosure conflict when a server keeps its model parameters confidential and a client withholds raw forget data from that server. Noisy model publication can limit disclosure, but it also changes the unlearning update. We propose DARC (Denoised Authorized Reparameterized Coding), coupling model publication with authorized update recovery through executors granted temporary access to client data. The server releases m model copies using correlated Gaussian perturbations and function-preserving coordinate transformations, under an isolation assumption limiting each coalition to $t < m$ same-round copies. After aligning all m returned updates, inverse-scale aggregation cancels the common first-order perturbation response of a shared locally smooth update map. We prove the decoder's uniqueness among normalized linear rules with universal first-order cancellation and analyze curvature, optimizer, and stochastic errors. Under a fixed public positive-definite noise covariance and independent key/code sampling, we derive exact neighboring-checkpoint Rényi divergence for key-revealed views and extend checkpoint-privacy bounds to complete executor transcripts under the stated auxiliary-release conditions. Our optimizer analysis identifies exact-arithmetic Adam-family compatibility conditions and quantifies discrepancies from continuous attention rotations. Experiments on Llama and Qwen series models with TOFU and MUSE benchmarks assess forgetting, retained utility, and update fidelity. For Llama-1B on TOFU, the six non-collapsed trainers have utility scores within 0.003 of their noise-free references. Targeted diagnostics reveal numerical compatibility gaps and closely matching outputs from individual published views on the tested prompts. DARC connects bounded checkpoint disclosure with optimizer-aware approximation of a chosen unlearning method's noise-free update.

1 INTRODUCTION

Machine unlearning aims to approximate retraining after removing designated data (Cao & Yang, 2015; Ginart et al., 2019; Bourtoule et al., 2021). Practical LLM methods update parameters through objectives targeting forgetting while preserving other capabilities (Jang et al., 2023; Zhang et al., 2024; Fan et al., 2024; Dong et al., 2025; Shen et al., 2025). Their execution needs both the model parameters, or *checkpoint*, and the records designated for forgetting.

This access becomes difficult when the model and data have different owners. We consider a server holding a confidential checkpoint and a client withholding raw forget data from that server. The client needs to authorize third-party *executors* to process its data temporarily, while the server limits the model information they receive. These executors follow the protocol but may combine their observations. The challenge is to control checkpoint disclosure while obtaining an update faithful to the chosen unlearning procedure.

We develop optimization techniques for the problem. Adding noise to a checkpoint changes the gradients and subsequent training trajectory, so removing the initial perturbation is different from recovering the update that training would have produced without it. Even a change of parameter coordinates that preserves predictions can alter the behavior of an adaptive optimizer such as Adam (Kingma & Ba, 2014; Zhang et al., 2026; Xie et al., 2025). Protecting the published parameters

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.

and preserving the unlearning update must therefore be analyzed together. We should explore *how noisy model publication can limit checkpoint disclosure while allowing an authorized server to approximate the chosen trainer's noise-free update.*

We propose DARC (*Denoised Authorized Reparameterized Coding*), a protocol coupling model publication with update aggregation. The server creates m copies, called model views, with correlated Gaussian noise and function-preserving coordinate transformations, giving one to each executor. Enforced isolation limits every coalition to at most $t < m$ same-round copies. The server maps all m returned updates back to the original coordinates and weights them inversely to their noise scales. The weighted perturbations sum to zero, canceling their common first-order response under a shared, locally smooth update map. We call this *authorized denoising* and characterize the remaining curvature, optimizer, and stochastic errors.

Contributions.

1. **Checkpoint privacy under bounded coalition access.** For fixed public positive-definite noise covariance and coordinate transformations independent of the checkpoint and noise, we derive exact neighboring-checkpoint Rényi divergence even with revealed transformations. We extend the bound to coalition observations when auxiliary inputs are derived from those views or separately protected, and analyze worst-case coalitions and conditional multi-round composition (Theorem 1).
2. **First-order cancellation with explicit residual control.** Among normalized linear decoders, we prove that inverse-scale weighting uniquely cancels the common first-order response for every zero-sum perturbation collection. Our analysis bounds curvature and optimizer errors and separates mean-update cancellation from variation across independent training runs (Theorem 2; Proposition 1).
3. **Optimizer-aware Transformer reparameterization.** We establish exact-arithmetic Adam-family compatibility for signed-permutation transformations with consistently transformed objectives, auxiliary state, parameter groups, and randomness. For continuous attention rotations, we derive direction and trajectory error bounds. We instantiate the transformations for SwiGLU, grouped-query attention, rotary position embeddings, and LoRA (Theorem 3).

Experiments on Llama and Qwen with TOFU and MUSE examine forgetting, retained utility, and update fidelity. In the Llama-1B TOFU comparison, the six non-collapsed trainers have utility scores within 0.003 of their noise-free references. Targeted diagnostics also reveal numerical compatibility gaps and closely matching outputs from published views on bounded prompt suites. Checkpoint privacy is quantified for executor observations. This links disclosure accounting with the accuracy of authorized update recovery.

2 Related Work

LLM unlearning. Approximate methods include gradient ascent/difference (Jang et al., 2023; Liu et al., 2022a), NPO and SimNPO (Zhang et al., 2024; Fan et al., 2024), preference or substitute-response training (Mekala et al., 2025), logit shaping (Dong et al., 2025; Wang et al., 2025), activation redirection (Shen et al., 2025), and multi-objective updates (Jin et al., 2025; Pan et al., 2025). TOFU and MUSE evaluate behavioral forgetting, leakage, and retained utility (Maini et al., 2024; Shi et al., 2025). DARC retains the existing forgetting objective and controls publication, execution, and aggregation of the parameter updates.

Privacy and confidential collaboration. Differential privacy requires an explicit neighboring relation and calibrated sensitivity (Dwork et al., 2006b); RDP accounts for Gaussian mechanisms and composition (Mironov, 2017). Distributed noise and secure aggregation address bounded collusion or hidden sums (Dwork et al., 2006a; Bonawitz et al., 2017). Federated unlearning keeps data local but usually publishes a global model and returns updates (Liu et al., 2020; 2022b; Nguyen et al., 2024), which may reveal training text or images (Zhu et al., 2019; Geiping et al., 2020; Petrov et al., 2024). Offsite-Tuning, secret sharing, MPC, and restricted forward computation impose different trust and computation costs (Xiao et al., 2023; You et al., 2026; Luo et al., 2026). Our theorem concerns only the server-to-executor checkpoint transcript under a view threshold; update privacy remains separate.

Parameter symmetries and optimizer dependence. Hidden-unit permutations and changes to the attention basis preserve network functions while changing parameter coordinates (Kůrková & Kainen, 1994; Entezari et al., 2021; Ainsworth et al., 2022). Pairwise assignment and Procrustes alignment are tractable, while unrestricted multiway alignment problems can be hard (Burkard et al., 2012; Ling, 2021); these complexity results do not prove hardness for the specific key family here and are not a privacy proof. Coordinatewise adaptive optimizers are generally basis-dependent, whereas their signed-permutation symmetry is much larger than their generic rotation symmetry (Kingma & Ba, 2014; Loshchilov & Hutter, 2017; Zhang et al., 2026; Xie et al., 2025). Recent work formulates optimizer–symmetry compatibility for Transformer parameter blocks and reports that Adam can separate gauge-equivalent Transformer initializations from the first step (Lau & Su, 2026; Singh, 2026). We specialize this distinction to the exact keys used by DARC and make the resulting mismatch an explicit decoder error term.

3 Setting and Threat Model

The model holder (server) holds $\theta_{r-1} \in \mathbb{R}^d$ in round r; the data holder (client) holds the forget set $\mathcal{D}_\mathrm{f}$. The client authorizes executors $E_1, \ldots, E_m$ to access and temporarily store $\mathcal{D}_\mathrm{f}$ for unlearning. Local denotes this authorized processing outside the server; raw records go only to these executors (Table 1).

Table 1: **Protocol information access** for $|S| \le t < m$.

	Server	Client	Executor E_k	Coalition S
Canonical model	Holds	Not sent	Not sent	Not sent
Raw forget data $\mathcal{D}_\mathrm{f}$	Not sent	Holds	Temporary use	Members' use
Coded views	Generates all	Not required	One assigned	Up to t pooled
Local computation	Aggregates	Authorizes	Unlearns	Members unlearn
Returned updates	Receives all	Not required	Produces own	Members' updates

A trainer $\mathcal{A}$ returns the displacement

$$\Delta_{\mathcal{A}}(\vartheta; \mathcal{D}_\mathrm{f}, \xi) := \mathcal{A}(\vartheta; \mathcal{D}_\mathrm{f}, \xi) - \vartheta, \tag{1}$$

where ξ contains minibatch order and other algorithmic randomness. Each executor receives one directly assigned same-round coded view and returns its update to the server.

Coalition and authorization semantics. For $S \subset [m]$, $\mathsf{View}_{S,r}$ contains every tensor, optimizer trace, prediction, and output available to executors in S, plus arbitrary public information. Executors in S may pool their observations, but isolation limits their combined access to the coded views assigned to S, including access through relays or persistent storage. Each round's coalition satisfies $|S_r| \le t < m$ and is fixed before fresh coding, with selection based only on prior observed history. Multi-round composition additionally requires the conditional-release assumptions in Appendix B.4. The parameter t limits pooled observations, whereas authorized recovery uses an m-of-m zero-noise direction.

Closed publication transcript. The theorem covers the *complete* checkpoint-dependent executor input. Every private reference model, teacher, adapter base, optimizer state, mask, or calibration statistic must either be public and checkpoint-independent, be generated as post-processing of the executor's coded view, or be included in a jointly accounted release. Server feedback derived from all copies, including a final canonical checkpoint, is an additional release unless it satisfies this same condition. The auxiliary-state audit supplies source and local-reference checks (Appendix J.1).

Protected object, profiles, and adversaries. The direct neighboring input is the server checkpoint under a declared whitened metric. A record-level server-data statement additionally requires a transcript-uniform conditional sensitivity certificate. The local forget set is held fixed in this theorem. The clean formal mode uses a fixed public covariance profile. We assume executors are honest-but-curious and may attempt checkpoint or functional recovery, and the server is honest-but-curious toward $\mathcal{D}_\mathrm{f}$ and may inspect all returned updates.

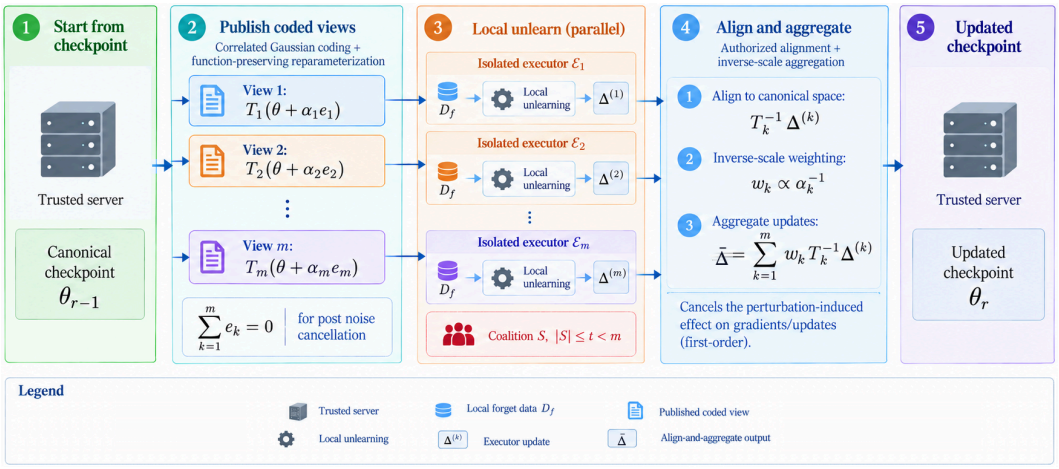


Figure 1: **DARC from publication to checkpoint update.** With $\theta = \theta_{r-1}$, all m returned displacements are aligned and aggregated using normalized inverse-scale weights $w_k = \alpha_k^{-1} / \sum_j \alpha_j^{-1}$ (Eq. (6)). The common first-order response cancels under Theorem 2; residual errors are shown in Figure 2. View isolation bounds pooled executor access as in Section 3; returned updates remain server-visible.

4 DARC: Coded Publication and Authorized Denoising

4.1 Correlated Gaussian publication code

Partition all protected checkpoint coordinates into L blocks and choose a positive-definite profile

$$\Sigma_r = \mathrm{blkdiag}(b_{r,1}^2 I_{d_1}, \dots, b_{r,L}^2 I_{d_L}), \qquad b_{r,\ell} \ge b_{\min} > 0. \tag{2}$$

For ordinary RDP, the profile is fixed independently of the protected checkpoint. It uses $b_{r,\ell} = \mathrm{RMS}(\theta_{r-1,\ell} - \theta_{\mathrm{ref},\ell})$. Declassification and a positive floor are analytical protocol choices. Protection covers all privately varying published tensors.

Draw $z_k^{(r)} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \kappa^2 \Sigma_r)$, let $\bar{z}^{(r)} = m^{-1} \sum_j z_j^{(r)}$, and set

$$e_k^{(r)} = \sqrt{\frac{m}{m-1}} \left(z_k^{(r)} - \bar{z}^{(r)} \right), \qquad \sum_{k=1}^{m} e_k^{(r)} = 0. \tag{3}$$

Every marginal is $\mathcal{N}(0, \kappa^2 \Sigma_r)$. Copy k uses $\epsilon_k^{(r)} = \alpha_k e_k^{(r)}$ with $\alpha_k > 0$; the supplied schedule is $\alpha_k = 1 + (k-1)/(m-1)$. The full copy covariance is singular, but every strict principal submatrix is positive definite.

4.2 Function-preserving coordinate keys

For every copy, the server samples an orthogonal parameter symmetry $T_{k,r}$ satisfying

$$f_{T_{k,r}\theta}(x) = f_\theta(x) \quad \forall x, \qquad T_{k,r}^{-1} = T_{k,r}^\top. \tag{4}$$

The optimizer-safe family $\mathcal{G}_{\mathrm{safe}}$ contains MLP/SwiGLU hidden-channel permutations and signed-permutation attention keys, including RoPE quarter turns. As global maps on the vectorized parameters, these preserve the abstract coordinate-wise optimizer dynamics under the conditions of Theorem 3. The richer family $\mathcal{G}_{\mathrm{rot}}$ contains continuous attention head-space rotations compatible with grouped-query attention and the RoPE commutant. It preserves the network function and SGD dynamics but is generally not Adam-equivariant.

The published view is

$$\theta_{\mathrm{pub}}^{(k,r)} = T_{k,r} \left(\theta_{r-1} + \alpha_k e_k^{(r)} \right). \tag{5}$$

Given public information and prior observed history, the key vector is independent of the checkpoint and current Gaussian code (Appendix B.1). Key revelation gives a conservative privacy bound; functional recovery is tested separately in Appendix J.9. Transformer/LoRA formulas appear in Appendix D.

4.3 Local execution and decoding

Executor E_k runs the chosen local trainer and returns $\Delta^{(k,r)} = \Delta_{\mathcal{A}}(\theta_{\mathrm{pub}}^{(k,r)}; \mathcal{D}_{\mathrm{f}}, \xi_{k,r})$. A private frozen reference required by the objective must be derived from the same coded view or be jointly protected. Likewise, a fresh zero optimizer state is automatically compatible with safe keys, while transferred first moments, second moments, masks, and parameter groups must be transformed consistently.

The server canonicalizes and aggregates

$$\widehat{\Delta}^{(k,r)} = T_{k,r}^{-1}\Delta^{(k,r)}, \qquad w_k = \frac{\alpha_k^{-1}}{\sum_j \alpha_j^{-1}}, \qquad \theta_r = \theta_{r-1} + \eta_{\mathrm{srv}} \sum_k w_k \widehat{\Delta}^{(k,r)}. \tag{6}$$

This weighted arithmetic average uses the inverse-scale normalizer, motivating the term *harmonic weights*. Proposition 1 proves their uniqueness for universal first-order cancellation. A two-pass streaming server uses $O(d)$ auxiliary storage, independent of m, with isolation enforced separately. Algorithm 1 states the round; executors use data under client authorization and return displacements to the server.

Algorithm 1 One optimizer-aware DARC round

Require: Server model θ_{r-1}, isolated executors with local access to $\mathcal{D}_{\mathrm{f}}$, declared profile mode, $\{\alpha_k\}$, κ, η_{srv}, optimizer-compatible key family

Sample zero-sum $\{e_k^{(r)}\}$ and keys as in Section 4.2; set $S_0 \leftarrow 0$, $S_1 \leftarrow 0$
for $k = 1, \ldots, m$ **do**
  Send $T_{k,r}(\theta_{r-1} + \alpha_k e_k^{(r)})$ only to E_k
  Construct every private auxiliary object according to Assumption 1
  E_k runs local unlearning with fresh or correctly transformed state and returns $\Delta^{(k,r)}$
  $\widehat{\Delta}^{(k,r)} \leftarrow T_{k,r}^{-1}\Delta^{(k,r)}$
  $S_0 \leftarrow S_0 + \alpha_k^{-1}$, $S_1 \leftarrow S_1 + \alpha_k^{-1}\widehat{\Delta}^{(k,r)}$
end for
$\theta_r \leftarrow \theta_{r-1} + \eta_{\mathrm{srv}} S_1/S_0$
return θ_r

5 Privacy and Optimizer-Aware Decoding

5.1 Privacy of incomplete coalitions

For a fixed public profile $\Sigma_r \succ 0$, set $d_r(\theta, \theta') = \left\|\Sigma_r^{-1/2}(\theta - \theta')\right\|_2$ and declare $\theta \simeq_r \theta'$ when $d_r \le \Delta_r$. For a coalition S chosen before the current code is drawn, write $s = |S| < m$, $B_S = \frac{m}{m-1} I_s - \frac{1}{m-1}\mathbf{1}\mathbf{1}^\top$, and $D_S = \mathrm{diag}(\alpha_k : k \in S)$.

Theorem 1 (Exact coalition divergence)**.** *For $\kappa > 0$ and keys sampled as in Section 4.2, the canonicalized views Y_S satisfy*

$$D_\lambda(Y_S(\theta) \| Y_S(\theta')) = \frac{\lambda}{2\kappa^2}\gamma(S) d_r(\theta, \theta')^2, \tag{7}$$

$$\gamma(S) = \frac{m-1}{m}\sum_{k\in S}\alpha_k^{-2} + \frac{m-1}{m(m-s)}\left(\sum_{k\in S}\alpha_k^{-1}\right)^2. \tag{8}$$

Key revelation preserves this divergence. Under publication closure (Section 3) with local data and other non-server inputs fixed, the complete single-round executor transcript is $(\lambda, \lambda\Delta_r^2\gamma(S)/(2\kappa^2))$-RDP.

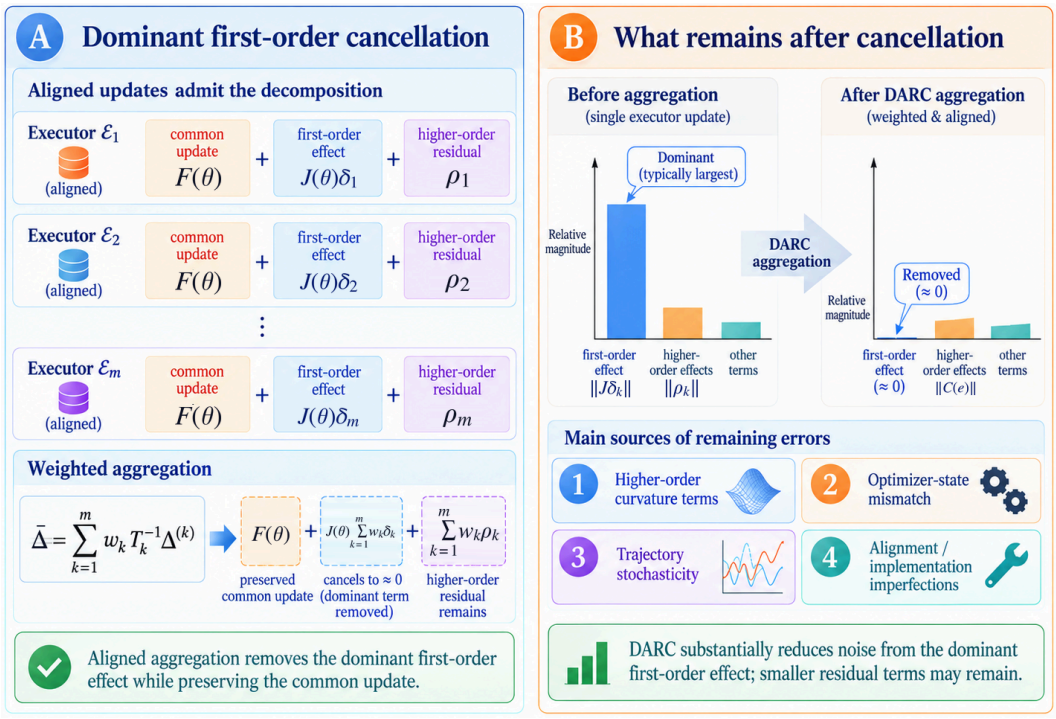


Figure 2: **First-order cancellation and residual errors.** A shows the shared-map, locally smooth, compatible/coupled case of Theorems 2 and 3. Here $\delta_k = \alpha_k e_k$, $J(\theta) = J_F(\theta)$ (abbreviated J in B), and ρ_k is the curvature-only Taylor remainder (Eqs. (49)–(50); Appendix C.2). Normalized weights cancel the common linear response, leaving $C(e) = \sum_k w_k \rho_k$. B's bars are schematic: "dominant" denotes local Taylor leading order. Figure 3 and Appendix J.6 report total-update and optimizer/state discrepancies; component magnitudes are unmeasured.

The proof in Appendix B uses the equal-covariance Gaussian divergence and B_S^{-1}. For equal unit scales, $\gamma_t = t(m-1)/(m-t)$; the worst heterogeneous coalition contains the t smallest α_k. All m aligned views recover θ exactly, so the key-revealed full bundle has infinite divergence for distinct checkpoints. An unnoised private coordinate or a deterministically disclosed changing profile can likewise invalidate ordinary RDP. The supplied task-vector profile supports only a common-profile-stratum statement.

Conditional composition. Across rounds, additive RDP requires a bound on each release *conditional on the coalition's actual observed history*, including the effect of hidden all-copy updates. Appendix B.4 gives sufficient deterministic-state and conditional-coupling conditions. Final canonical checkpoints and other all-copy feedback require separate accounting.

5.2 First-order cancellation and residual error

Fix local data and one shared local-randomness realization. Let $F(x)$ be the *complete* canonical displacement map, including the prescribed construction of frozen references, adapter initialization, and optimizer state. Define

$$R_T(x) = T^{-1}\Delta_{\mathcal{A}}(Tx; \mathcal{D}_{\mathrm{f}}, \tau_T \xi) - F(x), \tag{9}$$

where auxiliary objects and coordinate-indexed randomness are transformed consistently. A different private-reference rule defines a different F.

Theorem 2 (Authorized decoding on a smooth neighborhood)**.** *Suppose J_F is L_J-Lipschitz on a convex neighborhood $\mathcal{N}$ of θ. On every realization whose segments $[\theta, \theta + \alpha_k e_k]$ lie in $\mathcal{N}$, inverse-*

scale decoding satisfies

$$\left\|\bar{\Delta} - F(\theta)\right\|_2 \le \frac{L_J}{2} \sum_k w_k \alpha_k^2 \left\|e_k\right\|_2^2 + \sum_k w_k \left\|R_{T_k}(\theta + \alpha_k e_k)\right\|_2 . \tag{10}$$

If the same remainder bound holds globally, the expected curvature envelope is $\frac{L_J \kappa^2 \operatorname{tr}(\Sigma)}{2} (\sum_k \alpha_k)/(\sum_k \alpha_k^{-1})$. *For local smoothness only, the Gaussian neighborhood-escape probability and the extra tail-moment term are required (Appendix C.4).*

The common linear term cancels because $w_k \alpha_k$ is constant and $\sum_k e_k = 0$. Gaussian quadratic-form moments control the nonnegative error *envelope*; the actual nonlinear residual has a separate variance. A quadratic-map example retains a common quadratic bias as copy count increases (Appendix C.7).

Proposition 1 (Unique normalized linear decoder)**.** *If* $\sum_k w_k = 1$ *and* $\sum_k w_k \alpha_k e_k = 0$ *for every zero-sum collection, then* $w_k = \alpha_k^{-1} / \sum_j \alpha_j^{-1}$.

With independent local seeds, exact cancellation applies to the *mean update map*, while realized Jacobians vary. Appendix C.6 separates curvature bias from conditional training variance and derives the additional $O(\kappa^2)$ Jacobian-variation term. For a fixed RDP order budget, equal copy scales minimize the derived worst-coalition privacy–curvature factor

$$\Gamma_t(\alpha) \frac{\sum_k \alpha_k}{\sum_k \alpha_k^{-1}}, \qquad \Gamma_t(\alpha) = \max_{|S| \le t} \gamma(S), \tag{11}$$

The optimization concerns this bound rather than the measured error.

5.3 Optimizer compatibility and Adam error bounds

For the following theorem, assume a symmetry-covariant local objective and auxiliary-state construction, preserved parameter groups, and scalar hyperparameters, compatible clipping, and zero or correctly mapped optimizer state. A couple of minibatches and every coordinate-indexed random variable under the symmetry action.

Theorem 3 (Exact optimizer compatibility)**.** *Under the preceding compatibility conditions and in exact arithmetic, SGD, momentum, and Nesterov are pathwise equivariant under orthogonal keys; AdaGrad, RMSProp, Adam, AdamW, and AMSGrad are pathwise equivariant under signed permutations.*

MLP/SwiGLU channel permutations and RoPE quarter turns satisfy the signed-permutation condition, so $R_T \equiv 0$ in exact arithmetic under the stated compatibility conditions. For signed permutations, Adam's first and second moments transform as $m' = Tm$ and $v' = |T|v$, respectively, so its coordinate-wise operations commute with the key.

With uncoupled but symmetry-invariant randomness, and the other compatibility conditions unchanged, equivariance holds in distribution: the mean symmetry bias is zero, but the realized training variance remains (Proposition 6).

Dense rotations require a different analysis. For a fixed common gradient history, let $\widehat{m}_h$ be the corrected first moment and $\widehat{C}_h$ the full second-moment matrix whose diagonal Adam stores. With $A_\varepsilon(C) = \operatorname{Diag}((\sqrt{\operatorname{diag} C} + \varepsilon)^{-1})$, the exact normalized-direction defect is

$$q_h(T) = \left[T^\top A_\varepsilon(T \widehat{C}_h T^\top) T - A_\varepsilon(\widehat{C}_h)\right] \widehat{m}_h. \tag{12}$$

For block b, write $\widehat{C}_b = \mu_b I + A_b$. If every canonical and rotated diagonal is at least $\nu_b > 0$, then

$$\|q_h(T)\|_2^2 \le 4 \sum_b L_{\varepsilon,b}^2 \left\|A_b\right\|_{\mathrm{op}}^2 \left\|\widehat{m}_{h,b}\right\|_2^2, \quad L_{\varepsilon,b} = \frac{1}{2\sqrt{\nu_b}(\sqrt{\nu_b} + \varepsilon)^2}. \tag{13}$$

The mismatch depends on anisotropy and second-moment denominators as well as noise scale. Appendix E gives random-rotation moments, a floor-free bound, a nonzero one-step example, and lifted-state trajectory bounds. Applying fixed-history bounds along a key-dependent trajectory additionally requires uniform stability control. Signed-permutation keys avoid the structural rotation defect. Section 6.4 evaluates finite-precision agreement and reports unresolved tolerance exceedances and non-informative numerical bounds.

Table 2: **Noisy publication versus denoising on TOFU Forget01.** Llama-1B/3B, seven trainers, $m = 2$, $\kappa = .01$, seed 0. Bold compares only DARC and Noised within each trainer, model and metric; ties are included. Pale-blue Clean rows are unranked noise-free references. FQ/FTR/MU are ranked upward; reported PL (PrivLeak) is ranked by distance to zero. Highlighting follows the descriptive benchmark convention.

		Llama-3.2-1B				Llama-3.2-3B			
Trainer	Setting	FQ	FTR	MU	PL	FQ	FTR	MU	PL
GradAscent	Noised	2.81e-8	0.246	0.000	**58.9**	**5.91e-14**	**0.157**	**7.71e-5**	53.0
	DARC	**5.41e-2**	**0.468**	**2.31e-4**	69.6	4.47e-16	0.142	2.28e-5	**45.9**
	Clean (ref.)	6.58e-5	0.355	0.000	65.8	3.06e-11	0.187	8.23e-5	50.6
GradDiff	Noised	0.266	0.533	0.461	**73.3**	2.86e-2	0.476	0.384	**106.0**
	DARC	**0.405**	**0.547**	**0.464**	77.2	**5.41e-2**	**0.485**	**0.409**	107.0
	Clean (ref.)	0.405	0.535	0.461	77.1	5.41e-2	0.504	0.419	108.0
DPO	Noised	0.165	0.620	**0.595**	**-19.8**	**0.766**	**0.691**	**0.634**	-15.3
	DARC	**0.266**	**0.641**	0.591	-28.9	0.579	0.683	0.633	**-3.95**
	Clean (ref.)	0.165	0.637	0.591	-25.5	0.579	0.685	0.634	-10.2
NPO	Noised	0.766	**0.640**	**0.600**	32.9	**0.990**	**0.636**	**0.653**	62.1
	DARC	**0.919**	0.628	0.597	**28.2**	**0.990**	0.634	0.651	**57.5**
	Clean (ref.)	0.919	0.624	0.599	30.6	0.990	0.635	0.650	61.2
SimNPO	Noised	5.41e-2	0.522	0.592	**-70.2**	**9.71e-2**	0.550	**0.662**	-72.3
	DARC	**9.71e-2**	**0.525**	**0.598**	-71.8	**9.71e-2**	**0.554**	0.656	**-64.3**
	Clean (ref.)	5.41e-2	0.526	0.598	-68.4	9.71e-2	0.550	0.658	-62.4
UnDIAL	Noised	**1.43e-2**	0.527	0.614	**-77.4**	**5.41e-2**	**0.566**	0.692	**-71.8**
	DARC	**1.43e-2**	**0.529**	**0.615**	-78.0	2.86e-2	0.563	**0.693**	-73.3
	Clean (ref.)	1.43e-2	0.530	0.613	-76.4	2.86e-2	0.567	0.693	-72.0
SatImp	Noised	**6.76e-3**	0.470	0.597	-99.1	1.43e-2	0.507	**0.661**	-100.0
	DARC	**6.76e-3**	**0.476**	**0.601**	**-98.9**	**2.86e-2**	**0.512**	0.659	**-99.9**
	Clean (ref.)	3.02e-3	0.474	0.600	-98.9	1.43e-2	0.510	0.659	-99.7

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

We compare DARC with SINGLE-NOISED, which trains from a perturbed initialization without authorized decoding. DIRECT is the same trainer on the noise-free checkpoint and serves as the reference. Experiments cover seven trainers, Llama/Qwen models, TOFU Forget01/05/10, and MUSE Books/News (Grattafiori et al., 2024; Yang et al., 2024; Hu et al., 2021; Maini et al., 2024; Shi et al., 2025); complete results are in Appendix H.

Additional Llama-3.2-1B/TOFU experiments use paired 40-step runs and separately specified evaluations (Appendix J). GPU workloads use one NVIDIA RTX PRO 6000 with 96 GB memory. FQ is a descriptive nominal KS p-value (Appendix H.2.2). MU measures retained utility. The experiments assess update fidelity and behavior.

6.2 RECOVERY TOWARD A NOISE-FREE REFERENCE

For Llama-1B (Table 2), DARC matches or exceeds Noised FQ for all seven trainers: five increases and two ties. The six non-collapsed objectives have MU within .003 of Clean; GradAscent remains collapsed. For 3B, GradDiff MU increases from .384 to .409, toward Clean's .419; the other five non-collapsed trainers stay within .002 of Clean MU. DPO and UnDIAL have lower FQ than Noised, and leakage changes are mixed. The one-copy control matches mean copy scale but differs in perturbation variance and computation, so the comparison combines these effects with decoding.

6.3 Noise level, copy count, and repeated runs

Copy count and round allocation have trainer-dependent effects (Figure 7; Appendix I). In three paired GradDiff seeds, $\kappa = .01$ gives more consistent utility preservation than $\kappa = .10$, whose mean paired change is negative and includes a zero-MU endpoint. Local-reference NPO has a small mean change with a posthoc interval containing zero (Appendix Table 35). These comparisons are confined to three pairs of seeds. Public-base scaling raises GradDiff MU while increasing clean-relative update error; profile shape and energy vary together (Appendix J.8). Thus, update fidelity and retained utility provide distinct evaluation axes.

6.4 Decoder and optimizer mechanisms

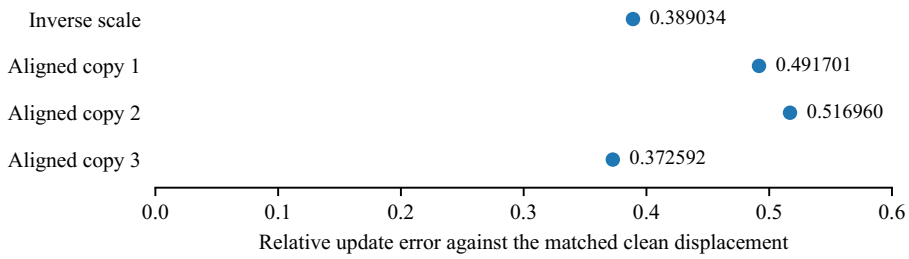


Figure 3: **Total error of saved aligned updates.** GradDiff on Llama-1B/TOFU, one fixed 40-pair traversal, $m = 3$, $\alpha = (1, 1.5, 2)$ and $\kappa = .01$. Four estimators share the saved updates and matched clean displacement (Appendix Table 41), with error normalized by the latter's norm. Each individual copy displacement is applied to the clean checkpoint. The inverse-scale error is close to, and slightly above, the lowest individual-copy error.

The inverse-scale error is .3890, compared with .4917, .5170 and .3726 for the aligned copies (Figure 3). Aggregation improves on two copies and approaches the lowest error. These are total displacement errors. The shared, extremely small GradDiff FQ values also limit the forgetting performance of this protocol (Appendix J.5).

Paired transformed/canonical runs share a noisy initialization. Continuous-rotation AdamW has a mean final relative discrepancy 0.687, versus 0.0538 for FFN-permutation AdamW. Even theoretically compatible pairs exceed numerical tolerances: FFN/AdamW in two copies and FFN/RMSProp in all three at the $.05$ endpoint threshold. Their causes remain unresolved. Fixed-history rotation diagnostics quantify conditional direction defects, with bounds too loose to certify negligible error (Appendix J.6). At a fixed student, reference replacement changes GradDiff/SimNPO fixed-KL gradients by 0.9663% and 10.8533%, while NLL branches are unchanged (Appendix J.1).

7 Conclusion

DARC couples model publication and authorized update recovery for unlearning between separate model and data holders. It combines correlated Gaussian views, client-authorized execution, and server-side inverse-scale aggregation. Under the stated publication and isolation conditions, our analysis quantifies bounded-coalition checkpoint privacy. For a shared locally smooth update map, aggregation cancels the common first-order perturbation response, with curvature, optimizer, and stochastic errors analyzed separately. The optimizer analysis distinguishes exact-arithmetic compatibility for signed permutations from discrepancies induced by continuous attention rotations. Empirically, In the Llama-1B TOFU comparison, the six non-collapsed trainers have utility scores within 0.003 of their noise-free references. Targeted diagnostics reveal numerical compatibility gaps and closely matching outputs from individual published views on tested prompts. These findings distinguish update fidelity, retained utility, and functional accessibility. Together, these results support the joint design of model publication and update recovery, connecting bounded checkpoint disclosure with optimizer-aware approximation of the chosen unlearning method's noise-free update.

AI USE STATEMENT

In this work, generative AI tools were used for language editing, code workings, figure drawing, and the drafting of candidate mathematical derivations. Authors conducted the verification of the final proofs, citations, code, and experiment provenance. The authors take responsibility for the final content of this work, including text, claims, code, and all other contents produced with the help of generative AI.

ETHICS STATEMENT

The intended use of DARC is to reduce unnecessary model disclosure while processing legitimate deletion requests near the data holder. The threshold guarantee depends on enforcing view isolation and on an explicit sensitivity bound; it must not be presented as cryptographic model secrecy. The protocol does not certify successful forgetting, and returned updates can reveal the client's forget data. Deployments should document the protected asset, adjacency relation, coalition threshold, scale-profile disclosure, cumulative privacy budget, and enforcement mechanism, and should conduct independent model-recovery and update-leakage audits before making legal or product privacy claims.

REPRODUCIBILITY STATEMENT

The appendix specifies the unlearning objectives, benchmark settings, separate mechanism protocols, inference configuration, and full mathematical proofs. The paper source includes scalar-value tables, presentation tables, and the quantitative figure's data mapping and generation script. Noise-free references, single-noisy controls, aligned updates, reused endpoints, and faulty oracle constructions are identified separately. Exported per-sample and per-tensor summaries support independent arithmetic checks in the experimental records.(Appendix J.6).

REFERENCES

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International conference on machine learning*, pp. 394–403. PMLR, 2018.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment problems: revised reprint*. SIAM, 2012.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In Dekai Wu, Marine Carpuat, Xavier Carreras, and Eva Maria Vecchi (eds.), *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012/.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8827–8840, 2025.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006b.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

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

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

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

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, 2023.

Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11278–11294, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Věra Kůrková and Paul C Kainen. Functionally equivalent feedforward neural networks. *Neural Computation*, 6(3):543–558, 1994.

Tim Tsz-Kit Lau and Weijie Su. Symmetry-compatible principle for optimizer design: Embeddings, lm heads, swiglu mlps, and moe routers. *arXiv preprint arXiv:2605.18106*, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Shuyang Ling. Near-optimal bounds for generalized orthogonal procrustes problem via generalized power method. *arXiv preprint arXiv:2112.13725*, 2021.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022a.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federated unlearning. *arXiv preprint arXiv:2012.13891*, 2020.

Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pp. 1749–1758. IEEE, 2022b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jinglong Luo, Zhuo Zhang, Yehong Zhang, Shiyu Liu, Ye Dong, Hui Wang, Yue Yu, Xun Zhou, and Zenglin Xu. Secp-tuning: Efficient privacy-preserving prompt tuning for large language models via mpc. In *International Conference on Learning Representations*, volume 2026, pp. 78760–78780, 2026.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Anmol Reddy Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid A Hasan, and Elita AA Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3732–3752, 2025.

Ilya Mironov. On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pp. 650–661, 2012.

Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, 2017.

Thai-Hung Nguyen, Hong-Phuc Vu, Dung Thuy Nguyen, Tuan Minh Nguyen, Khoa D Doan, and Kok-Seng Wong. Empirical study of federated unlearning: Efficiency and effectiveness. In *Asian Conference on Machine Learning*, pp. 959–974. PMLR, 2024.

Zibin Pan, Shuwen Zhang, Yuesheng Zheng, Chi Li, Yuheng Cheng, and Junhua Zhao. Multi-objective large language model unlearning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Ivo Petrov, Dimitar I Dimitrov, Maximilian Baader, Mark N Müller, and Martin Vechev. Dager: Exact gradient inversion for large language models. *Advances in neural information processing systems*, 37:87801–87830, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International conference on machine learning*, pp. 4596–4604. PMLR, 2018.

William F Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D Lane. Lunar: Llm unlearning via neural activation redirection. *arXiv preprint arXiv:2502.07218*, 2025.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, volume 2025, pp. 27797–27818, 2025.

Devender Singh. The loss does not see the basis, but adam does. *arXiv preprint arXiv:2608.05136*, 2026.

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. Balancing forget quality and model utility: A reverse kl-divergence knowledge distillation approach for better unlearning in llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1306–1321, 2025.

Guangxuan Xiao, Ji Lin, and Song Han. Offsite-tuning: Transfer learning without full model. *arXiv preprint arXiv:2302.04870*, 2023.

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 7937–7965, 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring criteria of loss reweighting to enhance llm unlearning. *arXiv preprint arXiv:2505.11953*, 2025.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Zhichao You, Xuewen Dong, Ke Cheng, Xutong Mu, Jiaxuan Fu, Shiyang Ma, Qiang Qu, and Yulong Shen. Prifft: Privacy-preserving federated fine-tuning of large language models via hybrid secret sharing. *IEEE Transactions on Dependable and Secure Computing*, 2026.

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

Tianyue Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding adam requires better rotation dependent assumptions. *Advances in Neural Information Processing Systems*, 38:115111–115152, 2026.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A Appendix Roadmap and Claim Scope

The appendix gives the privacy and decoding proofs, optimizer-state analysis, and Transformer/LoRA transformations, followed by complete historical measurements and the additional mechanism experiments. Experiments are organized by scientific question, with explicit distinctions between single-batch diagnostics, actual training, saved-update arithmetic and endpoint evaluation.

The formal checkpoint-publication claim applies to any coalition observing at most $t < m$ isolated same-round views. Ordinary RDP requires a fixed-public profile, or a separately privatized profile, together with the joint key/code sampling condition in Appendix B.1. Declassifying the historical task-vector profile would support a profile-stratified statement. The ideal Gaussian mechanism is distinguished from finite-precision sampling, and multi-round composition additionally requires conditional bounds given the coalition's observed history. The decoder theorem separately depends on valid perturbation neighborhoods, coupled randomness or controlled stochastic variance, and either optimizer-compatible keys or a controlled optimizer defect.

B PRIVACY THEORY FOR CORRELATED MODEL VIEWS

This appendix separates four mechanisms that are easy to conflate: (i) the Gaussian publication code, (ii) the optional parameter-coordinate key, (iii) the nonlinear local unlearning map, and (iv) the authorized estimator. Privacy is proved for (i) and is preserved under the adversary's post-processing. The coordinate key is therefore revealed in the proof. The decoder analysis is logically separate from the privacy proof.

B.1 COMPLETE-TRANSCRIPT CLOSURE AND PROTOCOL

Joint key/code sampling. Conditional on public information and the coalition's previously observed history h, the joint coordinate-key draw is independent of the current Gaussian code and has the same law for neighboring protected inputs. For the fixed-public-profile mechanism, writing $T_{1:m,r} = (T_{1,r}, \dots, T_{m,r})$ and $z^{(r)}_{1:m} = (z^{(r)}_1, \dots, z^{(r)}_m)$, the required sampling law is

$$\mathcal{L}\Big(T_{1:m,r}, z^{(r)}_{1:m} \mid \theta_{r-1} = \theta, h\Big) = \nu_{r,h}(dT_{1:m,r}) \prod_{k=1}^{m} \mathcal{N}(0, \kappa_r^2 \Sigma_r(h))(dz^{(r)}_k).$$

Here $\nu_{r,h}$ does not depend on the protected input or hidden server state. Keys may share randomness across copies or architectural blocks. Independence of the key's marginal law from the checkpoint alone is insufficient.

The primary Gaussian theorem is meaningful only for the complete server-to-executor publication. Let $Y_{S,r}(\theta)$ denote the canonicalized checkpoint views and $\widetilde{Y}_{S,r}(\theta) = (T_{S,r}, Y_{S,r}(\theta))$ their key-revealed version, where $T_{S,r} = (T_{k,r} : k \in S)$. The latter is equivalent to the published tensors together with their keys. Let $A_{S,r}(\theta)$ collect all other checkpoint-dependent objects available to coalition S: frozen reference or teacher models, adapter bases, optimizer states, masks, calibration statistics, and metadata. We impose the following closure condition.

Assumption 1 (Closed publication transcript)**.** *Conditional on public information and prior observed history, the auxiliary transcript is a randomized post-processing of the key-revealed coded views:*

$$\theta \longrightarrow \widetilde{Y}_{S,r}(\theta) \longrightarrow A_{S,r}(\theta) \qquad \textit{forms a Markov chain for every } |S| \le t. \tag{14}$$

Alternatively, an auxiliary mechanism may be separately released only if its conditional RDP cost is included by adaptive composition.

Under Assumption 1, data processing extends Theorem 1 from $\widetilde{Y}_{S,r}$ to $(\widetilde{Y}_{S,r}, A_{S,r})$ and hence to the actual hidden-key transcript. In particular, auxiliary objects derived from the published tensors satisfy this condition. The condition cannot be omitted: if a deterministic auxiliary statistic $a(\theta)$ is published and $a(\theta) \neq a(\theta')$ for an adjacent pair, then the joint outputs have disjoint support in that coordinate and their Rényi divergence is infinite. In particular, sending an uncoded private canonical reference model invalidates checkpoint privacy even when the trainable model view is correctly coded.

For reference-based unlearning, three implementations are formally distinct: (i) a genuinely public reference incurs no additional checkpoint release; (ii) freezing the executor's own coded initialization introduces no additional server message, although the objective is then defined relative to that coded initialization; and (iii) sending a second private reference requires a joint or sequentially composed coded mechanism. Appendix J.1 reports a source-level and single-batch audit of these alternatives. Historical trainer bindings and deployment isolation remain uncertified; the later local-reference experiments specify their own frozen-reference rule.

Algorithm 1 in Section 4 gives the complete round.

B.2 COALITION COVARIANCE AND SUFFICIENT STATISTICS

Fix one round, suppress the round index, and let the positive-definite scale profile be Σ. Stack the base perturbations as $e = [e_1; \dots; e_m]$. The construction in Eq. (3) gives

$$B_m = \frac{m}{m-1} I_m - \frac{1}{m-1} \mathbf{1}\mathbf{1}^\top, \qquad \mathrm{Cov}(e) = \kappa^2 B_m \otimes \Sigma. \tag{15}$$

Let $D = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$. The scaled perturbations have covariance

$$\mathrm{Cov}((D \otimes I_d)e) = \kappa^2 (DB_m D) \otimes \Sigma. \tag{16}$$

For a coalition $S \subset [m]$ of size s, let B_S and D_S denote the corresponding principal submatrices. Then

$$B_S = \frac{m}{m-1} I_s - \frac{1}{m-1} \mathbf{1}\mathbf{1}^\top. \tag{17}$$

The vector $\mathbf{1}$ has eigenvalue $(m-s)/(m-1)$, and every vector orthogonal to $\mathbf{1}$ has eigenvalue $m/(m-1)$. Hence $B_S \succ 0$ if and only if $s < m$, and

$$B_S^{-1} = \frac{m-1}{m} I_s + \frac{m-1}{m(m-s)} \mathbf{1}\mathbf{1}^\top. \tag{18}$$

Under the sampling condition in Appendix B.1, revealing and inverting the keys leaves the canonicalized tensors distributed as

$$Y_S(\theta) = \mathbf{1}_s \otimes \theta + \varepsilon_S, \qquad \varepsilon_S \sim \mathcal{N}\big(0, \kappa^2 M_S \otimes \Sigma\big), \quad M_S := D_S B_S D_S. \tag{19}$$

The full key-revealed observation is $(T_S, Y_S(\theta))$; its law is $\nu_S \otimes P_{\theta,S}$, where ν_S is the common key law and $P_{\theta,S} = \mathcal{L}(Y_S(\theta))$. The generalized least-squares statistic

$$Z_S := \frac{(\mathbf{1}^\top M_S^{-1}) \otimes I_d}{\gamma(S)} Y_S, \qquad \gamma(S) := \mathbf{1}^\top M_S^{-1} \mathbf{1}, \tag{20}$$

is sufficient for the common mean and obeys

$$Z_S = \theta + \zeta_S, \qquad \zeta_S \sim \mathcal{N}\left(0, \frac{\kappa^2}{\gamma(S)} \Sigma\right). \tag{21}$$

Equation (21) gives a useful interpretation: an incomplete coalition receives one effective noisy checkpoint whose precision is $\gamma(S)/\kappa^2$. It is not an additional release and is used only for analysis.

Corollary 1 (Exact two-checkpoint testing difficulty)**.** *Under the joint sampling condition in Appendix B.1, give the coalition the keys and equal prior probability on two candidate checkpoints θ and θ'. Let*

$$\mu_S(\theta, \theta') := \frac{\sqrt{\gamma(S)}}{\kappa} \left\| \Sigma^{-1/2} (\theta - \theta') \right\|_2. \tag{22}$$

The minimum possible candidate-identification error is

$$P_{\mathrm{e}}^*(S; \theta, \theta') = \Phi\left(-\frac{\mu_S(\theta, \theta')}{2} \right), \tag{23}$$

where Φ is the standard-normal CDF. Consequently, for every adjacent pair with distance at most Δ, the error is at least $\Phi(-\sqrt{\gamma(S)}\Delta/(2\kappa))$.

Proof. The independent keys have the same law under both candidates. Thus the sufficient statistic in Eq. (21) reduces the problem to discriminating two Gaussians with common covariance $\kappa^2\Sigma/\gamma(S)$. Their Mahalanobis separation is μ_S, and the equal-prior likelihood-ratio test has error $\Phi(-\mu_S/2)$. Revealing the keys only strengthens the attacker, so this is also a conservative interpretation of the actual hidden-key channel. □

This corollary is an operational interpretation for a fixed candidate pair, not a replacement for RDP: it assumes the adversary knows the two candidates and uses an average Bayes error, whereas RDP quantifies all neighboring pairs and composes adaptively.

Zero-noise and frozen coordinates. The positive-definite profile is substantive. For a Gaussian covariance $K \succeq 0$, two equal-covariance releases have finite divergence only if their mean difference lies in $\mathrm{range}(K)$; on that subspace the formula uses $K^\dagger$. Otherwise their affine supports are disjoint and the divergence is infinite. This follows by orthogonally diagonalizing K and separating noisy and deterministic coordinates. Consequently, an unperturbed private tensor or a zero task-vector-RMS block is safe only when it is identical across the declared adjacent inputs. A LoRA run that trains only adapters does not automatically protect its frozen base. The protected vector comprises all privately varying published tensors, with aliases counted once, regardless of which tensors are optimized.

B.3 Exact pairwise Rényi divergence

Proof of Theorem 1. For $s < m$, the covariance in Eq. (19) is positive definite. Equal-covariance Gaussians satisfy

$$D_\lambda(\mathcal{N}(\mu, K)\|\mathcal{N}(\mu', K)) = \frac{\lambda}{2}(\mu - \mu')^\top K^{-1}(\mu - \mu'), \qquad \lambda > 1. \tag{24}$$

Writing $\delta = \theta - \theta'$ and using the Kronecker inverse,

$$D_\lambda(Y_S(\theta)\|Y_S(\theta')) = \frac{\lambda}{2\kappa^2}\left(\mathbf{1}^\top D_S^{-1} B_S^{-1} D_S^{-1}\mathbf{1}\right)\left(\delta^\top \Sigma^{-1}\delta\right) \tag{25}$$

$$= \frac{\lambda}{2\kappa^2}\gamma(S)\left\|\Sigma^{-1/2}\delta\right\|_2^2. \tag{26}$$

Substituting Eq. (18) gives Eq. (8). The joint sampling condition gives

$$D_\lambda(\nu_S \otimes P_{\theta,S}\|\nu_S \otimes P_{\theta',S}) = D_\lambda(P_{\theta,S}\|P_{\theta',S}),$$

since the common key factor cancels in the likelihood ratio and integrates to one. Applying the invertible keys to Y_S is reversible when the keys are retained; hiding them is then post-processing and cannot increase divergence. With the forget data and other non-server inputs fixed across adjacent worlds, Assumption 1 makes the auxiliary transcript and every local trace further post-processing. This proves the key-revealed equality and the bound for the actual closed coalition view. □

The theorem is more naturally read as a metric privacy kernel:

$$D_\lambda(\mathsf{View}_S(\theta)\|\mathsf{View}_S(\theta')) \le \frac{\lambda\gamma(S)}{2\kappa^2} d_\Sigma(\theta,\theta')^2, \qquad d_\Sigma(\theta,\theta') = \left\|\Sigma^{-1/2}(\theta - \theta')\right\|_2. \tag{27}$$

Ordinary RDP follows after declaring an adjacency ball $d_\Sigma \le \Delta$. This distinction is important: the theorem controls neighboring checkpoints under a stated metric and does not make arbitrary, functionally distant checkpoints indistinguishable.

Lemma 1 (Worst coalition for ordered copy scales)**.** *Suppose $0 < \alpha_1 \le \cdots \le \alpha_m$. Among all coalitions with $|S| \le t < m$, $\gamma(S)$ is maximized by $S = \{1, \ldots, t\}$.*

Proof. Put $x_k = \alpha_k^{-1} > 0$. For a fixed size s, Eq. (8) is coordinatewise increasing in every x_k, so the s largest x_k are worst. Adding a positive coordinate increases both $\sum x_k^2$ and $\left(\sum x_k\right)^2/(m-s)$, including through the smaller denominator. Hence the maximum over $s \le t$ occurs at $s = t$ and uses the t smallest α_k. □

For equal unit scales, $\gamma_t = t(m-1)/(m-t)$ follows directly. As $t \uparrow m$, the smallest eigenvalue of B_S tends to zero and the privacy cost diverges.

B.4 Conversion and conditional multi-round composition

For (λ, ρ_λ)-RDP, the standard conversion (Mironov, 2017) gives

$$\epsilon \le \rho_\lambda + \frac{\log(1/\delta)}{\lambda - 1}. \tag{28}$$

When $\rho_\lambda = \lambda c$, optimization over $\lambda > 1$ gives $\epsilon \le c + 2\sqrt{c\log(1/\delta)}$. For one fixed coalition and the ideal fixed-profile Gaussian mechanism, the effective observation in Eq. (21) also gives the sharper analytical Gaussian privacy profile (Balle & Wang, 2018):

$$\delta_{\mathrm{G}}(\epsilon;\mu) = \Phi\left(\frac{\mu}{2} - \frac{\epsilon}{\mu}\right) - e^\epsilon \Phi\left(-\frac{\mu}{2} - \frac{\epsilon}{\mu}\right), \qquad \mu = \frac{\sqrt{\gamma(S)}\Delta}{\kappa}. \tag{29}$$

The expression is the exact pairwise hockey-stick divergence at separation μ and an upper bound over an adjacency ball of radius Δ; set it to zero when $\mu = 0$. It follows by integrating the likelihood-ratio test region for the two equal-covariance Gaussians. This improves one-round calibration, not the

assumptions needed to compose adaptive releases. The supplied RDP accountant remains a valid conservative conversion once those assumptions are certified.

Let X denote the protected input (a checkpoint or a server dataset), and let H_{r-1} be precisely the coalition's observed history, including prior messages, feedback, and coalition choices. The next subset S_r may depend on that history, but is selected before the current code is drawn; any fresh adversarial randomization can be conditioned on as independent public randomness. Profiles, scales, and κ_r are fixed conditional on that history or separately accounted. For every neighboring X, X' and every common feasible history h, composition requires the *conditional release kernels* to satisfy

$$D_\lambda(K_r(X,h)\|K_r(X',h)) \le \rho_{r,\lambda}. \tag{30}$$

It is not sufficient to check only their unconditional marginals.

A simple sufficient case is a deterministic checkpoint $G_r(X,h)$ given the protected input and observed history, satisfying

$$\sup_{X\simeq X'} \Big\|\Sigma_r(h)^{-1/2}[G_r(X,h)-G_r(X',h)]\Big\|_2 \le \Delta_r \quad \text{uniformly in } h. \tag{31}$$

In the actual iterative protocol, however, the server also uses updates from unobserved executors. Its checkpoint need not be deterministic given (X,h). Let $\pi_r^X(\cdot \mid h)$ be its conditional law, after marginalizing all hidden server state. The following condition accommodates that case without assuming the hidden state has the same posterior in neighboring worlds.

Proposition 2 (Conditional-mixture composition)**.** *Suppose, for every neighboring X, X' and common history h, the conditional checkpoint laws admit a coupling $\omega_{r,h}^{X,X'}$ with*

$$\Big\|\Sigma_r(h)^{-1/2}(\theta-\theta')\Big\|_2 \le \Delta_r \quad \omega_{r,h}^{X,X'}\text{-almost surely}. \tag{32}$$

Assume fresh Gaussian coding and the joint key/code sampling law in Appendix B.1 conditional on the checkpoint and h, with the same $\nu_{r,h}$ for both protected inputs and all coupled checkpoints. Assume same-round transcript closure and coalition selection as above. Then Eq. (30) holds with $\rho_{r,\lambda} = \lambda\Delta_r^2\Gamma_{t,r}/(2\kappa_r^2)$, and

$$\rho_\lambda^{1:R} \le \sum_{r=1}^{R} \frac{\lambda\Delta_r^2\Gamma_{t,r}}{2\kappa_r^2}. \tag{33}$$

Any separately protected auxiliary or feedback release adds its conditional RDP cost.

Proof. Fix (X, X', h) and use the same latent law ω for the pair (θ, θ') in two auxiliary mechanisms. In the first, sample the key-revealed view at θ; in the second, at θ', using the same independent key law $\nu_{r,h}$. The conditional divergence at each latent pair is bounded by Theorem 1. The order-λ divergence of the joint latent-pair/output distributions is

$$\frac{1}{\lambda-1}\log \mathbb{E}_\omega \exp\Big((\lambda-1)D_\lambda(\widetilde{Y}_{S_r}(\theta)\|\widetilde{Y}_{S_r}(\theta'))\Big) \le \rho_{r,\lambda}. \tag{34}$$

Marginalizing the latent pair and hiding the keys gives the actual conditional kernels, with any closed local transcript obtained by post-processing, so data processing proves Eq. (30). Uniform conditional RDP then composes adaptively (Mironov, 2017), giving Eq. (33). Dirac checkpoint laws recover Eq. (31). □

Why fresh randomness alone is insufficient. Let $Y_1 = X + Z$ and $Y_2 = Z + N$, with independent standard Gaussians Z, N. The marginal law of Y_2 is independent of X, and N is fresh. Nevertheless, (Y_1, Y_2) has mean $(X, 0)$ and covariance $[\begin{smallmatrix}1 & 1\\ 1 & 2\end{smallmatrix}]$. Its divergence between X and X' is $\lambda(X-X')^2$, twice that of Y_1 alone. Conditional on $Y_1 = y$, the second mean is $y - X$, so its conditional privacy cost is not zero. This example does not prove failure of DARC; it shows why marginal guarantees plus fresh noise are not a composition proof.

One must not replace the observed history by an all-copy history and then charge only the incomplete-coalition cost: the all-copy history already contains the decoding direction. Nor does using the same hidden random seed in an unconditioned sensitivity test establish Eq. (32). A final canonical checkpoint or server feedback depending on every view needs its own analysis. The existing multi-round experiments remain utility evidence until this conditional requirement is certified.

Fixed and adaptive same-round access. The single-round result quantifies every *fixed* subset S; thus it also covers a subset selected from prior history before fresh noise. It does not by itself cover selecting views after examining current unrevealed copies, or accumulating more than t views through retries or sequential worker reuse. Such access requires an interactive conditional analysis or a stricter deployment boundary.

B.5 Scale profiles: fixed, disclosed, implicit, or privatized

A data-dependent covariance is part of the randomized mechanism. Four regimes should be distinguished.

Fixed-public profile. If Σ is selected independently of the protected checkpoint, Theorem 1 gives ordinary metric RDP directly. A public floor $b_{\min} > 0$ should be applied to every block scale to keep $\Sigma \succ 0$.

Deterministically disclosed profile. Let $s(\theta)$ be a deterministic profile that is included in the output. If $s(\theta) \neq s(\theta')$, the two output distributions put unit mass on disjoint profile coordinates. Therefore

$$D_\lambda\big((s(\theta), Y_S(\theta))\|(s(\theta'), Y_S(\theta'))\big) = \infty. \tag{35}$$

A theorem conditioned on $s(\theta) = s(\theta')$ is valid only within that profile stratum; it is not ordinary DP over pairs with different disclosed scales.

Implicit checkpoint-dependent covariance. If the profile is not separately disclosed but controls the covariance, a finite unequal-covariance calculation is possible under scale-ratio restrictions. Let

$$\Sigma(\theta) = \mathrm{blkdiag}(b_1^2 I_{d_1}, \dots, b_L^2 I_{d_L}), \tag{36}$$

$$\Sigma(\theta') = \mathrm{blkdiag}(b_1'^2 I_{d_1}, \dots, b_L'^2 I_{d_L}), \qquad r_\ell := b_\ell'^2 / b_\ell^2. \tag{37}$$

Theorem 4 (Exact RDP with an implicit data-dependent block profile)**.** *For a strict coalition of size s and order $\lambda > 1$, assume the same joint key/code independence as in Appendix B.1, now with Gaussian profile $\Sigma(\theta)$ conditional on θ and the same key law under both inputs. Suppose $c_{\lambda,\ell} := \lambda r_\ell - (\lambda - 1) > 0$ for every block. With $P_{\theta,S} = \mathcal{L}(Y_S(\theta))$, the directional divergence, unchanged by revealing the keys, is*

$$D_\lambda(P_{\theta,S}\|P_{\theta',S}) = \sum_{\ell=1}^{L} \left[\frac{s d_\ell}{2(\lambda-1)} \log\left(\frac{r_\ell^\lambda}{c_{\lambda,\ell}}\right) \right. \tag{38}$$

$$\left. + \frac{\lambda \gamma(S)}{2\kappa^2 b_\ell^2 c_{\lambda,\ell}} \|\theta_\ell - \theta_\ell'\|_2^2 \right]. \tag{39}$$

If any $c_{\lambda,\ell} \leq 0$, the divergence is infinite. Both directions are finite only when

$$\frac{\lambda - 1}{\lambda} < r_\ell < \frac{\lambda}{\lambda - 1} \quad \textit{for every } \ell. \tag{40}$$

The actual hidden-key mechanism has no larger divergence.

Proof. The common independent key factor leaves the divergence unchanged, as in Theorem 1. Apply the invertible copy-space whitening $M_S^{-1/2} \otimes I$. For each scalar model coordinate in block ℓ, the two s-dimensional distributions have isotropic variances $\kappa^2 b_\ell^2$ and $\kappa^2 b_\ell'^2$, while their mean difference has squared norm $\gamma(S)(\theta_j - \theta_j')^2$. Direct Gaussian integration gives, for $P = \mathcal{N}(\mu, \sigma^2 I_n)$ and $Q = \mathcal{N}(\mu', \sigma'^2 I_n)$,

$$D_\lambda(P\|Q) = \frac{n}{2(\lambda-1)} \log \frac{r^\lambda}{\lambda r - (\lambda - 1)} + \frac{\lambda \|\mu - \mu'\|_2^2}{2\sigma^2[\lambda r - (\lambda - 1)]}, \quad r = \sigma'^2/\sigma^2, \tag{41}$$

provided the denominator is positive. Summing over coordinates gives Eq. (39). Swapping the two distributions gives the upper ratio condition in Eq. (40). □

The ratio window becomes narrow at high RDP orders. Therefore this optional theorem does not justify treating arbitrary task-vector-dependent scales as free; it instead shows precisely which additional scale-stability assumptions would be required.

Privately released profile. Suppose a randomized profile mechanism M_{prof} is $(\lambda, \rho_{\mathrm{prof}})$-RDP, and conditional on every profile value its subsequent model-view mechanism is uniformly $(\lambda, \rho_{\mathrm{view}})$-RDP. Adaptive composition gives

$$\rho_{\mathrm{joint}}(\lambda) \le \rho_{\mathrm{prof}}(\lambda) + \rho_{\mathrm{view}}(\lambda). \tag{42}$$

This route requires a concrete profile mechanism and is left as an extension.

B.6 Full-view failure and a Bayesian interpretation

For $S = [m]$, B_m is singular with null vector $\mathbf{1}$, and DB_mD has null vector $D^{-1}\mathbf{1}$. After key inversion,

$$\frac{\sum_{k=1}^{m} \alpha_k^{-1}(\theta + \alpha_k e_k)}{\sum_{k=1}^{m} \alpha_k^{-1}} = \theta. \tag{43}$$

Conditional on the revealed keys, distinct checkpoints are supported on disjoint affine subspaces in this direction, so the key-revealed full-view RDP is infinite. This is the intended authorization boundary of the theorem. Marginalizing hidden keys could only reduce divergence, but no finite full-view bound is established or claimed; Appendix J.9 reports bounded public-anchor and oracle recovery tests. They show functional proximity in the tested setting, not a finite full-view privacy bound or general resistance to recovery.

For intuition only, assume the stylized prior $\theta \sim \mathcal{N}(\mu_0, \tau^2\Sigma)$. Equation (21) yields posterior covariance

$$\mathrm{Cov}(\theta \mid Y_S) = \left(\tau^{-2} + \gamma(S)/\kappa^2\right)^{-1} \Sigma, \tag{44}$$

and Bayesian MMSE $\mathrm{tr}(\Sigma)/(\tau^{-2} + \gamma(S)/\kappa^2)$. The corresponding mutual information is

$$I(\theta; Y_S) = \frac{d}{2} \log\left(1 + \frac{\tau^2 \gamma(S)}{\kappa^2}\right). \tag{45}$$

These prior-dependent quantities make the coalition-size effect explicit, but neither is a substitute for the worst-case RDP statement.

C Authorized Denoising and Its Error Terms

C.1 Zero-sum Gaussian moments

For one block, draw $z_{k,\ell} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \kappa^2 b_\ell^2 I)$ and set

$$e_{k,\ell} = \sqrt{\frac{m}{m-1}} \left(z_{k,\ell} - \frac{1}{m} \sum_{j=1}^{m} z_{j,\ell} \right). \tag{46}$$

Then $\sum_k e_{k,\ell} = 0$ and

$$\mathrm{Var}(e_{k,\ell}) = \kappa^2 b_\ell^2 I, \tag{47}$$

$$\mathrm{Cov}(e_{k,\ell}, e_{j,\ell}) = -\frac{\kappa^2 b_\ell^2}{m-1} I, \qquad k \ne j. \tag{48}$$

C.2 Proof of optimizer-aware decoding

Proof of Theorem 2. Let $\delta_k = \alpha_k e_k$. The aligned update is $F(\theta + \delta_k) + R_{T_k}(\theta + \delta_k)$. Taylor's theorem with integral remainder gives

$$F(\theta + \delta_k) = F(\theta) + J(\theta)\delta_k + \rho_k, \qquad \rho_k = \int_0^1 [J(\theta + u\delta_k) - J(\theta)]\delta_k \, du, \tag{49}$$

and the Jacobian-Lipschitz condition in Theorem 2 implies

$$\|\rho_k\|_2 \le \frac{L_J}{2} \alpha_k^2 \|e_k\|_2^2. \tag{50}$$

Since $w_k\alpha_k = (\sum_j \alpha_j^{-1})^{-1}$,

$$\sum_k w_k J(\theta)\alpha_k e_k = \frac{J(\theta)}{\sum_j \alpha_j^{-1}} \sum_k e_k = 0. \tag{51}$$

The triangle inequality gives Eq. (10). Finally, when the same remainder bound is globally valid, $\mathbb{E}\,\|e_k\|_2^2 = \kappa^2 \operatorname{tr}(\Sigma)$ and $\sum_k w_k \alpha_k^2 = (\sum_k \alpha_k)/(\sum_k \alpha_k^{-1})$, proving the expectation statement in Theorem 2. □

Exact optimizer equivariance reduces Eq. (10) to

$$\left\|\bar{\Delta} - F(\theta)\right\|_2 \le \frac{L_J}{2} \sum_k w_k \alpha_k^2 \,\|e_k\|_2^2 . \tag{52}$$

C.3 Mean, variance, and a high-probability curvature bound

Define $W = \operatorname{diag}(w_1\alpha_1^2, \ldots, w_m\alpha_m^2)$ and

$$Q := \sum_k w_k \alpha_k^2 \,\|e_k\|_2^2 = e^\top (W \otimes I_d) e. \tag{53}$$

Let $H = B_m^{1/2} W B_m^{1/2} \succeq 0$. Gaussian quadratic-form identities give

$$\mathbb{E}Q = \kappa^2 \operatorname{tr}(H)\operatorname{tr}(\Sigma) = \kappa^2 \frac{\sum_k \alpha_k}{\sum_k \alpha_k^{-1}} \operatorname{tr}(\Sigma), \tag{54}$$

$$\operatorname{Var}(Q) = 2\kappa^4 \operatorname{tr}(H^2)\operatorname{tr}(\Sigma^2). \tag{55}$$

Moreover, the Gaussian quadratic-form inequality (Hsu et al., 2012) gives, with probability at least $1-\delta$,

$$Q \le \mathbb{E}Q + 2\kappa^2 \sqrt{\operatorname{tr}(H^2)\operatorname{tr}(\Sigma^2)\log(1/\delta)} \tag{56}$$

$$+ 2\kappa^2 \,\|H\|_{\mathrm{op}} \,\|\Sigma\|_{\mathrm{op}} \log(1/\delta). \tag{57}$$

The nonnegative scalar $U = (L_J/2)Q$ is an *upper envelope* for $\|C(e)\|$ where the smoothness condition holds. Its mean is $(L_J/2)\mathbb{E}Q$ and its variance is $(L_J^2/4)\operatorname{Var}(Q)$. Domination by U does not imply that the actual residual has variance bounded by $\operatorname{Var}(U)$. Under global remainder control, the valid second-moment consequence is

$$\mathbb{E}\,\|C(e)\|_2^2 \le \frac{L_J^2}{4}\left[(\mathbb{E}Q)^2 + \operatorname{Var}(Q)\right]. \tag{58}$$

For local smoothness, the same inequality holds with $\mathbf{1}_{\mathcal{E}_R}$ on the left. Neither statement includes optimizer mismatch.

C.4 Local smoothness under unbounded Gaussian perturbations

A Gaussian has unbounded support. Therefore a Lipschitz Jacobian only near θ does not justify an unconditional expectation bound over all publication noises. Suppose J_F is L_J-Lipschitz on the ball $B(\theta, R)$, and define

$$\mathcal{E}_R = \{\max_k \alpha_k \,\|e_k\|_2 \le R\}. \tag{59}$$

On this event every segment used in Taylor's theorem lies inside the ball. For $\delta_{\mathrm{loc}} \in (0,1)$, put $u = \log(m/\delta_{\mathrm{loc}})$ and

$$R_{\delta_{\mathrm{loc}}} = \alpha_{\max}\kappa \left[\operatorname{tr}\Sigma + 2\sqrt{\operatorname{tr}(\Sigma^2)u} + 2\,\|\Sigma\|_{\mathrm{op}}\, u\right]^{1/2}. \tag{60}$$

If $R \ge R_{\delta_{\mathrm{loc}}}$, the Gaussian quadratic-form tail bound (Hsu et al., 2012) and a union bound give $\Pr(\mathcal{E}_R) \ge 1 - \delta_{\mathrm{loc}}$. Independence between copies is not needed for this union bound. Combining it with the event in Eq. (57), of failure probability δ_Q, gives Theorem 2 with the stated quadratic envelope with probability at least $1 - \delta_{\mathrm{loc}} - \delta_Q$.

Define the actual curvature residual

$$C(e) = \sum_k w_k \big[F(\theta + \alpha_k e_k) - F(\theta) - J_F(\theta)\alpha_k e_k\big]. \tag{61}$$

The valid local expectation statement is

$$\mathbb{E}[\|C(e)\| \ \mathbf{1}_{\mathcal{E}_R}] \le \frac{L_J}{2}\mathbb{E}Q. \tag{62}$$

Thus its conditional expectation on $\mathcal{E}_R$ is bounded by $(L_J/2)\mathbb{E}Q/(1-\delta_{\mathrm{loc}})$, not automatically by $(L_J/2)\mathbb{E}Q$. If F is globally defined and a separate growth argument supplies $\mathbb{E}\,\|C(e)\|^2 \le M_C^2 < \infty$, Cauchy–Schwarz yields

$$\mathbb{E}\,\|C(e)\| \le \frac{L_J}{2}\mathbb{E}Q + M_C\sqrt{\delta_{\mathrm{loc}}}. \tag{63}$$

A global Jacobian-Lipschitz bound is an alternative that removes the tail qualification. No unconditional guarantee follows from local smoothness alone. The neighborhood event is used only in the utility analysis: rejecting or truncating code samples to enforce it would change the Gaussian publication mechanism and would require a new privacy proof.

Smoothness of the complete trainer. The map F includes all local steps, frozen-reference construction, masks, and state initialization, not just the bare loss at one parameter vector. If an update is the finite composition $\Phi_H \circ \cdots \circ \Phi_1$ of continuously twice differentiable augmented-state maps, with a continuously twice differentiable initialization map and bounded derivatives on a neighborhood of all trajectories, then the composition has a bounded second derivative there. Its displacement map consequently has a Lipschitz Jacobian. Smooth activations, positive normalization stabilizers, and positive adaptive-moment floors provide useful sufficient conditions, but they do not cover hard clipping thresholds, discrete sampling changes, or moment coordinates at zero automatically. Those cases must stay inside one smooth branch or be analyzed separately. In particular, safe-key *equivariance* can be exact even where this separate smoothness assumption is not verified.

C.5 Uniqueness, random optimizer error, and the single-copy control

Proof of Proposition 1. For one scalar coordinate, zero-sum vectors form $\mathcal{S} = \{e \in \mathbb{R}^m : \mathbf{1}^\top e = 0\}$. Cancellation for every $e \in \mathcal{S}$ requires $(w_1\alpha_1, \dots, w_m\alpha_m)^\top \in \mathcal{S}^\perp = \mathrm{span}\{\mathbf{1}\}$. Thus $w_k\alpha_k = c$. The normalization $\sum_k w_k = 1$ gives $c = (\sum_j \alpha_j^{-1})^{-1}$. □

Suppose the optimizer defect decomposes as $R_{T_k}(x_k) = b_k + \zeta_k$ conditional on the initializations, with $\mathbb{E}\zeta_k = 0$. Then

$$\mathbb{E}\sum_k w_k R_{T_k}(x_k) = \sum_k w_k b_k, \tag{64}$$

$$\mathrm{Cov}\left(\sum_k w_k R_{T_k}(x_k)\right) = \sum_{i,j} w_i w_j\,\mathrm{Cov}(\zeta_i, \zeta_j). \tag{65}$$

Independent, equal-covariance errors reduce by $\sum_k w_k^2 = \sum_k \alpha_k^{-2}/(\sum_k \alpha_k^{-1})^2$, but a shared bias or correlated error remains.

Two distinct single-copy controls. If a baseline commits the noisy initialization as well as the displacement, its next state is

$$\theta^+_{\mathrm{single,state}} = \theta + \varepsilon + \eta_{\mathrm{srv}}F(\theta+\varepsilon) = \theta + \eta_{\mathrm{srv}}F(\theta) + (I + \eta_{\mathrm{srv}}J)\varepsilon + \eta_{\mathrm{srv}}\rho(\varepsilon). \tag{66}$$

If it commits only the displacement onto the clean server state, it instead has

$$\theta^+_{\mathrm{single,update}} = \theta + \eta_{\mathrm{srv}}F(\theta+\varepsilon) = \theta + \eta_{\mathrm{srv}}F(\theta) + \eta_{\mathrm{srv}}J\varepsilon + \eta_{\mathrm{srv}}\rho(\varepsilon). \tag{67}$$

Both generally retain a first-order response, but only the first contains the identity term.

Matching a mean copy scale is also not variance matching. For a uniformly chosen copy, the covariance is $\kappa^2\mathbb{E}_k[\alpha_k^2]\Sigma$, whereas scaling a single Gaussian by $\mathbb{E}_k\alpha_k$ gives $\kappa^2(\mathbb{E}_k\alpha_k)^2\Sigma$. With $(\alpha_1, \alpha_2) = (1, 2)$ these factors are 2.5 and 2.25, respectively. A variance-matched single-copy comparator under this uniform-copy convention uses $\kappa\sqrt{2.5}$, rather than 1.5κ. Other matching targets, such as weighted decoded-coordinate energy, must be declared explicitly.

C.6 INDEPENDENT LOCAL SEEDS: MEAN-MAP CANCELLATION AND VARIANCE

The distributional optimizer identity in Proposition 6 compares two runs at the *same* initialization. In a coded round, $x_k = \theta + \alpha_k e_k$ varies with k, so equal covariance across copies cannot be assumed. Under safe keys, write the aligned output as $Y_k = F_{\xi_k}(x_k)$ and define $\mu(x) = \mathbb{E}_\xi F_\xi(x)$ and $C_{ij}(e) = \mathrm{Cov}(Y_i, Y_j \mid e)$.

Proposition 3 (Conditional bias–variance decomposition)**.** *Suppose the conditional means are $\mathbb{E}[Y_k \mid e] = \mu(x_k)$ and second moments are finite. Then*

$$\mathbb{E}\left[\left\|\sum_k w_k Y_k - \mu(\theta)\right\|_2^2 \middle| e\right] = \left\|\sum_k w_k \mu(x_k) - \mu(\theta)\right\|_2^2 + \mathrm{tr}\sum_{i,j} w_i w_j C_{ij}(e). \tag{68}$$

If J_μ is L_μ-Lipschitz on the relevant segments, the squared bias is at most $L_\mu^2 Q^2/4$. For conditionally independent seeds, the covariance term is $\sum_k w_k^2 \,\mathrm{tr}\, C_{kk}(e)$. It reduces to $\sigma^2 \sum_k w_k^2$ only when every trace equals σ^2 (or is bounded by it for an upper bound).

Proof. Subtract the conditional mean of the aggregate. The centered cross term has zero conditional expectation, and the centered squared norm is the trace of its covariance. Apply Taylor's theorem to μ; its common Jacobian cancels by $\sum_k w_k \alpha_k e_k = 0$. Conditional independence eliminates only the off-diagonal covariances, not the initialization dependence of $C_{kk}(e)$. □

A random-Jacobian term that need not cancel. For a local expansion at θ, suppose $J_k = J_{F_{\xi_k}}(\theta)$ are independent and identically distributed, have finite second moments, and are independent of the publication noises. Put $\bar{J} = \mathbb{E}J_k$ and $V_J = \mathbb{E}\,\mathrm{tr}[(J_k - \bar{J})\Sigma(J_k - \bar{J})^\top]$. The first-order random component is

$$L = \sum_k w_k \alpha_k (J_k - \bar{J}) e_k, \qquad \mathbb{E}\,\|L\|_2^2 = \kappa^2 \sum_k w_k^2 \alpha_k^2 V_J = \frac{\kappa^2 m}{(\sum_k \alpha_k^{-1})^2} V_J. \tag{69}$$

For $i \neq j$, independence and centering of $J_i - \bar{J}$ and $J_j - \bar{J}$ make the cross terms zero even though e_i, e_j are correlated. The diagonal terms use $\mathrm{Cov}(e_k) = \kappa^2\Sigma$. For equal unit scales this is $\kappa^2 V_J/m$: it is second order in noise amplitude (first order in its variance), unlike the $O(\kappa^4)$ squared curvature envelope. With a common coupled Jacobian, by contrast, L vanishes pathwise. This calculation does not assert that other training-noise or curvature terms are absent.

For a scalar check, let $F_\xi(x) = \xi x$, where ξ is an independent random sign, and let $\theta = 0$, $\Sigma = 1$, $\alpha_k = 1$. The aggregate is identically zero with a shared sign, but has exact mean-square error κ^2/m with independent signs. Thus zero structural Adam mismatch alone does not imply zero realized stochastic decoding error.

C.7 SECOND-ORDER BIAS DOES NOT VANISH WITH MORE COPIES

Consider one scalar component $F(x) = \frac{1}{2}\|x\|_2^2$ at $\theta = 0$, equal scales $\alpha_k = a$, and $w_k = 1/m$. This globally smooth quadratic map has zero linear term and exact residual $C(e) = a^2 \sum_k \|e_k\|^2/(2m)$. The Gaussian moments above give

$$\mathbb{E}C(e) = \frac{a^2\kappa^2}{2}\,\mathrm{tr}\,\Sigma, \qquad \mathrm{Var}(C(e)) = \frac{a^4\kappa^4}{2(m-1)}\,\mathrm{tr}(\Sigma^2). \tag{70}$$

Hence the MSE equals the square of the first expression plus the second. As m increases, variance decreases but the curvature bias remains. More generally, if component F_j is quadratic with Hessian H_j, its mean residual is $\frac{1}{2}\kappa^2(\sum_k \alpha_k)/(\sum_k \alpha_k^{-1})\,\mathrm{tr}(H_j\Sigma)$. This provides a sharp example of what authorized first-order cancellation does *not* remove; the bounds are not claims of exact recovery after nonlinear local training.

C.8 Privacy–curvature calibration and the role of copy scales

Combining Theorem 1 with Eq. (54) gives Eq. (11). The following result clarifies the heterogeneous α_k schedule.

Proposition 4 (Equal scales optimize the worst-coalition privacy–curvature factor)**.** *For positive $\alpha_1, \dots, \alpha_m$ and $1 \le t < m$, define*

$$\Gamma_t(\alpha) = \max_{|S| \le t} \gamma(S), \qquad C(\alpha) = \frac{\sum_k \alpha_k}{\sum_k \alpha_k^{-1}}. \tag{71}$$

Then

$$\Gamma_t(\alpha) C(\alpha) \ge \frac{t(m-1)}{m-t}, \tag{72}$$

with equality when all copy scales are equal. Equality is unique up to the common scale under the strict positivity assumptions.

Proof. Let $x_k = \alpha_k^{-1}$, let $X = \sum_k x_k$, and let A be the sum of the t largest x_k. For their index set $S_\star$, Cauchy–Schwarz gives

$$\gamma(S_\star) \ge \frac{m-1}{m} \left(\frac{A^2}{t} + \frac{A^2}{m-t} \right) = \frac{m-1}{t(m-t)} A^2. \tag{73}$$

Since $A \ge (t/m)X$, $\Gamma_t \ge t(m-1)X^2/[m^2(m-t)]$. Multiplying by $C = (\sum_k \alpha_k)/X$ and using $(\sum_k \alpha_k)(\sum_k \alpha_k^{-1}) \ge m^2$ proves Eq. (72). Equality in both inequalities requires all α_k to be equal. □

Thus heterogeneous scales remain supported, for example when executors require different perturbation levels, but they cannot improve this symmetric worst-case privacy–curvature upper bound.

D Transformer Reparameterization Details

D.1 Grouped-query attention

Let H_Q and H_{KV} be the numbers of query and key/value heads, d_h the head width, and $\pi : [H_Q] \to [H_{KV}]$ the grouped-query assignment. Sample $S_j \in O(d_h)$ for each key/value head and define

$$S_{KV} = \mathrm{blkdiag}(S_1, \dots, S_{H_{KV}}), \qquad U_\pi = \mathrm{blkdiag}(S_{\pi(1)}, \dots, S_{\pi(H_Q)}). \tag{74}$$

Using the convention that activations right-multiply projection matrices, transform

$$W_Q' = W_Q U_\pi, \quad W_K' = W_K S_{KV}, \quad W_V' = W_V S_{KV}, \quad W_O' = U_\pi^\top W_O. \tag{75}$$

Biases, if present, are right-multiplied by the same head matrices. For query head i, $Q_i' = Q_i S_{\pi(i)}$ and $K'_{\pi(i)} = K_{\pi(i)} S_{\pi(i)}$, so orthogonality preserves the attention logits. The value output is right-multiplied by $S_{\pi(i)}$, and concatenation produces U_π, which is canceled by W_O'.

D.2 RoPE commutant and optimizer-safe quarter turns

At position p, RoPE acts in a head as

$$\Phi(p) = \mathrm{blkdiag}(R(\omega_1 p), \dots, R(\omega_{d_h/2} p)), \quad R(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \tag{76}$$

A sufficient condition is $S_j \Phi(p) = \Phi(p) S_j$ for every p. With distinct frequencies, an orthogonal member of the commutant is

$$S_j = \mathrm{blkdiag}(R(\varphi_{j,1}), \dots, R(\varphi_{j,d_h/2})). \tag{77}$$

The continuous angles define $\mathcal{G}_{\mathrm{rot}}$. Restricting each angle to $q\pi/2$ makes every plane a signed permutation; these quarter turns define the attention component of $\mathcal{G}_{\mathrm{safe}}$. When frequencies repeat, the commutant may be larger, but only explicitly verified transformations should be used.

D.3 Standard and gated feed-forward blocks

For $f(x) = W_2\sigma(W_1x + b_1) + b_2$ and a hidden-channel permutation P, use

$$W_1' = PW_1, \quad b_1' = Pb_1, \quad W_2' = W_2P^\top, \quad b_2' = b_2. \tag{78}$$

Elementwise nonlinearities satisfy $\sigma(Pz) = P\sigma(z)$. In SwiGLU/GEGLU, the same P is applied to the gate and up branches; since $(Pa) \odot (Pb) = P(a \odot b)$, the down projection cancels it. Importantly, Eq. (78) is an ordinary permutation of the entries of the global parameter vector. No dense mixing occurs.

D.4 LoRA factors

For $W_{\text{eff}} = W + BA$, an input/output symmetry $W' = P_{\text{out}}WP_{\text{in}}^\top$ is extended by

$$B' = P_{\text{out}}B, \qquad A' = AP_{\text{in}}^\top. \tag{79}$$

The rank-space gauge $B' = BQ$, $A' = Q^{-1}A$ also preserves BA, but a dense Q is generally not compatible with coordinatewise optimizer states. The optimizer-safe variant restricts Q to signed permutations.

D.5 Residual paths, normalization, tying, and fused storage

The MLP and attention keys act only in internal hidden-channel or head coordinates. Their transformed branch output is exactly the same tensor in the residual-stream coordinates, so residual addition and pre/post normalization remain unchanged. The construction leaves the residual-stream basis, token embeddings, final normalization, and language-model head fixed; no commutation property of RMSNorm or LayerNorm with a dense residual rotation is assumed.

Implementation aliases must nevertheless be respected. When input embeddings and the language-model head share one parameter object, the protocol perturbs and serializes that shared object once. If a framework materializes tied weights as duplicated tensors, the same perturbation and any allowed key action must be applied consistently to both aliases. Likewise, fused QKV projections and packed gate/up projections must apply the prescribed block transform to the correct slices. Breaking a tie or transforming only one packed slice destroys exact functional equivalence and becomes an implementation defect rather than optimizer mismatch.

All transformations above are linear isometries when the constituent matrices are orthogonal. Hence $T(a + b) = T(a) + T(b)$ and $T^{-1}\Delta$ correctly maps a displacement into canonical coordinates.

E Optimizer Equivariance and Adam Rotation Error

E.1 Exact pathwise equivariance

Let the local loss and all auxiliary constructions be symmetry covariant. For an output-based objective with a fixed public reference, this reduces to $U(T\theta) = U(\theta)$; when a coded reference is constructed from the initialization, the same identity applies jointly to the trainable parameters and that reference. A parameter-indexed penalty or mask must also transform consistently. For orthogonal T,

$$\nabla U(T\theta) = T\nabla U(\theta). \tag{80}$$

Pathwise statements require coupled minibatches and every coordinate-indexed source of randomness, such as dropout masks, to be transformed consistently.

For AdamW, write

$$m_h = \beta_1 m_{h-1} + (1 - \beta_1)g_h, \tag{81}$$

$$v_h = \beta_2 v_{h-1} + (1 - \beta_2)g_h^{\odot 2}, \tag{82}$$

$$\theta_{h+1} = (1 - \eta_h\lambda_{\text{wd}})\theta_h - \eta_h \frac{\widehat{m}_h}{\sqrt{\widehat{v}_h} + \varepsilon}. \tag{83}$$

The same argument covers Adam by replacing decoupled decay with the chosen regularization convention.

Proof of Theorem 3. For SGD, Eq. (80) and orthogonality imply that the transformed update, mapped back by $T^\top$, equals the canonical update. Scalar momentum recurrences, Nesterov look-ahead, global ℓ_2 clipping (or another orthogonally invariant clipping rule), and scalar Euclidean weight decay commute with T, so induction proves trajectory equivariance.

Let T be a signed permutation and $P = |T|$ its unsigned permutation. If $g'_h = Tg_h$, induction gives $m'_h = Tm_h$ and $v'_h = Pv_h$. Bias correction is scalar, square roots and coordinatewise maxima commute with P, and

$$\frac{Tm}{\sqrt{Pv} + \varepsilon} = T\left(\frac{m}{\sqrt{v} + \varepsilon}\right). \tag{84}$$

Decoupled scalar weight decay also commutes with T. The same coordinatewise argument applies to AdaGrad, RMSProp, and AMSGrad. Parameter groups must carry identical hyperparameters under the permutation, and nonzero states must be transformed as above.

The MLP/SwiGLU transformations in Eq. (78) are global parameter permutations, and RoPE quarter turns are signed permutations. Therefore $R_T \equiv 0$ for every realized safe key in exact arithmetic. In particular, for every $p > 0$,

$$\mathbb{E}_T \left\|R_T(x)\right\|_2^p = 0, \qquad \Pr_T[\left\|R_T(x)\right\|_2 > 0] = 0. \tag{85}$$

□

Because the safe-key identity holds pointwise in both the initialization and the key, Eq. (85) also remains zero after taking any joint expectation over the publication perturbation, the safe-key draw, and a data distribution, provided the coupled optimizer assumptions hold.

Proposition 5 (State-mapping error for signed-permutation Adam)**.** *Let T be a signed permutation and $P = |T|$. At one Adam-family step, let (m, v) be the canonical bias-corrected moments and let (m', v') be the moments supplied in transformed coordinates. Pull them back as $\bar{m} = T^\top m'$ and $\bar{v} = P^\top v'$, and put $\delta_m = \bar{m} - m$ and $\delta_v = \bar{v} - v$. For $u(m, v) = m \oslash (\sqrt{v} + \varepsilon)$, if every coordinate of v and $\bar{v}$ is at least $\nu > 0$, then*

$$\left\|T^\top u(m', v') - u(m, v)\right\|_2 \le \frac{\left\|\delta_m\right\|_2}{\sqrt{\nu} + \varepsilon} + L_{\varepsilon,\nu} \left\|m\right\|_2 \left\|\delta_v\right\|_\infty, \qquad L_{\varepsilon,\nu} = \frac{1}{2\sqrt{\nu}(\sqrt{\nu} + \varepsilon)^2}. \tag{86}$$

Without a positive floor, the always-valid fallback is

$$\left\|T^\top u(m', v') - u(m, v)\right\|_2 \le \varepsilon^{-1} \left\|\delta_m\right\|_2 + \varepsilon^{-2} \left\|m\right\|_2 \sqrt{\left\|\delta_v\right\|_\infty}. \tag{87}$$

The corresponding parameter-step discrepancy is multiplied by the learning rate. In particular, fresh zero state or exactly mapped state makes both residuals zero.

Proof. Signed-permutation algebra gives $T^\top u(m', v') = u(\bar{m}, \bar{v})$. Decompose

$$u(\bar{m}, \bar{v}) - u(m, v) = \delta_m \oslash (\sqrt{\bar{v}} + \varepsilon) + m \odot \left[f_\varepsilon(\bar{v}) - f_\varepsilon(v)\right], \tag{88}$$

where $f_\varepsilon(x) = (\sqrt{x} + \varepsilon)^{-1}$ acts coordinatewise. The floor-based result follows from the supremum and Lipschitz constants of f_ε on $[\nu, \infty)$. The floor-free result uses $f_\varepsilon \le \varepsilon^{-1}$ and $|f_\varepsilon(x) - f_\varepsilon(y)| \le \varepsilon^{-2}\sqrt{|x - y|}$. □

Proposition 6 (Distributional equivariance under uncoupled invariant randomness)**.** *Let ξ denote the complete local-randomness sequence, including minibatch order, dropout masks, sampling decisions, and randomized kernels. Suppose there is a measurable probability-preserving action τ_T such that*

$$T^{-1}\Delta_{\mathcal{A}}(Tx; \mathcal{D}_{\mathrm{f}}, \tau_T \xi) = \Delta_{\mathcal{A}}(x; \mathcal{D}_{\mathrm{f}}, \xi) \quad \textit{for every } \xi. \tag{89}$$

Then an independently sampled $\xi' \overset{d}{=} \xi$ satisfies

$$T^{-1}\Delta_{\mathcal{A}}(Tx; \mathcal{D}_{\mathrm{f}}, \xi') \overset{d}{=} \Delta_{\mathcal{A}}(x; \mathcal{D}_{\mathrm{f}}, \xi). \tag{90}$$

Let $\mu_F(x) = \mathbb{E}_\xi[\Delta_{\mathcal{A}}(x; \mathcal{D}_{\mathrm{f}}, \xi)]$ and $C_F(x) = \mathrm{Cov}_\xi(\Delta_{\mathcal{A}}(x; \mathcal{D}_{\mathrm{f}}, \xi))$, and let ξ, ξ' be independent. For the uncoupled discrepancy

$$\widetilde{R}_T(x; \xi, \xi') := T^{-1}\Delta_{\mathcal{A}}(Tx; \mathcal{D}_{\mathrm{f}}, \xi') - \Delta_{\mathcal{A}}(x; \mathcal{D}_{\mathrm{f}}, \xi), \tag{91}$$

whenever the second moment is finite,

$$\mathbb{E}[\widetilde{R}_T] = 0, \qquad \mathbb{E}\left\|\widetilde{R}_T\right\|_2^2 = 2\,\mathrm{tr}\,C_F(x), \tag{92}$$

$$\Pr\left[\left\|\widetilde{R}_T\right\|_2 \geq a\right] \leq \frac{2\,\mathrm{tr}\,C_F(x)}{a^2}, \qquad a > 0. \tag{93}$$

For m independent transformed runs with weights summing to one, the centered aggregate satisfies

$$\mathbb{E}\,\|\textstyle\sum_k w_k(Y_k - \mu_F(x))\|_2^2 = \left(\sum_k w_k^2\right)\mathrm{tr}\,C_F(x), \tag{94}$$

where every pulled-back Y_k has the same law as the canonical update.

Proof. Because τ_T preserves the law of the complete randomness sequence, Eq. (89) and a change of variables give Eq. (90). Thus the two terms in Eq. (91) are independent with common mean $\mu_F(x)$ and covariance $C_F(x)$. The first identity follows by centering, the second by adding the two covariance traces, and Eq. (93) follows from Markov's inequality applied to the squared norm. Independence eliminates cross-covariances in the weighted aggregate, proving Eq. (94). □

Independent rather than explicitly coupled i.i.d. dropout masks therefore weaken pathwise equality to distributional equality. However, using the same random seed is not necessarily a valid symmetry coupling because tensor traversal may assign masks to different hidden channels.

Table 3: Idealized optimizer compatibility with the key families. "Exact" assumes correctly transformed state, groups, masks, and auxiliary operations.

Optimizer/update rule	Largest key class used here	Status
SGD, momentum, Nesterov	orthogonal T	exact pathwise
Full-matrix preconditioner with conjugated state	compatible orthogonal T	exact pathwise
AdaGrad, RMSProp, Adam, AdamW, AMSGrad	signed permutations	exact pathwise
Lion; LARS/LAMB with preserved groups	signed permutations	exact pathwise
Factored Adafactor	signed row/column permutations	exact with mapped factor state
Coordinatewise adaptive methods	continuous rotations	anisotropy-controlled defect
Paged/fused/mixed-precision implementations	implementation dependent	empirical audit required

The signed-permutation result extends beyond the named optimizers. Any update rule assembled from scalar affine state recurrences, coordinatewise odd/even maps that transform consistently under sign changes, permutation-invariant scalar reductions within preserved parameter groups, coordinatewise maxima, scalar clipping, and scalar weight decay is signed-permutation equivariant. This covers Lion's sign update (Chen et al., 2023) and the layerwise trust ratios of LARS/LAMB when groups are preserved (You et al., 2017; 2019). Standard factored Adafactor (Shazeer & Stern, 2018) is exact for row/column signed permutations when its row and column factor states are permuted with the corresponding axes; an arbitrary vectorized permutation need not preserve its factorization. Conversely, a full-matrix preconditioner is orthogonally equivariant only when its state is transformed by conjugation rather than discarded or diagonalized.

E.2 The intrinsic Adam defect for a dense rotation

Assume that, up to step h, a canonical trajectory and a pulled-back rotated trajectory share the same gradient history. Define the bias-corrected moments

$$\widehat{m}_h = \sum_{s=1}^{h} b_{h,s} g_s, \qquad \widehat{C}_h = \sum_{s=1}^{h} a_{h,s} g_s g_s^\top, \tag{95}$$

where

$$a_{h,s} = \frac{(1-\beta_2)\beta_2^{h-s}}{1-\beta_2^h}, \qquad b_{h,s} = \frac{(1-\beta_1)\beta_1^{h-s}}{1-\beta_1^h}. \tag{96}$$

Adam stores $\mathrm{diag}(\widehat{C}_h)$, while a rotated run stores $\mathrm{diag}(T\widehat{C}_h T^\top)$. Put

$$A_\varepsilon(C) := \mathrm{Diag}\Big((\sqrt{\mathrm{diag}(C)} + \varepsilon\mathbf{1})^{-1}\Big). \tag{97}$$

The pulled-back difference between the two normalized adaptive directions is exactly

$$q_h(T) = \left[T^\top A_\varepsilon(T\widehat{C}_h T^\top)T - A_\varepsilon(\widehat{C}_h)\right]\widehat{m}_h. \tag{98}$$

The one-step parameter discrepancy generated at a common state is $\eta_h q_h(T)$. Scalar AdamW weight decay cancels from this comparison.

E.3 A FIRST-STEP EXAMPLE: RANDOM ROTATIONS NEED NOT BE UNBIASED

Take a two-coordinate gradient $g = (a, 0)^\top$, $a > 0$, fresh zero Adam moments, and no weight decay. Bias correction gives the canonical direction $u(g) = (a/(a+\varepsilon), 0)^\top$. For a uniformly random planar rotation R_φ, the first coordinate of the pulled-back direction is

$$\left[R_\varphi^\top u(R_\varphi g)\right]_1 = \frac{a\cos^2\varphi}{a|\cos\varphi| + \varepsilon} + \frac{a\sin^2\varphi}{a|\sin\varphi| + \varepsilon}. \tag{99}$$

Its second-coordinate expectation is zero by reflection symmetry. As $\varepsilon/a \to 0$, dominated convergence gives

$$\mathbb{E}_\varphi[R_\varphi^\top u(R_\varphi g) - u(g)] \longrightarrow (4/\pi - 1, 0)^\top \neq 0. \tag{100}$$

At $\varphi = \pi/4$, the first-coordinate defect tends to $\sqrt{2} - 1$. These are normalized-direction errors; parameter errors are multiplied by the learning rate. This explicit example complements prior observations of Adam's basis dependence (Singh, 2026). It rules out the claim that averaging random continuous keys necessarily removes optimizer bias, even with zero publication noise. The full second moment here is rank one, explaining why a positive rotation-uniform denominator floor fails at this step. MLP permutations and quarter turns do not have this defect.

E.4 DETERMINISTIC ANISOTROPY BOUND

Partition T and $\widehat{C}_h$ into the independently rotated blocks. In block b of dimension q_b, define

$$\mu_b = \frac{\operatorname{tr}(\widehat{C}_b)}{q_b}, \qquad A_b = \widehat{C}_b - \mu_b I. \tag{101}$$

Assume every canonical and rotated diagonal second moment is at least $\nu_b > 0$ and set

$$L_{\varepsilon,b} = \sup_{x \geq \nu_b} \left|\frac{d}{dx}\frac{1}{\sqrt{x} + \varepsilon}\right| = \frac{1}{2\sqrt{\nu_b}(\sqrt{\nu_b} + \varepsilon)^2}. \tag{102}$$

Proposition 7 (Anisotropy controls the dense-rotation defect)**.** *Under the preceding floor condition,*

$$\|q_h(T)\|_2^2 \leq 4\sum_b L_{\varepsilon,b}^2 \|A_b\|_{\mathrm{op}}^2 \|\widehat{m}_{h,b}\|_2^2. \tag{103}$$

In particular, $q_h(T) = 0$ for every rotation whenever each $\widehat{C}_b$ is isotropic.

Proof. Subtract the scalar preconditioner $(\sqrt{\mu_b} + \varepsilon)^{-1} I$, which commutes with T_b. By scalar Lipschitzness, every diagonal entry of the remaining canonical or rotated preconditioner is at most $L_{\varepsilon,b}\|A_b\|_{\mathrm{op}}$ in magnitude. The two terms in Eq. (98) therefore contribute at most $2L_{\varepsilon,b}\|A_b\|_{\mathrm{op}}\|\widehat{m}_{h,b}\|_2$. Squaring blockwise proves the result. □

The floor assumption can be made rotation-uniform by taking $\nu_b \leq \lambda_{\min}(\widehat{C}_b)$, because every diagonal entry of $T_b\widehat{C}_b T_b^\top$ is a Rayleigh quotient. Early in optimization $\widehat{C}_b$ is often rank deficient, so this universal floor can be zero even when the currently stored diagonal is positive. The following fallback remains valid without a spectral floor.

Proposition 8 (Floor-free Hölder bound)**.** *For every positive-semidefinite $\widehat{C}_h$, every block-orthogonal T, and every $\varepsilon > 0$,*

$$\|q_h(T)\|_2^2 \leq \frac{4}{\varepsilon^4}\sum_b \|A_b\|_{\mathrm{op}} \|\widehat{m}_{h,b}\|_2^2. \tag{104}$$

Proof. For $f_\varepsilon(x) = (\sqrt{x} + \varepsilon)^{-1}$ and $x, y \geq 0$,

$$|f_\varepsilon(x) - f_\varepsilon(y)| \leq \frac{|\sqrt{x} - \sqrt{y}|}{\varepsilon^2} \leq \frac{\sqrt{|x - y|}}{\varepsilon^2}. \tag{105}$$

Use $y = \mu_b$ for every canonical and rotated diagonal entry. Since $|[T_b A_b T_b^\top]_{ii}| \leq \|A_b\|_{\mathrm{op}}$, each of the two preconditioner deviations has operator norm at most $\varepsilon^{-2}\sqrt{\|A_b\|_{\mathrm{op}}}$. The triangle inequality and blockwise squaring prove Eq. (104). □

The denominator floor is substantive. A tiny second-moment coordinate can make the Lipschitz constant large even when absolute anisotropy is small. The floor-free bound is mathematically robust but is usually conservative because Adam uses a small ε and the coefficient scales as ε^{-4}. A diagnostic should therefore report the universal spectral floor $\lambda_{\min}(\widehat{C}_b)$, a lower quantile of the realized canonical and rotated diagonals, the relative anisotropy $\|A_b\|_F / \left\|\widehat{C}_b\right\|_F$, and the exact defect in Eq. (98).

A directly measurable negligibility certificate. Let $u_h = A_\varepsilon(\widehat{C}_h)\widehat{m}_h$ be the canonical Adam direction and choose only a numerical reporting floor $\epsilon_{\mathrm{cert}} > 0$. Define

$$\chi_h := \frac{2\left[\sum_b L_{\varepsilon,b}^2 \|A_b\|_{\mathrm{op}}^2 \|\widehat{m}_{h,b}\|_2^2\right]^{1/2}}{\|u_h\|_2 + \epsilon_{\mathrm{cert}}}. \tag{106}$$

When a positive rotation-uniform floor is unavailable, Proposition 8 gives the fallback certificate

$$\chi_h^{\mathrm{H}} := \frac{2\varepsilon^{-2}\left[\sum_b \|A_b\|_{\mathrm{op}} \|\widehat{m}_{h,b}\|_2^2\right]^{1/2}}{\|u_h\|_2 + \epsilon_{\mathrm{cert}}}. \tag{107}$$

Propositions 7 and 8 give the corresponding relative upper bounds. Thus "negligible at step h up to tolerance τ" is certified when a mathematically valid certificate is at most τ; if both apply, use their minimum. The multi-step statement additionally requires the stability factor in Proposition 10. These are regularized ratios: when $\|u_h\|$ is close to zero, a small value relative to $\|u_h\| + \epsilon_{\mathrm{cert}}$ is not a small unregularized relative error. Report the absolute defect too. Empirical optimizer and fixed-history diagnostics are reported in Appendix J.6.

E.5 Expectation and probability over random rotations

Proposition 9 (Haar-rotation expectation and probability)**.** *Condition on a key-independent common gradient history, and let each marginal T_b be Haar on $O(q_b)$. Blocks may share the same rotation when required by an architectural key. Under a floor valid uniformly over the sampled rotations, define*

$$B_h^{\mathrm{lin}} := 2\sum_b L_{\varepsilon,b}^2 \|\widehat{m}_{h,b}\|_2^2 \left(\|\mathrm{diag}(A_b)\|_2^2 + \frac{2}{q_b + 2}\|A_b\|_F^2\right). \tag{108}$$

Without any positive floor, define

$$B_h^{\mathrm{H}} := \frac{2}{\varepsilon^4}\sum_b \|\widehat{m}_{h,b}\|_2^2 \left(\|\mathrm{diag}(A_b)\|_1 + \sqrt{\frac{2q_b}{q_b + 2}}\|A_b\|_F\right). \tag{109}$$

Then $\mathbb{E}_T \|q_h(T)\|_2^2$ is bounded by every applicable quantity in $\{B_h^{\mathrm{lin}}, B_h^{\mathrm{H}}\}$. Let B_h be their minimum when both apply, and otherwise the valid one. For every $\delta \in (0, 1)$,

$$\Pr_T\left[\|q_h(T)\|_2 > \sqrt{B_h/\delta}\right] \leq \delta. \tag{110}$$

The same diagonal-moment identities hold for a uniformly random planar angle when $q_b = 2$.

Proof. For a traceless symmetric $A \in \mathbb{R}^{q\times q}$ and a uniformly random unit vector u, $\mathbb{E}(u^\top Au)^2 = 2\|A\|_F^2/[q(q+2)]$. Summing over the q rows of a Haar matrix gives

$$\mathbb{E}_T \left\|\mathrm{diag}(TAT^\top)\right\|_2^2 = \frac{2}{q+2}\|A\|_F^2 . \tag{111}$$

Under a floor, subtract the isotropic scalar preconditioner and apply scalar Lipschitzness. The inequality $\|u-v\|^2 \le 2\|u\|^2 + 2\|v\|^2$, followed by $\|Dx\| \le \|D\|_F \|x\|$, yields Eq. (108).

For the floor-free bound, Hölder continuity gives $\|D_T\|_{\mathrm{op}}^2 \le \varepsilon^{-4}\left\|\mathrm{diag}(TAT^\top)\right\|_1$ and $\|D_0\|_{\mathrm{op}}^2 \le \varepsilon^{-4}\|\mathrm{diag}(A)\|_1$. Cauchy–Schwarz and Eq. (111) imply

$$\mathbb{E}_T \left\|\mathrm{diag}(TAT^\top)\right\|_1 \le \sqrt{q\,\mathbb{E}_T\|\mathrm{diag}(TAT^\top)\|_2^2} = \sqrt{\frac{2q}{q+2}}\|A\|_F . \tag{112}$$

Applying the same two-term inequality proves Eq. (109). Markov's inequality applied to $\|q\|^2$ proves Eq. (110); when $B_h = 0$, the defect is zero almost surely and the strict-threshold event is empty. □

These expectations are conditional and counterfactual: they isolate one optimizer step at a shared history and do not assume that two dense-rotation Adam trajectories remain aligned automatically. The floor-free expression remains valid for rank-deficient early moments, but its ε^{-4} dependence can make it numerically vacuous. For RoPE the natural blocks are often two-dimensional, so $2/(q_b+2) = 1/2$; there is no strong high-dimensional averaging effect, and empirical calibration is essential.

E.6 A STOCHASTIC-GRADIENT CONDITION UNDER WHICH THE DEFECT SHRINKS

Let $X_s = g_s g_s^\top - \sigma_b^2 I$ in block b. Assume (X_s) is a matrix martingale-difference sequence with

$$\mathbb{E}[X_s \mid \mathcal{F}_{s-1}] = 0, \qquad \mathbb{E}[\|X_s\|_F^2 \mid \mathcal{F}_{s-1}] \le \tau_b^2, \tag{113}$$

and assume the deterministic bounds $\|\widehat{m}_{h,b}\| \le M_b$ and a fixed positive, key-uniform denominator floor as above. The normalized EMA weights in Eq. (95) satisfy

$$\sum_{s=1}^{h} a_{h,s}^2 = \frac{1}{n_{\mathrm{eff}}(h,\beta_2)}, \qquad n_{\mathrm{eff}}(h,\beta_2) = \frac{1+\beta_2}{1-\beta_2}\frac{1-\beta_2^h}{1+\beta_2^h}. \tag{114}$$

Martingale orthogonality and projection onto the traceless subspace imply

$$\mathbb{E}\|A_b\|_F^2 \le \frac{\tau_b^2}{n_{\mathrm{eff}}(h,\beta_2)}. \tag{115}$$

Combining this with Eq. (103) gives the conservative root-mean-square control

$$\left(\mathbb{E}\|q_h(T)\|_2^2\right)^{1/2} \le 2\left[\sum_b \frac{L_{\varepsilon,b}^2 M_b^2 \tau_b^2}{n_{\mathrm{eff}}(h,\beta_2)}\right]^{1/2}. \tag{116}$$

Thus the intrinsic defect is $O(n_{\mathrm{eff}}^{-1/2})$ under the stated strong block-isotropic model and a positive denominator floor. Without that floor, Proposition 8, Jensen's inequality, and Eq. (115) give the weaker but always valid control

$$\left(\mathbb{E}\|q_h(T)\|_2^2\right)^{1/2} \le \frac{2}{\varepsilon^2}\left[\sum_b \frac{M_b^2\tau_b}{\sqrt{n_{\mathrm{eff}}(h,\beta_2)}}\right]^{1/2}, \tag{117}$$

which scales only as $O(n_{\mathrm{eff}}^{-1/4})$ and is normally loose for practical Adam ε. For fixed $\beta_2 < 1$, n_{eff} saturates at $(1+\beta_2)/(1-\beta_2)$, so increasing the horizon alone does not force these bounds to zero. Directional unlearning gradients, nonstationarity, heavy tails, or very short local runs can invalidate the isotropic model; Eqs. (116) and (117) are sufficient controls, not assumptions made about the reported experiments.

E.7 From one-step defects to complete Adam trajectories

A dense rotation cannot be pulled back on Adam's *diagonal* second-moment vector alone: the missing cross moments are needed to form $\operatorname{diag}(TCT^\top)$. A valid common-state comparison instead uses an analytical lift, not a modification of the implemented optimizer. With fresh zero states, let

$$x_h = (\theta_h, m_h, C_h), \qquad C_h = \beta_2 C_{h-1} + (1-\beta_2) g_h g_h^\top. \tag{118}$$

The canonical algorithm uses $\operatorname{diag} C_h$; the pulled-back rotated algorithm uses $\operatorname{diag}(TC_hT^\top)$, with C_h evaluated along *its own* pulled-back gradient history. Both lifted algorithms update (m, C) by the same recurrences at a common state, and differ only in the parameter direction there. The lift exactly represents standard Adam with zero state because taking its diagonal recovers the implemented recurrence. Nonzero initialization requires a compatible full-moment lift, not merely an arbitrary pair of diagonal states. The lift is for proof; implementations need not store a $d \times d$ matrix.

Let Φ_h and $\widetilde{\Phi}_{h,T}$ be the resulting canonical and pulled-back one-step maps, with the same covariant loss, minibatch, and auxiliary construction. Here $h = 0, \dots, H-1$ indexes transitions: $q_h(T; x)$ uses the newly updated, bias-corrected moments for transition $h+1$, not the incoming state moments. Use a weighted product norm on (θ, m, C) with positive weights and unit parameter weight. At a common state their discrepancy is

$$\left\| \widetilde{\Phi}_{h,T}(x) - \Phi_h(x) \right\| = \eta_h \left\| q_h(T; x) \right\|_2 . \tag{119}$$

Let x_h follow the canonical trajectory, which is independent of the key. Assume that throughout a region containing both trajectories, and *uniformly over admissible keys*,

$$\left\| \widetilde{\Phi}_{h,T}(x) - \widetilde{\Phi}_{h,T}(y) \right\| \le K_h \left\| x - y \right\|, \qquad K_h \ge 0. \tag{120}$$

Unlike a parameter-only gradient step, an augmented Adam step is not automatically $(1 + O(\eta_h))$-Lipschitz: its moment recurrences also change when the gradient changes.

Proposition 10 (Lifted-state stability propagation)**.** *Starting from identical lifted states, under Eq. (120),*

$$\left\| \widetilde{x}_H - x_H \right\| \le \sum_{h=0}^{H-1} \eta_h \left\| q_h(T; x_h) \right\|_2 \prod_{j=h+1}^{H-1} K_j. \tag{121}$$

If B_h bounds $\mathbb{E}_T \left\| q_h(T; x_h) \right\|_2^2$ on the canonical histories and the K_h are deterministic, then

$$\left(\mathbb{E}_T \left\| \widetilde{x}_H - x_H \right\|^2 \right)^{1/2} \le \sum_{h=0}^{H-1} \eta_h \sqrt{B_h} \prod_{j=h+1}^{H-1} K_j =: M_H. \tag{122}$$

Thus $\Pr(\left\| \widetilde{x}_H - x_H \right\| > a) \le \min\{1, M_H^2/a^2\}$ for $a > 0$. These bounds also control the parameter displacement error.

Proof. Insert $\widetilde{\Phi}_{h,T}(x_h)$ between the two next states. The key-uniform Lipschitz bound gives $E_{h+1} \le K_h E_h + \eta_h \left\| q_h(T; x_h) \right\|$, evaluated on the *canonical* history. Unroll from $E_0 = 0$ to obtain Eq. (121). Minkowski's inequality gives Eq. (122); no independence between steps is needed. Markov's inequality gives the probability statement. Because the canonical history is independent of T, Proposition 9 can supply B_h when its floor or floor-free conditions hold. The key-uniform trajectory stability condition remains an additional requirement. □

Sufficient conditions for finite stability constants. On a region where the minibatch gradient is L_g-Lipschitz and bounded by G, differences after updating uncorrected moments satisfy

$$\delta_m^+ \le \beta_1 \delta_m + (1-\beta_1) L_g \delta_\theta, \tag{123}$$

$$\delta_C^+ \le \beta_2 \delta_C + 2(1-\beta_2) G L_g \delta_\theta, \tag{124}$$

where δ_C uses the Frobenius norm. If corrected adaptive diagonals have a key-uniform lower bound $\nu > 0$ and corrected first moments are bounded by M, then

$$\delta_\theta^+ \le |1 - \eta_h \lambda_{\mathrm{wd}}| \delta_\theta + \eta_h a_h \delta_m^+ + \eta_h b_h \delta_C^+, \tag{125}$$

with $a_h = [(1-\beta_1^{h+1})(\sqrt{\nu}+\varepsilon)]^{-1}$ and $b_h = L_{\varepsilon,\nu} M/(1-\beta_2^{h+1})$. The first two inequalities follow from the moment recurrences and $\left\|gg^\top - \widetilde{g}\widetilde{g}^\top\right\|_F \le 2G\,\|g-\widetilde{g}\|$. The last follows from the state-mismatch argument; rotation preserves the relevant norms. The induced norm of this nonnegative three-variable recurrence supplies a valid K_h. Only when one separately proves $K_h \le 1+\eta_h L_h$ with $L_h \ge 0$ may the products be replaced by $\exp(\sum_j \eta_j L_j)$.

If the region is valid only on a high-probability event, its failure probability must be added to the trajectory bound. Evaluating a fixed-history rotation expectation on the already diverged, key-dependent history without this construction is not justified. Similarly, reporting one small direction defect does not establish that the product of stability factors is small over a long local run.

E.8 Finite precision and numerical implementation

Exact safe-key equivariance is an exact-arithmetic identity, not a probability law for GPU discrepancies. If the two implemented updates have extra per-step errors bounded by ξ_h, Eq. (121) acquires ξ_h in addition to $\eta_h\,\|q_h\|$. No distribution-free numerical failure probability is asserted. Floating-point failures of privacy mechanisms are a separate issue (Mironov, 2012).

Quantization versus finite-precision sampling. A fixed, checkpoint-independent rounding map applied to an *ideal* Gaussian release is post-processing and preserves its RDP upper bound, though not necessarily the exact divergence equality. Sampling approximate noise and adding it in floating point is not automatically the same mechanism. If its output distribution is uniformly within total variation τ of such an ideal post-processed mechanism, then an ideal (ϵ,δ) guarantee implies $(\epsilon, \delta+(1+e^\epsilon)\tau)$-DP for the implementation. Indeed, for any event, replacing each of the two distributions in the DP inequality changes the bound by at most τ and $e^\epsilon\tau$, respectively. No such τ is certified for the supplied GPU implementation; the formal Gaussian theorem is therefore an ideal-mechanism statement.

If floating-point coding leaves $r_{\mathrm{sum}} = \sum_k e_k \neq 0$, the canceled first-order term is instead $J_F(\theta) r_{\mathrm{sum}}/(\sum_k \alpha_k^{-1})$. Its norm is bounded by $\|J_F(\theta)\|_{\mathrm{op}}\,\|r_{\mathrm{sum}}\|\,/(\sum_k \alpha_k^{-1})$, in addition to curvature and inverse-key numerical errors. This residual can be measured without additional model training.

F Full-View Recovery and Alignment Audit

If all executors collude, an adversary observes $\theta_{\mathrm{pub}}^{(k)} = T_k(\theta + \alpha_k e_k)$ and may solve

$$\min_{\widehat{T}_1,\dots,\widehat{T}_m \in \mathcal{G}} \sum_{i<j} \left\| \widehat{T}_i^{-1}\theta_{\mathrm{pub}}^{(i)} - \widehat{T}_j^{-1}\theta_{\mathrm{pub}}^{(j)} \right\|_F^2 . \tag{126}$$

Pairwise FFN permutation matching can reduce to linear assignment, and two-set orthogonal Procrustes has an SVD solution. Unrestricted multi-index assignment and generalized Procrustes problems have worst-case hardness results (Burkard et al., 2012; Ling, 2021), but no reduction from those problems to the exact architecture-constrained key family in Eq. (126) is established here. We therefore make no NP-hardness claim for this recovery problem. Even such a reduction would not prove average-case security, exclude approximate recovery, or prevent output-space distillation. They are therefore used only to motivate the empirical full-view audit in Appendix J.9; formal RDP intentionally ends at $s = m$.

G Streaming Memory and Computation

The server needs only

$$S_0 = \sum_{k=1}^{m} \alpha_k^{-1}, \qquad S_1 = \sum_{k=1}^{m} \alpha_k^{-1}\widehat{\Delta}_k, \tag{127}$$

so it can receive, invert, accumulate, and discard one view at a time. To generate the centered Gaussian code without storing all copies, a first pass accumulates $\bar{z}$ using regeneratable private random streams, and a second pass regenerates each z_k and forms e_k. This needs a mean-noise

buffer as well as the current view and update accumulator: $O(d)$ auxiliary storage with a constant number of model-sized tensors, not exactly two tensors. It avoids m simultaneously resident models. Communication and local optimization remain linear in m. This memory optimization does not itself enforce view isolation: a stateful sequential worker that can retain every same-round view violates the threat model.

H Experimental Details and Complete Results

The historical tables report the seed-0 benchmark experiments. Clean, Noised and DARC correspond to the noise-free DIRECT reference, SINGLE-NOISED control and DARC, respectively. They use the historical task-vector RMS profile, hardware and metric conventions; repeated entries reuse the same observations. MUSE changes depend on method and domain.

Table 4: Complete Llama-1B/3B MUSE-Books and News measurements, six trainers, $m = 2$, $\kappa = .01$, seed 0. DIRECT is a noise-free reference. ES: extraction strength; F-KM/R-KM: forget/retain KnowMem ROUGE; PL: reported PrivLeak scale. Bold is the original descriptive Direct/DARC comparison; it does not denote statistical significance.

		Llama-3.2-1B-Instruct				**Llama-3.2-3B-Instruct**			
Algorithm	Setting	ES ↓	F-KM ↓	R-KM ↑	PL → 0	ES ↓	F-KM ↓	R-KM ↑	PL → 0
MUSE-Books									
GradAscent	Direct	**0.00915**	**0**	**0**	**-32.0**	**0.00915**	**0**	**0**	**52.3**
	DARC	**0.00915**	**0**	**0**	-34.5	**0.00915**	**0**	**0**	56.6
GradDiff	Direct	**0.00915**	0.0455	**0.146**	**28.9**	**0.00915**	**0**	0.115	**7.42**
	DARC	**0.00915**	**0.0331**	0.141	31.7	**0.00915**	**0**	**0.146**	35.4
NPO	Direct	**0.0125**	**0.172**	0.188	**-23.7**	0.0155	0.234	0.313	**-71.9**
	DARC	0.0127	0.181	**0.191**	-24.0	**0.015**	**0.226**	**0.333**	-73.2
SimNPO	Direct	**0.0122**	0.161	0.206	-30.1	0.0143	0.237	**0.32**	-84.2
	DARC	0.0125	**0.152**	**0.218**	**-28.5**	**0.0141**	**0.227**	0.316	**-83.6**
UnDIAL	Direct	**0.0112**	**0.167**	**0.204**	**-56.9**	0.0109	0.243	**0.359**	**-89.6**
	DARC	**0.0112**	0.183	0.195	-57.2	**0.0106**	**0.229**	0.316	**-89.6**
SatImp	Direct	**0.00924**	0.152	**0.202**	-2.22	**0.00915**	**0.221**	**0.315**	**-28.8**
	DARC	**0.00924**	**0.147**	0.198	**-1.87**	**0.00915**	0.239	0.3	-29.8
MUSE-News									
GradAscent	Direct	**0.00975**	**0**	**0**	19.3	**0.00975**	**0**	**0**	**14.2**
	DARC	**0.00975**	**0**	**0**	**19.2**	**0.00975**	**0**	**0**	15.4
GradDiff	Direct	**0.00985**	0.0381	**0.0292**	**79.7**	0.0142	**0**	0	**64.1**
	DARC	0.01	**0.0346**	0.0123	82.9	**0.0132**	**0**	**0.026**	68.3
NPO	Direct	**0.0279**	**0.163**	**0.178**	-28.0	**0.0271**	0.234	**0.255**	**-64.5**
	DARC	0.028	0.167	**0.178**	**-27.5**	0.0283	**0.213**	0.233	-64.9
SimNPO	Direct	0.0282	**0.155**	**0.181**	**-52.0**	**0.0282**	0.236	0.213	-82.7
	DARC	**0.0279**	0.169	0.175	**-52.0**	0.0288	**0.231**	**0.216**	**-82.6**
UnDIAL	Direct	**0.016**	**0.136**	**0.21**	**-49.0**	0.0136	**0.236**	**0.276**	**-81.6**
	DARC	**0.016**	0.137	0.207	**-49.0**	**0.0134**	0.238	0.256	**-81.6**
SatImp	Direct	0.0281	0.173	**0.179**	-52.2	0.0288	**0.224**	**0.213**	**-82.5**
	DARC	**0.028**	**0.17**	0.176	**-52.1**	**0.0278**	0.232	0.212	-82.8

H.1 Benchmarks

TOFU (Task of Fictitious Unlearning), introduced by Maini et al. (2024), is a question–answer (QA) benchmark for evaluating unlearning in large language models. The dataset comprises QA pairs derived from synthetic autobiographies of 200 fictitious authors, with all content generated by GPT-4 to ensure exclusion from the pretraining corpora of existing LLMs. Each author profile contains 20 QA pairs covering biographical attributes such as name, birthplace, gender, birth year, literary genre, awards, and parental occupations, with book titles seeded from the Goodreads Books dataset to enhance topical diversity.

MUSE (Machine Unlearning Six-Way Evaluation) (Shi et al., 2025) evaluates language-model unlearning against six criteria relevant to data owners and model deployers. From the data owner's perspective, it measures verbatim memorization, knowledge memorization, and privacy leakage (assessed via membership inference attacks). For deployers, it measures retained model utility, scaling to large forget sets, and performance under sequential unlearning requests. Evaluated on unlearning Harry Potter books and news articles, MUSE reveals that while current algorithms can reduce memorization, they struggle significantly with preventing privacy leakage, maintaining model utility, and scaling effectively.

H.2 METRICS

The metrics cover memorization, privacy leakage, and retained utility.

H.2.1 MEMORIZATION METRICS

These metrics measure the degree to which information from the training data remains encoded in the model after unlearning.

Probability We quantify the model's confidence in generating correct answers by measuring the conditional probability $P(a \mid q)$ assigned to the ground-truth answer a given the question q, evaluated on **Retain Set**. The resulting score is reported as a normalized probability in $[0, 1]$. Following standard practice (Cho et al., 2014), we normalize for answer length by exponentiating the sequence probability by $1/|a|$, as formalized in Equation 128:

$$P_{\text{retain}}(x) = P(a \mid q)^{1/|a|}. \tag{128}$$

For the **Real Authors** and **World Facts** subsets, we evaluate model performance using a relative probability formulation in a multiple-choice setting:

$$P_{\text{real/world}}(x) = \frac{P(a_1 \mid q)}{\sum_{i=1}^{n} P(a_i \mid q)}, \tag{129}$$

where q denotes a multiple-choice question with candidate answers $\{a_1, \ldots, a_n\}$, and a_1 corresponds to the ground-truth correct answer. The score is the share of candidate-answer probability assigned to the correct option.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) We measure overlap between generated and ground-truth answers with ROUGE-L recall (Lin, 2004), which measures similarity via the length of the longest common subsequence (LCS). This metric provides a robust estimate of answer correctness in QA settings by tolerating minor paraphrasing and surface-form variations.

$$\text{ROUGE}_L(x) = \frac{\text{LCS}(a, \hat{a})}{|a|}, \tag{130}$$

where $\hat{a}$ is the generated answer, and $\text{LCS}(a, \hat{a})$ is the LCS length between a and $\hat{a}$.

Truth Ratio For a given question, we define the *truth ratio* R_{Truth} to approximate the relative likelihood assigned by the model to correct versus incorrect answers. Since the model is fine-tuned on a specific canonical phrasing of the ground-truth answer, the corresponding probability may be artificially inflated compared to alternative but semantically equivalent formulations. We therefore evaluate a paraphrase of the correct answer instead of its original wording.

For the incorrect-answer comparison, we average the probabilities of syntactically similar answers containing factual errors. This design yields a more stable and representative estimate of the model's preference for incorrect information.

Formally, let $\tilde{a}$ denote a paraphrased correct answer to question q, and let $\tilde{x} = [q, \tilde{a}]$. Let A_{err} be a set of incorrect answers $\{a_{\text{err}}\}$, constructed by preserving the general textual structure of $\tilde{a}$ while introducing factual errors. The truth ratio R_{Truth} is then defined as:

$$R_{\text{truth}}(x) = \frac{\frac{1}{|A_{\text{err}}|}\sum_{a_{\text{err}} \in A_{\text{err}}} P(a_{\text{err}} \mid q)^{1/|a_{\text{err}}|}}{P(\tilde{a} \mid q)^{1/|\tilde{a}|}}. \tag{131}$$

Equation 132 rescales the metric to $[0, 1]$, with larger values indicating better model performance.

$$R_{\text{adjusted}}(x) = \max(0, 1 - R_{\text{Truth}}(x)). \tag{132}$$

Verbatim Memorization (Data Extraction) To evaluate whether a model inappropriately extracts or replicates exact details from a removed dataset, the MUSE benchmark utilizes the Verbatim Memorization ($VerbMem$) metric. This is measured by prompting the model f with the first l tokens of a sequence from the forget set, denoted as $x_{[:l]}$. The model's generated continuation, $f(x_{[:l]})$, is then compared to the actual ground-truth continuation, $x_{[l+1:]}$, using the ROUGE-L F1 score. The overall metric is calculated by averaging these scores across the entire forget set, formalized as:

$$VerbMem(f, \mathcal{D}) := \frac{1}{|\mathcal{D}_{forget}|} \sum_{x \in \mathcal{D}_{forget}} ROUGE(f(x_{[:l]}), x_{[l+1:]})$$

H.2.2 Privacy Metrics

These metrics evaluate whether sensitive information from the forget set can still be inferred or extracted from the model after unlearning. Such metrics often assume perfectly i.i.d. holdout samples or access to an oracle retain model. These assumptions may not hold in deployment.

Membership Inference Leakage We evaluate final-model membership leakage through membership inference attacks (MIAs), which assess a model's tendency to memorize training data. An MIA tests whether an adversary can distinguish examples drawn from the forget set $\mathcal{D}_{\text{forget}}$ (members) and unseen examples from a holdout set $\mathcal{D}_{\text{holdout}}$ (non-members), based on model confidence or loss statistics.

Ideally, a model that has not been trained on $\mathcal{D}_{\text{forget}}$ should yield an AUC of 0.5, indicating indistinguishability between member and non-member samples. In practice, however, constructing perfectly i.i.d. holdout splits is challenging, and even retrained models may exhibit nontrivial membership signals. Following MUSE, we calibrate the membership-inference AUC against a retrained reference model instead of interpreting the absolute AUC alone.

Unlearning generally increases the loss on forgotten samples, but privacy leakage may still arise in two failure modes: ***(i)* under-unlearning**, where the loss increase is insufficient, and membership information remains detectable; and ***(ii)* over-unlearning**, where the loss becomes abnormally large, again yielding a distinguishable signal. Both cases induce separable loss distributions between $\mathcal{D}_{\text{forget}}$ and $\mathcal{D}_{\text{holdout}}$.

To quantify this effect, we compare the AUC-ROC achieved by the unlearned model f_{unlearn} against that of a retrained reference model f_{retrain}, and define the relative privacy leakage as

$$\text{PrivLeak}_{\text{ratio}} := \frac{\text{AUC}(f_{\text{Unlearn}}; \mathcal{D}_{\text{Forget}}, \mathcal{D}_{\text{Holdout}}) - \text{AUC}(f_{\text{Retrain}}; \mathcal{D}_{\text{Forget}}, \mathcal{D}_{\text{Holdout}})}{\text{AUC}(f_{\text{Retrain}}; \mathcal{D}_{\text{Forget}}, \mathcal{D}_{\text{Holdout}})}. \tag{133}$$

Legacy PrivLeak scale. A well-behaved unlearning algorithm should yield a PrivLeak value close to zero, indicating privacy leakage comparable to retraining. In contrast, under-unlearning and over-unlearning result in large negative and positive deviations, respectively, reflecting increased membership distinguishability.

Forget Quality Following TOFU, Forget Quality compares the distribution of Truth Ratio values produced by the unlearned model with the distribution produced by a retain-only reference model, both evaluated on the forget set. Let $F_U(x)$ and $F_R(x)$ be the corresponding empirical CDFs from n and m examples. The two-sample Kolmogorov–Smirnov statistic is

$$D_{n,m} = \sup_x |F_U(x) - F_R(x)|. \tag{134}$$

We retain the benchmark's nominal two-sample KS p-value as a descriptive comparison statistic. Standard KS null calibration assumes independent samples from continuous distributions. Here both models are evaluated on the same questions, so the observations are paired. For the independent continuous-sample reference, a common approximate asymptotic cutoff at nominal level $\alpha \in (0, 1)$ is

$$d_{\text{approx}}(\alpha; n, m) = c_{\text{approx}}(\alpha)\sqrt{\frac{n+m}{nm}}, \qquad c_{\text{approx}}(\alpha) = \sqrt{-\frac{1}{2}\ln\Big(\frac{\alpha}{2}\Big)}. \tag{135}$$

This cutoff is not the exact asymptotic KS quantile and is not equivalent to a finite-sample exact p-value. Larger nominal FQ is favored only under the benchmark's descriptive convention. We claim neither calibrated significance for this paired protocol nor evidence of distributional equivalence, data deletion, or differential privacy; reported FQ values are unchanged.

H.2.3 Utility Metrics

The objective of unlearning is to effectively remove the influence of targeted data while preserving the model's performance on non-forget data. Utility metrics evaluate whether the unlearned model maintains its capabilities on tasks beyond the retain set, thereby ensuring that unlearning does not degrade general performance on real-world data distributions.

Model Utility Model Utility (MU) measures the retained performance of a model after unlearning, covering both the closely related retain set and broader general-knowledge tasks.

Following the TOFU benchmark protocol, MU is computed as the harmonic mean of nine metrics spanning three data levels: **Retain Set**, **Real Authors**, and **World Facts**. At each level, three metrics are evaluated: **Probability**, **ROUGE**, and **Truth Ratio**, to ensure balanced assessment across memorization, semantic accuracy, and factual correctness.

$$U_{\text{model}} = \frac{9}{\sum_{m \in M} \frac{1}{m}}, \tag{136}$$

where M denotes the set of all evaluated metrics.

H.3 Unlearning Algorithms

We use the following seven unlearning objectives within DARC.

Notation Let $\mathcal{D}_{\text{forget}}$ and $\mathcal{D}_{\text{retain}}$ denote the forget and retain sets, respectively, and let f_θ be the model being updated during unlearning. For an input–output (prompt–completion) pair (x, y) under a causal LM, we define

$$\log p_\theta(y \mid x) = \sum_{t=1}^{|y|} \log p_\theta(y_t \mid y_{<t}, x), \quad \ell_{\text{CE}}(y \mid x; f_\theta) = -\log p_\theta(y \mid x). \tag{137}$$

Unless otherwise stated, we optionally include a retain regularizer:

$$\mathcal{L}_{\text{retain}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{retain}}} \ell_{\text{CE}}(y \mid x; f_\theta), \tag{138}$$

and optimize $\min_\theta \ \mathcal{L}_{\text{forget}}(\theta) + \alpha\, \mathcal{L}_{\text{retain}}(\theta)$, where $\alpha \geq 0$ controls the forgetting–retention trade-off.

H.3.1 GradAscent (Jang et al., 2023)

Gradient Ascent (GradAscent) directly "reverses" standard likelihood training on the forget set by maximizing the cross-entropy (equivalently, minimizing the log-likelihood). Under the minimization convention, the forget objective is

$$\mathcal{L}^{\text{GA}}_{\text{forget}}(\theta) = -\mathbb{E}_{(x,y_{\text{f}})\sim\mathcal{D}_{\text{forget}}} \ell_{\text{CE}}(y_{\text{f}} \mid x; f_\theta) = \mathbb{E}_{(x,y_{\text{f}})\sim\mathcal{D}_{\text{forget}}} \log p_\theta(y_{\text{f}} \mid x). \tag{139}$$

This objective is effective at reducing the model likelihood on targeted samples, but without an explicit retention constraint, it may cause collateral degradation on non-forgotten behavior.

H.3.2 GRADDIFF (LIU ET AL., 2022A)

GradDiff augments GradAscent with a retain loss to explicitly preserve utility on non-forgotten data. A common instantiation writes the full objective as

$$\min_{\theta} \mathcal{L}^{\mathrm{GradDiff}}(\theta) = -\lambda_{\mathrm{f}} \mathbb{E}_{(x,y_{\mathrm{f}})\sim\mathcal{D}_{\mathrm{forget}}} \ell_{\mathrm{CE}}(y_{\mathrm{f}} \mid x; f_\theta) + \lambda_{\mathrm{r}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{retain}}} \ell_{\mathrm{CE}}(y \mid x; f_\theta), \quad (140)$$

where $\lambda_{\mathrm{f}}, \lambda_{\mathrm{r}} > 0$ trade off forgetting strength and retention fidelity.

H.3.3 DPO (RAFAILOV ET AL., 2023)

Direct Preference Optimization (DPO) is a preference-learning objective originally proposed for alignment. Given a preference dataset $\mathcal{D}_{\mathrm{pref}}$ of triples (x, y_w, y_l) (preferred y_w vs. dispreferred y_l) and a reference model f_{ref}, DPO optimizes

$$\mathcal{L}^{\mathrm{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\mathrm{pref}}} \log \sigma\left(\beta \log \frac{p(y_w \mid x; f_\theta)}{p(y_w \mid x; f_{\mathrm{ref}})} - \beta \log \frac{p(y_l \mid x; f_\theta)}{p(y_l \mid x; f_{\mathrm{ref}})}\right), \quad (141)$$

where $\beta > 0$ controls the sharpness of preference separation. In unlearning-style instantiations, one can construct preference pairs so that generations containing forget targets are treated as dispreferred.

H.3.4 NPO (ZHANG ET AL., 2024)

Negative Preference Optimization (NPO) reformulates unlearning as a bounded, alignment-inspired objective that discourages the forget targets relative to a frozen reference model. Let f_{ref} denote the reference model (typically the pre-unlearning checkpoint) and $\sigma(\cdot)$ the sigmoid. The NPO's forget loss is

$$\mathcal{L}^{\mathrm{NPO}}_{\mathrm{forget}}(\theta) = -\frac{2}{\beta} \mathbb{E}_{(x,y_{\mathrm{f}})\sim\mathcal{D}_{\mathrm{forget}}} \log \sigma\left(-\beta \log \frac{p(y_{\mathrm{f}} \mid x; f_\theta)}{p(y_{\mathrm{f}} \mid x; f_{\mathrm{ref}})}\right), \quad (142)$$

where $\beta > 0$ is a temperature parameter. In practice, NPO is often combined with the retain regularizer $\alpha\, \mathcal{L}_{\mathrm{retain}}(\theta)$ to preserve utility.

H.3.5 SIMNPO (FAN ET AL., 2024)

SimNPO removes the explicit reference model and introduces a margin/offset term (denoted γ in the original paper) while retaining the stabilized log-sigmoid form:

$$\mathcal{L}^{\mathrm{SimNPO}}_{\mathrm{forget}}(\theta) = -\frac{2}{\beta} \mathbb{E}_{(x,y_{\mathrm{f}})\sim\mathcal{D}_{\mathrm{forget}}} \log \sigma\left(-\frac{\beta}{|y_{\mathrm{f}}|} \log p(y_{\mathrm{f}} \mid x; f_\theta) - \gamma\right), \quad (143)$$

where $|y_{\mathrm{f}}|$ is the target sequence length (used for normalization), and $\gamma \geq 0$ calibrates the separation threshold. As with other unlearning objectives, SimNPO is commonly paired with $\alpha\, \mathcal{L}_{\mathrm{retain}}(\theta)$.

H.3.6 UNDIAL (DONG ET AL., 2025)

UnDIAL (Unlearning via Self-Distillation on Adjusted Logits) stabilizes unlearning by defining an explicit, fixed target distribution and distilling the model toward it. Let f_{orig} be the frozen pre-unlearning model. For a forget example (x, y_{f}) and each position t, let $\mathbf{z}_t^{\mathrm{orig}}$ be the teacher logits and $\mathbf{e}_{y_{\mathrm{f},t}}$ the one-hot vector of the target token. UnDIAL constructs adjusted logits and a target distribution:

$$\mathbf{z}_t^{\mathrm{adj}} = \mathbf{z}_t^{\mathrm{orig}} - \gamma_{\mathrm{UD}} \mathbf{e}_{y_{\mathrm{f},t}}, \quad \mathbf{p}_t^{\mathrm{adj}} = \mathrm{softmax}\left(\mathbf{z}_t^{\mathrm{adj}}\right), \quad (144)$$

where $\gamma_{\mathrm{UD}} > 0$ controls the strength of demoting the memorized token. The self-distillation (cross-entropy) unlearning objective is

$$\mathcal{L}^{\mathrm{UnDIAL}}_{\mathrm{forget}}(\theta) = \mathbb{E}_{(x,y_{\mathrm{f}})\sim\mathcal{D}_{\mathrm{forget}}} \left[\sum_{t=1}^{|y_{\mathrm{f}}|} \mathcal{H}\left(\mathbf{p}_t^{\mathrm{adj}},\ p(\cdot \mid x, y_{\mathrm{f},<t}; f_\theta)\right)\right], \quad (145)$$

where $\mathcal{H}(\cdot,\cdot)$ is the cross-entropy between the fixed adjusted distribution and the student model's predictive distribution. Optionally, UnDIAL can also be combined with $\alpha\, \mathcal{L}_{\mathrm{retain}}(\theta)$ for utility preservation.

H.3.7 SATIMP (YANG ET AL., 2025)

Saturated Importance (SatImp) is a token-wise soft reweighting strategy for enhancing unlearning. At iteration h, with current parameters θ_h, it uses the surrogate

$$\mathcal{L}_{\text{forget},h}(\theta) = \mathbb{E}_{(x,y_\text{f})\sim\mathcal{D}_\text{forget}} \sum_{k=1}^{|y_\text{f}|} w_{x,y_\text{f},k}^{(h)} \log p(y_{\text{f},k} \mid y_{\text{f},<k}, x; f_\theta)\,. \tag{146}$$

Writing $p_{h,k} = p(y_{\text{f},k} \mid y_{\text{f},<k}, x; f_{\theta_h})$, the weights are

$$w_{x,y_\text{f},k}^{(h)} = p_{h,k}^{\beta_1}(1-p_{h,k})^{\beta_2}, \qquad \beta_1, \beta_2 \geq 0. \tag{147}$$

The weights are held constant when taking $g_h = \nabla_\theta \mathcal{L}_{\text{forget},h}(\theta)|_{\theta=\theta_h}$: no gradient propagates through their construction. They are recomputed at each iteration, so the complete training map still depends on the weight rule. The token reduction above and the optional retain term $\alpha\mathcal{L}_\text{retain}(\theta)$ are unchanged. This specifies the differentiation convention; it does not establish the historical run-to-code binding.

H.4 EXPERIMENTAL SETTINGS

H.4.1 TESTBED

Hardware Configuration The inherited benchmark experiments were conducted on two Elastic Compute Service (ECS) instances.

For experiments using **Llama-3.2-1B-Instruct** and **Qwen2.5-1.5B-Instruct**, we employ an instance equipped with an Intel Xeon Gold 6462C CPU (16 available cores), 128 GB RAM, 512 GB of available disk space, and an NVIDIA L20 GPU with 48 GB memory. For experiments using **Llama-3.2-3B-Instruct**, **Llama-3.1-8B-Instruct**, and **Qwen2.5-3B-Instruct**, we use an instance equipped with an Intel Xeon Platinum 8469C CPU (24 available cores), 128 GB RAM, 512 GB of available disk space, and an NVIDIA HGX H20 GPU with 96 GB memory. The GPU-based Llama-1B mechanism experiments in Appendix J used one NVIDIA RTX PRO 6000 with 96 GB memory.

Software Environment Those inherited benchmark experiments were performed on Ubuntu 22.04.5 LTS with NVIDIA driver version 570.195.03 and CUDA 12.8. DARC framework is implemented in Python 3.12.12 using PyTorch 2.9.1, and is developed on top of **OpenUnlearning** framework. All unlearning algorithms are instantiated using the implementations integrated within OpenUnlearning. The baseline unlearning models are initialized from the publicly released checkpoints provided by OpenUnlearning on Hugging Face: `https://huggingface.co/open-unlearning`.

Table 5: Default configurations and hyperparameters used in our experiments.

Notation	Hyperparameter	Value	Explanation
		General Training Configuration	
–	Random Seed	0	Seed for reproducibility
–	Precision	`bfloat16`	Numerical precision used for training
–	Attention Backend	FlashAttention-2	Optimized attention implementation
–	Optimizer	`paged_adamw_32bit`	Memory-efficient AdamW variant
λ_wd	Weight Decay	0.01	ℓ_2 regularization coefficient
B	Batch Size	32	Number of samples per batch
		Server-Side Configuration	
R	Global Communication Rounds	See Table 10	Total number of server–client rounds
m	Copy Number	2	Number of perturbed model copies per round
κ	Noise Level	0.01	Global multiplier for block-wise noise scales $\{\sigma_\ell\}$
α_k	Noise Scaling	$1+\frac{k-1}{m-1}$	Used to match noise magnitude in the single-copy *Noised* baseline and DARC
η_srv	Server Step Size	1.0	Step size for applying the aggregated update $\bar{\Delta}^{(r)}$
		Executor Unlearning Configuration	
–	Unlearning Algorithms	See Appendix H.3	Unlearning trainer executed on each published copy
E_cli	Local Epoch	See Table 10	Local unlearning epochs per copy per round
η_cli	Local Learning Rate	1×10^{-5}	Learning rate used by the chosen unlearning trainer

Optimizer and schedule (Table 5). The local optimizer in the supplied runs is paged 32-bit AdamW, exactly the setting for which dense rotation keys require an equivariance audit. The schedules also differ substantially across objectives: some use one long local round, while DPO/SimNPO/SatImp use ten short rounds. Cross-algorithm comparisons therefore conflate objective and communication schedule; the optimizer analysis treats the complete scheduled trainer as the local map.

H.4.2 HYPERPARAMETERS

Table 5 lists the experimental hyperparameters.

We configure the training schedule by selecting the number of global communication rounds (**R**) and local client epochs per round (**E**) based on the client-side unlearning method. Specifically, **GradAscent**, **GradDiff**, and **UnDIAL** use a single-round schedule with $(R, E) = (1, 10)$; **SimNPO**, **DPO**, and **SatImp** use a more communication-intensive schedule with $(R, E) = (10, 1)$; and **NPO** adopts an intermediate configuration with $(R, E) = (2, 5)$. For any unspecified method, we default to $(R, E) = (1, 10)$.

H.4.3 PROMPT TEMPLATE

The chat template is enabled during training. Figure 4 and Figure 5 show the formats for user queries and assistant responses.

Prompt: Llama-3.2 Series Template

System Prompt: You are a helpful assistant.
System Prompt with Special Tokens:
`<|begin_of_text|><|start_header_id|>system`
`<|end_header_id|>\n\n You are a helpful assistant.<|eot_id|>`
User Start Tag: `<|start_header_id|>user<|end_header_id|>\n\n`
User End Tag: `<|eot_id|>`
Assistant Start Tag: `<|start_header_id|>assistant<|end_header_id|>\n\n`
Assistant End Tag: `<|eot_id|>`
Data String: `10 Apr 2025`

Figure 4: Prompt template for Llama-3.2 series.

Prompt: Qwen Series Template

System Prompt: You are a helpful assistant.
System Prompt with Special Tokens:
`<|im_start|>system\n You are a helpful assistant.`
`<|im_end|>\n`
User Start Tag: `<|im_start|>user\n`
User End Tag: `<|im_end|>\n`
Assistant Start Tag: `<|im_start|>assistant\n`
Assistant End Tag: `<|im_end|>\n`

Figure 5: Prompt template for Qwen2.5 series.

H.5 SUPPLEMENTARY EXPERIMENTAL RESULTS

Highlighting convention. In the three-setting TOFU table, bold compares only Noised and DARC within each algorithm, model and metric; pale-blue Clean cells are unranked references. Other paired tables retain their stated comparison sets. In within-benchmark noise/copy/schedule sweeps, bold marks the largest FQ, FTR and MU or the smallest absolute PrivLeak; all ties are included. Cross-forget-split tables and attack statistics with unspecified direction are not ranked. Highlighting is descriptive and does not establish statistical significance. Numerical values are unchanged.

Table 6: Performance comparison of different unlearning algorithms using the **Qwen2.5-1.5B**, **Llama-3.2-3B**, and **Qwen2.5-3B** models on the TOFU benchmark (Split99). Results are reported under three settings: **Clean**, a direct noise-free reference; **Noised**, a single-copy noise baseline; and DARC, using m=2 copies with noise level κ=0.01. Higher values indicate better performance for Forget Quality, Forget Truth Ratio, and Model Utility, while values of PrivLeak closer to 0 are preferred. Bold compares Noised and DARC only, with all ties; pale-blue Clean cells are unranked noise-free references.

Unlearning Algorithms	Forget Quality ↑			Forget Truth Ratio ↑			Model Utility ↑			PrivLeak		
	CLEAN (REF.)	NOISED	DARC	CLEAN (REF.)	NOISED	DARC	CLEAN (REF.)	NOISED	DARC	CLEAN (REF.)	NOISED	DARC
Qwen2.5-1.5B-Instruct												
GRADASCENT [ACL 2023]	1.43e-2	5.41e-2	**0.990**	0.516	0.536	**0.740**	9.75e-4	6.76e-4	**0.346**	13.7	62.1	**-21.7**
GRADDIFF [PMLR 2022]	6.58e-5	6.61e-6	**0.990**	0.622	0.586	**0.740**	0.297	0.322	**0.346**	80.9	76.7	**-21.6**
DPO [NEURIPS 2023]	0.266	0.405	**0.990**	0.757	**0.750**	0.738	0.305	0.335	**0.347**	86.0	87.5	**-21.7**
NPO [COLM 2024]	1.43e-2	2.86e-2	**0.919**	0.708	0.709	**0.739**	0.305	0.318	**0.346**	89.4	90.2	**-21.9**
SIMNPO [NEURIPS 2025]	0.990	**0.990**	**0.990**	0.738	**0.742**	0.738	0.349	**0.350**	0.347	-18.7	**-17.4**	-22.0
UNDIAL [NAACL 2025]	1.000	**1.000**	0.990	0.769	**0.767**	0.740	0.338	0.333	**0.346**	12.3	**14.1**	-21.8
SATIMP [ICML2025]	0.919	0.919	**0.990**	0.738	**0.740**	**0.740**	0.349	**0.348**	0.346	-21.8	**-20.5**	-21.5
Llama-3.2-3B-Instruct												
GRADASCENT [ACL 2023]	3.06e-11	**5.91e-14**	4.47e-16	0.187	**0.157**	0.142	8.23e-5	**7.71e-5**	2.28e-5	50.6	53.0	**45.9**
GRADDIFF [PMLR 2022]	5.41e-2	2.86e-2	**5.41e-2**	0.504	0.476	**0.485**	0.419	0.384	**0.409**	108.0	**106.0**	107.0
DPO [NEURIPS 2023]	0.579	**0.766**	0.579	0.685	**0.691**	0.683	0.634	**0.634**	0.633	-10.2	-15.3	**-3.95**
NPO [COLM 2024]	0.990	**0.990**	**0.990**	0.635	**0.636**	0.634	0.650	**0.653**	0.651	61.2	62.1	**57.5**
SIMNPO [NEURIPS 2025]	9.71e-2	**9.71e-2**	**9.71e-2**	0.550	0.550	**0.554**	0.658	**0.662**	0.656	-62.4	-72.3	**-64.3**
UNDIAL [NAACL 2025]	2.86e-2	**5.41e-2**	2.86e-2	0.567	**0.566**	0.563	0.693	0.692	**0.693**	-72.0	**-71.8**	-73.3
SATIMP [ICML2025]	1.43e-2	1.43e-2	**2.86e-2**	0.510	0.507	**0.512**	0.659	**0.661**	0.659	-99.7	-100.0	**-99.9**
Qwen2.5-3B-Instruct												
GRADASCENT [ACL 2023]	0.579	1.86e-23	**0.579**	0.709	6.4e-20	**0.709**	0.392	0.000	**0.391**	-28.7	**8.57**	-28.1
GRADDIFF [PMLR 2022]	0.766	3.02e-3	**0.766**	0.712	0.463	**0.710**	0.394	0.237	**0.395**	-28.0	67.7	**-27.1**
DPO [NEURIPS 2023]	0.766	**0.919**	0.766	0.711	**0.733**	0.709	0.397	**0.412**	0.397	-28.9	84.4	**-28.8**
NPO [COLM 2024]	0.766	0.266	**0.766**	0.712	**0.724**	0.711	0.398	**0.420**	0.397	-28.5	90.5	**-28.5**
SIMNPO [NEURIPS 2025]	0.766	**0.919**	0.766	0.712	**0.726**	0.710	0.398	0.388	**0.397**	-28.5	**-18.9**	-28.6
UNDIAL [NAACL 2025]	0.766	**0.919**	0.766	0.710	**0.734**	0.711	0.398	**0.405**	0.396	-28.1	**4.17**	-28.8
SATIMP [ICML2025]	0.766	**0.766**	**0.766**	0.711	**0.714**	0.712	0.397	0.387	**0.397**	-28.9	**-24.9**	-28.7

Model-dependent results (Table 6). The historical results differ across models. Qwen2.5-3B has nearly identical DIRECT/DARC values across the seven objectives. For Qwen2.5-1.5B, DARC achieves FQ at least 0.919 for every objective and MU between 0.346 and 0.347; MU exceeds DIRECT for five of seven objectives. The comparison describes the reported single seed and metrics; repeatability and the mechanisms behind the changes require paired controls.

Table 7: Performance comparison of different unlearning algorithms using the **Llama-3.2-1B** model under DARC on the TOFU benchmark (Split99), with varying noise levels $\kappa \in \{0, 0.01, 0.05, 0.1\}$ and fixed perturbed copies m=2. Higher values indicate better performance for Forget Quality, Forget Truth Ratio, and Model Utility, while values of PrivLeak closer to 0 are preferred.

Unlearning Algorithms	Forget Quality ↑				Forget Truth Ratio ↑				Model Utility ↑				PrivLeak			
	κ=0	κ=0.01	κ=0.05	κ=0.1	κ=0	κ=0.01	κ=0.05	κ=0.1	κ=0	κ=0.01	κ=0.05	κ=0.1	κ=0	κ=0.01	κ=0.05	κ=0.1
GRADASCENT	6.76e-3	5.41e-2	**0.579**	1.43e-2	0.440	0.468	**0.534**	0.430	1.39e-4	2.31e-4	**3.06e-4**	1.49e-4	**67.9**	69.6	68.0	69.1
GRADDIFF	**0.405**	**0.405**	0.266	0.400	0.546	0.547	0.544	**0.553**	0.468	0.464	**0.474**	0.472	**75.3**	77.2	76.5	76.7
DPO	**0.266**	**0.266**	**0.266**	0.165	0.640	**0.641**	**0.641**	0.637	0.592	0.591	**0.594**	**0.594**	-30.1	-28.9	-29.2	**-28.7**
NPO	**0.919**	**0.919**	**0.919**	**0.919**	0.621	**0.628**	0.624	0.618	0.595	0.597	**0.599**	0.596	27.7	28.2	**26.8**	28.1
SIMNPO	5.41e-2	**9.71e-2**	**9.71e-2**	5.41e-2	0.520	0.525	**0.526**	0.525	0.596	**0.598**	**0.598**	0.597	-72.4	-71.8	-72.1	**-71.7**
UNDIAL	**1.43e-2**	**1.43e-2**	**1.43e-2**	**1.43e-2**	0.528	**0.529**	0.527	0.526	0.614	**0.615**	**0.615**	0.613	-78.3	**-78.0**	-78.2	**-78.0**
SATIMP	**6.76e-3**	**6.76e-3**	3.02e-3	**6.76e-3**	0.474	**0.476**	0.475	0.475	**0.601**	**0.601**	0.600	0.600	-99.3	**-98.9**	-99.2	-99.2

Response to noise (Table 7). NPO keeps FQ 0.919 at every tested κ and MU remains in $[0.595, 0.599]$; UnDIAL and SatImp are similarly stable in MU. GradDiff utility is also stable, although its FQ fluctuates. DPO falls from FQ 0.266 to 0.165 at $\kappa = 0.1$, and GradAscent is strongly

non-monotonic with essentially zero utility. The stable objectives preserve utility across the tested noise levels; this observation is specific to the reported sweep.

Table 8: Single-copy no-denoise results for **Llama-3.2-1B** on TOFU Forget01. The displayed effective noise multipliers are $\{0.015, 0.075, 0.15\}$, obtained by multiplying nominal $\kappa \in \{0.01, 0.05, 0.1\}$ by the mean scale 1.5 of the two-copy DARC reference. This is a single-copy control; the table contains no zero-noise column. Higher values indicate better performance for Forget Quality, Forget Truth Ratio, and Model Utility, while values of PrivLeak closer to 0 are preferred.

Unlearning Algorithms	Forget Quality ↑			Forget Truth Ratio ↑			Model Utility ↑			PrivLeak		
	κ=0.015	κ=0.075	κ=0.15	κ=0.015	κ=0.075	κ=0.15	κ=0.015	κ=0.075	κ=0.15	κ=0.015	κ=0.075	κ=0.15
GradAscent	2.81e-8	1.12e-9	**1.23e-7**	0.246	0.190	**0.252**	0.000	0.000	**4.11e-5**	58.9	**57.4**	66.4
GradDiff	0.266	0.266	**0.405**	0.533	0.534	**0.546**	0.461	**0.466**	0.463	73.3	75.4	**72.3**
DPO	**0.165**	0.097	**0.165**	**0.620**	0.618	**0.620**	**0.595**	**0.595**	**0.595**	**-19.8**	-20.0	-20.0
NPO	**0.766**	**0.766**	**0.766**	**0.640**	0.623	0.622	**0.600**	0.599	0.597	**32.9**	35.4	33.3
SimNPO	**5.41e-2**	**5.41e-2**	**5.41e-2**	0.522	0.522	**0.523**	0.592	0.593	**0.597**	**-70.2**	-71.8	-72.1
UnDIAL	**1.43e-2**	**1.43e-2**	**1.43e-2**	**0.527**	0.526	**0.527**	0.614	0.614	**0.616**	-77.4	-77.3	**-77.2**
SatImp	**6.76e-3**	**6.76e-3**	**6.76e-3**	0.470	**0.471**	**0.471**	0.597	0.597	**0.598**	**-99.1**	-99.2	-99.2

Single-copy control (Table 8). The single-copy no-denoise control collapses GradAscent at all three matched scales, whereas GradDiff, DPO, NPO, SimNPO, UnDIAL, and SatImp remain broadly usable. The response is trainer-dependent. The control matches expected copy scale and follows a different stochastic trajectory, so both perturbation and trajectory effects enter this comparison.

Table 9: Performance comparison of different unlearning algorithms using the **Llama-3.2-1B** model under DARC on the TOFU benchmark (Split99), with varying numbers of perturbed copies $m \in \{2, 3, 4\}$ and noise level κ=0.01. Higher values indicate better performance for Forget Quality, Forget Truth Ratio, and Model Utility, while values of PrivLeak closer to 0 are preferred.

Unlearning Algorithms	Forget Quality ↑			Forget Truth Ratio ↑			Model Utility ↑			PrivLeak		
	m=2	m=3	m=4	m=2	m=3	m=4	m=2	m=3	m=4	m=2	m=3	m=4
GradAscent	5.41e-2	**0.165**	2.16e-5	0.468	**0.500**	0.334	**2.31e-4**	1.68e-4	0.000	69.6	64.0	**63.8**
GradDiff	0.405	0.405	**0.579**	0.547	0.555	**0.570**	0.464	**0.467**	**0.467**	**77.2**	77.7	77.4
DPO	0.266	**0.579**	0.165	**0.641**	0.631	0.615	0.591	0.592	**0.594**	-28.9	**-22.0**	-22.2
NPO	**0.919**	0.766	0.766	**0.628**	0.623	0.620	**0.597**	0.590	0.594	28.2	25.3	**25.1**
SimNPO	**9.71e-2**	5.41e-2	5.41e-2	**0.525**	0.518	0.519	0.598	0.597	**0.599**	**-71.8**	-74.6	-74.1
UnDIAL	**1.43e-2**	**1.43e-2**	6.76e-3	**0.529**	0.528	0.528	**0.615**	0.613	**0.615**	**-78.0**	**-78.0**	-78.4
SatImp	6.76e-3	**1.43e-2**	3.02e-3	**0.476**	0.473	0.471	**0.601**	0.598	0.599	**-98.9**	-99.2	-99.8

Copy-count changes (Table 9). Copy count is not monotonically beneficial. GradDiff FQ changes from 0.405 to 0.579 at $m = 4$, DPO peaks at $m = 3$ and then falls, and NPO decreases from 0.919 at $m = 2$ to 0.766 at $m = 3, 4$; GradAscent collapses at $m = 4$. In contrast, MU is nearly constant for every non-collapsed objective. These are behavioral comparisons across copy counts; the corresponding isolation thresholds are analytical.

Table 10: Performance comparison of different unlearning algorithms using the **Llama-3.2-1B** model under DARC on the TOFU benchmark, across varying global communication rounds (**R**) and local client epochs (**E**), with fixed perturbed copies $m{=}2$ and noise level $\kappa{=}0.01$. Higher values indicate better performance for Forget Quality and Model Utility, while values of PrivLeak closer to 0 are preferred.

Unlearning Algorithms	Forget Quality ↑				Model Utility ↑				PrivLeak			
	R1E10	R2E5	R5E2	R10E1	R1E10	R2E5	R5E2	R10E1	R1E10	R2E5	R5E2	R10E1
GradAscent	**5.41e-2**	1.86e-23	1.86e-23	1.86e-23	**2.31e-4**	0.00	0.000	0.000	69.6	35.2	**-2.24**	25.6
GradDiff	**0.405**	1.43e-2	5.04e-4	1.95e-10	0.464	0.484	0.494	**0.578**	**77.2**	86.8	88.9	88.9
DPO	5.41e-2	0.165	0.165	**0.266**	0.584	0.591	**0.592**	0.591	-49.5	-42.1	**-17.9**	-28.9
NPO	0.766	**0.919**	**0.919**	**0.919**	0.584	**0.597**	0.592	0.594	51.8	**28.2**	34.7	40.7
SimNPO	2.86e-2	2.86e-2	**9.71e-2**	**9.71e-2**	0.599	**0.601**	0.598	0.598	-78.0	-80.5	-75.6	**-71.8**
UnDial	**1.43e-2**	6.76e-3	**1.43e-2**	6.76e-3	0.615	**0.617**	0.612	0.615	-78.0	-82.8	**-77.6**	-84.2
SatImp	**6.76e-3**	3.02e-3	**6.76e-3**	**6.76e-3**	0.600	0.599	0.598	**0.601**	-99.9	-99.8	-99.3	**-98.9**

Allocation across rounds (Table 10). Holding ten local epochs fixed while changing round allocation changes the complete optimization map. GradAscent and GradDiff FQ decreases sharply with more rounds, while DPO improves toward its original ten-round schedule and NPO is strongest near $R2E5$. The scheduled trainer includes state resets and server commits; it also matters for privacy because every additional publication round requires conditional RDP accounting under Proposition 2.

Table 11: Performance comparison of different unlearning algorithms using the **Llama-3.2-1B** model under DARC on the TOFU benchmark, across varying split strategies (Forget01, Forget05, and Forget10), with fixed perturbed copies $m{=}2$ and noise level $\kappa{=}0.01$. Higher values indicate better performance for Forget Quality, Forget Truth Ratio, and Model Utility, while values of PrivLeak closer to 0 are preferred.

Unlearning Algorithms	Forget Quality ↑			Forget Truth Ratio ↑			Model Utility ↑			PrivLeak		
	Forget01	Forget05	Forget10	Forget01	Forget05	Forget10	Forget01	Forget05	Forget10	Forget01	Forget05	Forget10
GradAscent	5.41e-2	1.94e-119	1.06e-239	0.468	1.71e-23	1.74e-22	2.31e-4	0.000	0.000	69.6	-28.0	-21.6
GradDiff	0.405	5.99e-105	8.51e-237	0.547	0.002	3.09e-10	0.464	0.421	0.264	77.2	56.8	61.8
DPO	0.266	4.30e-3	9.07e-8	0.641	0.574	0.545	0.591	0.595	0.598	-28.9	-43.7	-59.0
NPO	0.919	0.112	4.46e-6	0.628	0.603	0.555	0.597	0.614	0.623	28.2	-4.00	-0.946
SimNPO	9.71e-2	4.61e-7	5.42e-13	0.525	0.514	0.498	0.598	0.595	0.598	-71.8	-77.7	-82.0
UnDial	1.43e-2	2.44e-10	7.98e-17	0.529	0.521	0.531	0.615	0.615	0.616	-78.0	-91.7	-94.9
SatImp	6.76e-3	1.33e-13	2.81e-20	0.476	0.471	0.469	0.601	0.593	0.597	-98.9	-99.9	-99.2

Larger forget sets (Table 11). Under unchanged hyperparameters, FQ deteriorates by many orders of magnitude from Forget01 to Forget05/10 for nearly every objective, even when MU remains around 0.59–0.62 for bounded methods. Thus retained utility can mask failure to match the retain-only forget distribution. Robustness claims should remain restricted to the evaluated deletion size, and larger requests require retuning rather than merely increasing the number of coded views.

Table 12: Baseline performance of the OpenUnlearning Framework without DARC, using the **Llama-3.2-1B** and **Llama-3.2-3B** models on TOFU Forget01, Forget05, and Forget10. Higher values indicate better performance for Forget Quality, Forget Truth Ratio, and Model Utility, while values of PrivLeak closer to 0 are preferred.

Unlearning Algorithms	Forget Quality ↑			Forget Truth Ratio ↑			Model Utility ↑			PrivLeak		
	FORGET01	FORGET05	FORGET10	FORGET01	FORGET05	FORGET10	FORGET01	FORGET05	FORGET10	FORGET01	FORGET05	FORGET10
Llama-3.2-1B-Instruct												
GRADASCENT	6.58e-5	1.94e-119	1.06e-239	0.355	2.12e-23	1.27e-22	0.000	0.000	0.000	65.8	-28.4	-21.1
GRADDIFF	0.405	5.99e-105	8.51e-237	0.535	0.004	7.22e-10	0.461	0.393	0.247	77.1	56.8	61.3
DPO	0.165	2.08e-3	3.08e-7	0.637	0.573	0.545	0.591	0.594	0.598	-25.5	-44.5	-57.8
NPO	0.919	0.112	4.58e-7	0.624	0.595	0.551	0.599	0.614	0.619	30.6	-2.11	3.78
SIMNPO	5.41e-2	1.46e-7	2.73e-12	0.526	0.510	0.498	0.598	0.593	0.597	-68.4	-76.6	-82.4
UNDIAL	1.43e-2	4.87e-10	4.24e-17	0.530	0.524	0.531	0.613	0.615	0.618	-76.4	-91.6	-94.8
SATIMP	3.02e-3	2.96e-13	1.12e-19	0.474	0.471	0.469	0.600	0.594	0.600	-98.9	-99.9	-99.2
Llama-3.2-3B-Instruct												
GRADASCENT	3.06e-11	1.94e-119	1.06e-239	0.187	6.63e-8	4.79e-20	8.23e-5	0.000	0.000	50.6	-22.5	-19.2
GRADDIFF	5.41e-2	1.94e-119	8.51e-237	0.504	2.01e-9	1.20e-10	0.419	0.593	0.345	108	56.3	64.6
DPO	0.579	1.18e-2	2.13e-6	0.685	0.588	0.594	0.634	0.635	0.634	-10.2	-62.0	-69.1
NPO	0.990	0.793	2.99e-2	0.635	0.617	0.581	0.650	0.675	0.673	61.2	9.83	-24.0
SIMNPO	9.71e-2	3.60e-9	1.03e-14	0.550	0.499	0.501	0.658	0.652	0.659	-62.4	-81.0	-86.3
UNDIAL	2.86e-2	3.08e-12	8.08e-22	0.567	0.507	0.515	0.693	0.701	0.696	-72.0	-92.1	-94.3
SATIMP	1.43e-2	2.96e-13	2.05e-24	0.510	0.463	0.469	0.659	0.649	0.659	-99.7	-100	-99.5

Direct baselines (Table 12). The direct baselines exhibit the same severe Forget05/10 degradation seen with coded execution, especially for GradAscent and GradDiff. This comparison prevents attributing every large-split failure to DARC. The scientifically relevant quantity is the paired change from the corresponding direct trainer under the same model and schedule, not the absolute rank of unrelated objectives.

Table 13: Results of Attack Tests among Six Methods with Qwen2.5-1.5B-Instruct on the TOFU Benchmark(Split99). The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	LOSS	ZLib	GradNorm	MinK	MinK++	Reference
GRADASCENT	0.39/0.587	0.388/0.608	0.565/0.727	0.385/0.59	0.431/0.632	0.28/0.993
GRADDIFF	0.383/0.589	0.416/0.609	0.479/0.734	0.378/0.589	0.388/0.636	0.128/0.992
NPO	0.363/0.589	0.359/0.611	0.546/0.731	0.354/0.591	0.39/0.64	0.246/0.999
SIMNPO	0.584/0.589	0.604/0.608	0.727/0.734	0.586/0.591	0.634/0.636	0.961/0.998
UNDIAL	0.46/0.589	0.47/0.608	0.379/0.733	0.512/0.591	0.609/0.639	0.463/0.999
SATIMP	0.589/0.588	0.606/0.611	0.732/0.734	0.585/0.589	0.641/0.642	0.977/0.995

Final-model attack statistics (Table 13). The six legacy attack-statistic pairs are reliable reported measurements on the final unlearned model, which is studied separately in Appendix J.9.

Table 14: Results of LoRA compatibility with Llama-3.1-8B-Instruct on the TOFU Benchmark(Split99). The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Forget Quality	Forget Truth Ratio	Model Utility	PrivLeak
GRADASCENT	0.919/0.919	0.721/0.721	0.427/0.426	-11.3/-10.9
GRADDIFF	0.919/0.919	0.721/0.724	0.427/0.426	-9.78/-8.85
NPO	0.919/0.99	0.723/0.724	0.429/0.429	-9.95/-9.55
SIMNPO	1.0/1.0	0.735/0.733	0.396/0.411	-12.0/-11.6
DPO	1.0/0.99	0.731/0.732	0.426/0.426	-8.03/-6.98
UNDIAL	0.99/0.99	0.733/0.731	0.376/0.381	-12.8/-12.8
SATIMP	0.99/1.0	0.733/0.734	0.394/0.411	-11.9/-12.0

LoRA results (Table 14). Both settings have FQ between 0.919 and 1.0; the largest displayed FQ difference is 0.071 (NPO). The maximum absolute MU difference is 0.017 (SatImp: 0.394 versus

0.411). These historical values support compatibility at the 8B LoRA scale, not a formal optimizer-equivariance result: rank-gauge keys, base/adaptor tying, reference construction, and optimizer-state mapping require the conditions in Appendices B.1 and E.

Table 15: Results of Qwen2.5-1.5B-Instruct on the TOFU(Split99) Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Forget Quality	Forget Truth Ratio	Model Utility	PrivLeak
GRADASCENT	1.43e-2/0.990	0.516/0.740	9.75e-4/0.346	13.7/-21.7
GRADDIFF	6.58e-5/0.990	0.622/0.740	0.297/0.346	80.9/-21.6
DPO	0.266/0.990	0.757/0.738	0.305/0.347	86.0/-21.7
NPO	1.43e-2/0.919	0.708/0.739	0.305/0.346	89.4/-21.9
SIMNPO	0.990/0.990	0.738/0.738	0.349/0.347	-18.7/-22.0
UNDIAL	1.000/0.990	0.769/0.740	0.338/0.346	12.3/-21.8
SATIMP	0.919/0.990	0.738/0.740	0.349/0.346	-21.8/-21.5

Qwen2.5-1.5B on TOFU (Table 15). This table repeats the DIRECT/DARC pairs from the Qwen2.5-1.5B block of the historical comparison. DARC gives FQ ≥ 0.919 for all seven objectives and higher MU than DIRECT for GradAscent, GradDiff, DPO, NPO, and UnDIAL. SimNPO and SatImp instead have MU decreases of 0.002 and 0.003. These are single-seed comparisons; paired repetitions would assess their repeatability and mechanism.

Table 16: Results of Qwen2.5-3B-Instruct on the TOFU(Split99) Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Forget Quality	Forget Truth Ratio	Model Utility	PrivLeak
GRADASCENT	0.579/0.579	0.709/0.709	0.392/0.391	-28.7/-28.1
GRADDIFF	0.766/0.766	0.712/0.710	0.394/0.395	-28.0/-27.1
DPO	0.766/0.766	0.711/0.709	0.397/0.397	-28.9/-28.8
NPO	0.766/0.766	0.712/0.711	0.398/0.397	-28.5/-28.5
SIMNPO	0.766/0.766	0.712/0.710	0.398/0.397	-28.5/-28.6
UNDIAL	0.766/0.766	0.710/0.711	0.398/0.396	-28.1/-28.8
SATIMP	0.766/0.766	0.711/0.712	0.397/0.397	-28.9/-28.7

Qwen2.5-3B on TOFU (Table 16). The Qwen2.5-3B results are closely matched: every FQ pair is identical at the displayed precision, FTR changes by at most 0.002, MU by at most 0.002, and PrivLeak remains close. The single-seed agreement describes this model and protocol; the cross-model differences motivate setting-specific assessment.

Table 17: Results of Llama-3.2-1B-Instruct on the MUSE-Books Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Extraction Strength	Forget KnowMem ROUGE	Retain KnowMem ROUGE	PrivLeak
GRADASCENT	0.00915/0.00915	0/0	0/0	-32.0/-34.5
GRADDIFF	0.00915/0.00915	0.0455/0.0331	0.146/0.141	28.9/31.7
NPO	0.0125/0.0127	0.172/0.181	0.188/0.191	-23.7/-24.0
SIMNPO	0.0122/0.0125	0.161/0.152	0.206/0.218	-30.1/-28.5
UNDIAL	0.0112/0.0112	0.167/0.183	0.204/0.195	-56.9/-57.2
SATIMP	0.00924/0.00924	0.152/0.147	0.202/0.198	-2.22/-1.87

Llama-1B on Books (Table 17). Most direct/DARC pairs are close, especially extraction strength. The largest visible differences occur for GradDiff and UnDIAL in the two ROUGE columns, but their direction is mixed. This supports broad MUSE compatibility rather than a consistent denoising gain, and the single-seed differences should not be ranked.

Table 18: Results of Llama-3.2-3B-Instruct on the MUSE-Books Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Extraction Strength	Forget KnowMem ROUGE	Retain KnowMem ROUGE	PrivLeak
GradAscent	0.00915/0.00915	0/0	0/0	52.3/56.6
GradDiff	0.00915/0.00915	0/0	0.115/0.146	7.42/35.4
NPO	0.0155/0.015	0.234/0.226	0.313/0.333	-71.9/-73.2
SimNPO	0.0143/0.0141	0.237/0.227	0.32/0.316	-84.2/-83.6
UnDIAL	0.0109/0.0106	0.243/0.229	0.359/0.316	-89.6/-89.6
SatImp	0.00915/0.00915	0.221/0.239	0.315/0.3	-28.8/-29.8

Llama-3B on Books (Table 18). NPO, SimNPO, UnDIAL, and SatImp remain close under coded execution, whereas GradDiff changes Retain KnowMem ROUGE from 0.115 to 0.146 and PrivLeak from 7.42 to 35.4.

Table 19: Results of Qwen2.5-1.5B-Instruct on the MUSE-Books Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Extraction Strength	Forget KnowMem ROUGE	Retain KnowMem ROUGE	PrivLeak
GradAscent	0.0148/0.0149	0.22/0.22	0.24/0.242	-29.7/-29.7
GradDiff	0.0143/0.014	0.218/0.214	0.281/0.277	-29.3/-28.9
NPO	0.0147/0.0144	0.221/0.223	0.306/0.304	-26.0/-25.8
SimNPO	0.0144/0.0147	0.221/0.221	0.294/0.297	-25.9/-26.0
UnDIAL	0.0146/0.0142	0.226/0.215	0.273/0.282	-27.8/-27.8
SatImp	0.0147/0.0146	0.227/0.218	0.297/0.303	-25.9/-25.9

Qwen2.5-1.5B on Books (Table 19). All six methods produce close direct/DARC pairs across extraction, forget ROUGE, retain ROUGE, and PrivLeak. These observations provide no indication of a large shift caused by the wrapper on this model. Repeated seeds would still be needed to test equivalence.

Table 20: Results of Qwen2.5-3B-Instruct on the MUSE-Books Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Extraction Strength	Forget KnowMem ROUGE	Retain KnowMem ROUGE	PrivLeak
GradAscent	0.0172/0.0175	0.214/0.218	0.238/0.247	-56.1/-55.9
GradDiff	0.0159/0.0155	0.191/0.213	0.229/0.201	-57.9/-57.1
NPO	0.0145/0.0148	0.206/0.186	0.194/0.198	-53.1/-52.9
SimNPO	0.0147/0.0146	0.182/0.186	0.205/0.2	-52.9/-53.0
UnDIAL	0.0136/0.0136	0.151/0.135	0.129/0.13	-48.2/-48.6
SatImp	0.0147/0.0147	0.193/0.208	0.192/0.219	-52.6/-53.1

Qwen2.5-3B on Books (Table 20). Extraction and PrivLeak are nearly unchanged. Some ROUGE pairs move by roughly 0.02–0.03 in opposite directions, for example GradDiff retain behavior and NPO forget behavior, so there is no uniform improvement. The sources of these single-seed differences remain unresolved.

Table 21: Results of Llama-3.2-1B-Instruct on the MUSE-News Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Extraction Strength	Forget KnowMem ROUGE	Retain KnowMem ROUGE	PrivLeak
GradAscent	0.00975/0.00975	0/0	0/0	19.3/19.2
GradDiff	0.00985/0.01	0.0381/0.0346	0.0292/0.0123	79.7/82.9
NPO	0.0279/0.028	0.163/0.167	0.178/0.178	-28.0/-27.5
SimNPO	0.0282/0.0279	0.155/0.169	0.181/0.175	-52.0/-52.0
UnDIAL	0.016/0.016	0.136/0.137	0.21/0.207	-49.0/-49.0
SatImp	0.0281/0.028	0.173/0.17	0.179/0.176	-52.2/-52.1

Llama-1B on News (Table 21). Most rows are close, with GradDiff again showing the clearest retain-score difference (0.0292 versus 0.0123) while both values remain low. Because PrivLeak also moves slightly away from zero, this is not a simple forgetting–utility improvement. This is a method-specific single-seed discrepancy.

Table 22: Results of Llama-3.2-3B-Instruct on the MUSE-News Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Extraction Strength	Forget KnowMem ROUGE	Retain KnowMem ROUGE	PrivLeak
GradAscent	0.00975/0.00975	0/0	0/0	14.2/15.4
GradDiff	0.0142/0.0132	0/0	0/0.026	64.1/68.3
NPO	0.0271/0.0283	0.234/0.213	0.255/0.233	-64.5/-64.9
SimNPO	0.0282/0.0288	0.236/0.231	0.213/0.216	-82.7/-82.6
UnDIAL	0.0136/0.0134	0.236/0.238	0.276/0.256	-81.6/-81.6
SatImp	0.0288/0.0278	0.224/0.232	0.213/0.212	-82.5/-82.8

Llama-3B on News (Table 22). The bounded objectives are close in extraction and privacy leakage. GradDiff changes retain ROUGE from zero to 0.026, but its forget ROUGE remains zero and its leakage remains large. This table again supports compatibility more strongly than quality improvement.

Table 23: Results of Qwen2.5-1.5B-Instruct on the MUSE-News Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Extraction Strength	Forget KnowMem ROUGE	Retain KnowMem ROUGE	PrivLeak
GradAscent	0.0242/0.0246	0.196/0.196	0.168/0.159	-18.7/-18.8
GradDiff	0.0279/0.0279	0.149/0.147	0.2/0.193	-19.0/-18.9
NPO	0.0278/0.0259	0.123/0.125	0.149/0.159	-15.7/-15.8
SimNPO	0.0263/0.028	0.129/0.129	0.159/0.156	-15.7/-15.6
UnDIAL	0.0259/0.0263	0.107/0.103	0.119/0.123	-15.3/-15.3
SatImp	0.0261/0.028	0.125/0.126	0.15/0.157	-15.6/-15.8

Qwen2.5-1.5B on News (Table 23). Every pair is close at the reported precision, including PrivLeak. Together with the corresponding Books table, this is the most consistent cross-domain evidence that coded execution need not materially alter a stable Qwen2.5-1.5B MUSE trainer, despite the unusual TOFU slice.

Table 24: Results of Qwen2.5-3B-Instruct on the MUSE-News Benchmark. The results are shown by comparison of CLEAN/DARC, with hyperparameter $m = 2$ and $\kappa = 0.01$ by default.

Algorithms	Extraction Strength	Forget KnowMem ROUGE	Retain KnowMem ROUGE	PrivLeak
GradAscent	0.0254/0.0254	0.042/0.0413	0.0738/0.067	-39.5/-39.4
GradDiff	0.031/0.0309	0.0658/0.076	0.0636/0.0588	-41.3/-41.0
NPO	0.0244/0.0278	0.063/0.0639	0.0694/0.0727	-28.5/-28.6
SimNPO	0.0244/0.0245	0.0644/0.0602	0.0702/0.0703	-28.6/-28.7
UnDIAL	0.023/0.0234	0.0653/0.0641	0.0713/0.0758	-27.7/-28.1
SatImp	0.0238/0.0279	0.0621/0.0634	0.0689/0.0664	-28.3/-28.5

Qwen2.5-3B on News (Table 24). The direct/DARC gaps are uniformly small across all four metrics and six objectives. This reinforces the Qwen2.5-3B TOFU result: on this model, the available seed-0 evidence primarily indicates preservation rather than systematic improvement or degradation.

I Analysis of the Supplied Results

All analyses in this section describe the historical seed-0 measurements in Appendix H. The auxiliary-state diagnostic has its own protocol (Appendix J.1). Numerical privacy calibration additionally requires a sensitivity estimate.

I.1 Copy count and threshold mode

Table 9 varies $m \in \{2, 3, 4\}$ at $\kappa = 0.01$. The associations $t = 1, 2, 3$ for $m = 2, 3, 4$ denote the theoretical observation limits; these runs measure behavioral outcomes. GradDiff rises from FQ 0.405 to 0.579 at $m = 4$, and DPO reaches its largest reported FQ, 0.579, at $m = 3$. NPO falls from 0.919 at $m = 2$ to 0.766 for $m = 3, 4$, while GradAscent collapses at $m = 4$. MU is stable for non-collapsed methods. The runs demonstrate execution with $m \geq 3$ and trainer-dependent responses to copy count.

I.2 Noise level and no-denoise control

Table 7 varies κ through 0.1. NPO, UnDIAL, and SatImp are nearly invariant in FQ and MU. DPO loses FQ at $\kappa = 0.1$, and GradAscent is strongly non-monotonic. The mean-scale-matched single-copy results in Table 8 show that committing a single noisy update is particularly damaging for GradAscent, while stable algorithms can remain similar under both procedures. The comparison combines perturbation control with differences in compute, variance and the committed-state convention. The sensitivity Δ_r needed for a numerical privacy budget remains unmeasured.

I.3 Round–epoch allocation

Table 10 holds ten local epochs fixed while changing their allocation across communication rounds. GradAscent and GradDiff FQ decreases sharply with more rounds, whereas DPO improves and NPO prefers R2E5. Changing round allocation also changes the complete trainer map, including state resets and server commits. Multi-round privacy composes across these rounds only after the conditional-release criterion in Appendix B.4 is established, making schedule choice relevant to both utility and cumulative privacy cost.

I.4 Forget-set and model scale

Tables 11 and 12 show that FQ generally becomes much worse at Forget05/10 under unchanged hyperparameters. Larger models provide additional headroom for NPO and DPO in some settings, while unstable objectives remain sensitive to deletion size. The observed robustness is specific to the evaluated split and hyperparameters.

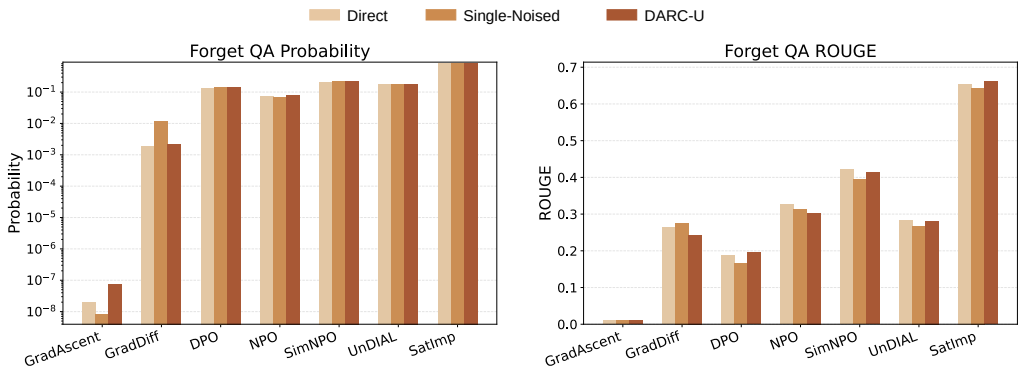


Figure 6: Colored memorization comparison on Llama-3.2-1B/TOFU Forget01 for the supplied seed-0 run. The figure reports Forget-QA probability and ROUGE under Direct, Single-Noised, and DARC. The figure preserves the historical seed-0 measurements and displays their memorization statistics.

I.5 Final-model membership/extraction audit

Table 13 contains LOSS, ZLib, GradNorm, MinK, MinK++, and reference-model statistics for the final unlearned model. These diagnostics measure final-model behavior, with favorable direction depending on the individual statistic. Checkpoint-publication accounting and published-view recovery are assessed separately.

I.6 LoRA and MUSE

Table 14 establishes compatibility with an 8B LoRA setting. The MUSE Books and News tables report all Direct/DARC pairs across Llama and Qwen models. Most pairs are close at the reported precision. These single-seed observations broaden the behavioral evidence across model sizes and domains; repeatability remains to be assessed.

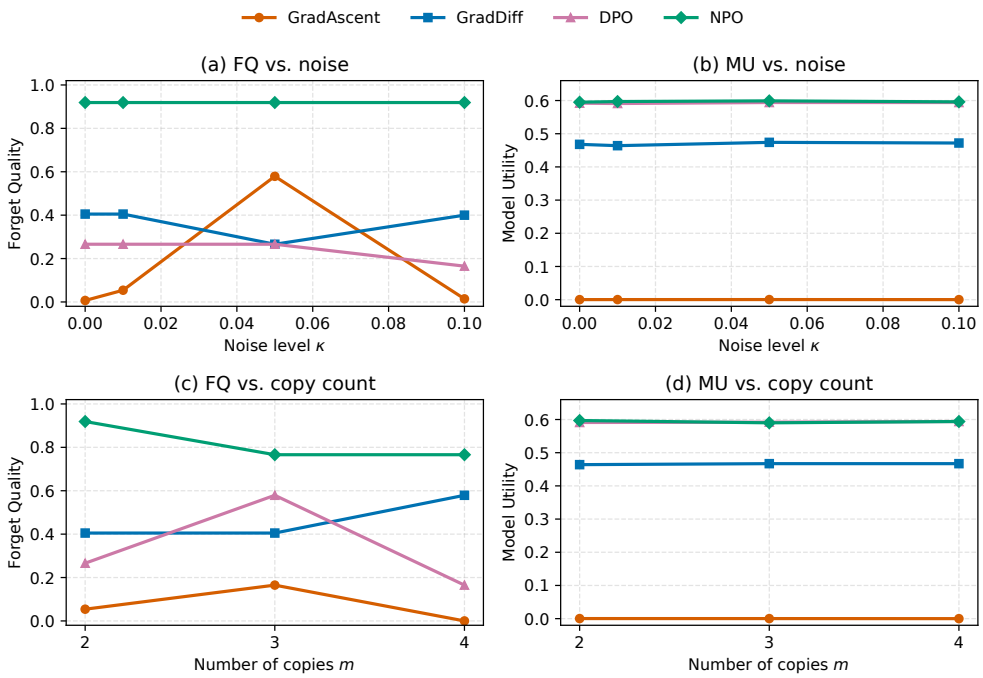


Figure 7: **Sensitivity to publication noise and copy count.** Llama-3.2-1B on TOFU Forget01. Top: $m = 2, \kappa \in \{0, .01, .05, .1\}$. Bottom: $\kappa = .01, m \in \{2, 3, 4\}$. Curves are supplied seed-0 behavioral measurements. FQ varies non-monotonically with m, and GradAscent exhibits utility collapse.

J Mechanism Experiments and Empirical Limits

The experiments below test reference construction, decoding, optimization and recovery. Single-batch checks, actual training, offline tensor arithmetic and endpoint evaluation address distinct questions; each protocol specifies its comparison objects and scope.

Protocol overview. The additional experiments use the OpenUnlearning TOFU-full Llama-3.2-1B-Instruct checkpoint as the canonical model and Meta Llama-3.2-1B-Instruct as the public base. Forget01/Retain99 roles and the retain-only FQ reference are specified below; exact asset revisions are listed in the companion reproducibility configuration. Tied embedding/head storage is counted once. GPU workloads used one NVIDIA RTX PRO 6000 with 96 GB memory; source inspections and offline arithmetic used CPUs. The historical L20/HGX H20 configurations remain as specified in Appendix H.4.1.

Table 25: Comparison objects and evidence types. Each comparison specifies its reference, initialization and evaluation protocol.

Object / experiment	Definition and permitted interpretation
Clean DIRECT	The same local trainer starts at the clean canonical checkpoint. It provides a noise-free reference for the same trainer.
Single-Noised	Historical one-copy noisy initialization with no authorized decoding. Its historical noisy-state commitment is distinguished from the new update-only controls.
Canonical-noisy aggregate	Combine updates from multiple noisy canonical-coordinate initializations.
Aligned aggregate	Invert the actual copy keys, combine returned displacements, and add the result to the clean server checkpoint.
Individual aligned copy	Clean checkpoint plus one aligned displacement, with the clean state as the commitment origin.
Single-batch / fixed-history	Local gradients or conditional optimizer arithmetic, without an independent full training trajectory.
Recovery / telemetry	Bounded public-prompt functional tests or component resource observations.

J.1 Auxiliary-state closure and reference-source diagnostics

These single-batch experiments separate accessible checkpoint-dependent objects, the effect of changing a frozen reference, and numerical coordinate-transform controls.

Inputs and computation. The full checkpoint and TOFU assets are identified in the protocol overview and companion configuration. Forget01 and Retain99 contribute original row0, fixed once with no shuffle, replacement or truncation. The cap is 512, right pair-max padding is used and ignore index is -100. Forget/retain have 37/30 shifted answer targets. DPO uses the same question with a newly fixed chosen refusal, "I don't know.", containing six targets; the original answer is rejected. Unpadded lengths are 101, 90 and 70, not the padded tensor widths. This does not reconstruct a historical DPO pair.

One saved noisy view and FFN key are reused from a three-copy initialization with scales $(1, 1.5, 2)$, $\kappa = .01$, and noise/key seeds 101001/202001. This diagnostic does not run the other views or aggregate their updates. It uses checkpoint-dependent task-vector RMS, seed 303001, FP32 forward/backward, eager attention, train-mode gradients, zero native dropout and no AMP, TF32, checkpointing or compilation. There are zero optimizer updates. References are frozen without gradients in separate caches. Only summary reductions use FP64. The recorded runtime is PyTorch 2.12.1+cu132 and Transformers 5.14.1.

Table 26: Explicit single-batch diagnostic coefficients, not observed outcomes. γ_{f} and α_{r} multiply forget and retain components.

Trainer	γ_{f}	α_{r}	β	Retain / additional setting
GradDiff	1	1	–	NLL; separate fixed KL
DPO	1	1	0.1	NLL
NPO	1	1	0.1	NLL
SimNPO	0.125	1	4.5	NLL; separate fixed KL; $\delta = 0$
UnDIAL	1	0	10	weight zero
NEW 1B LoRA	1	1	0.1	NLL; NPO objective

Objectives and source-level limits. The loss is $\gamma_{\mathrm{f}} L_{\mathrm{f}} + \alpha_{\mathrm{r}} L_{\mathrm{r}}$, with coefficients in Table 26. NLL averages shifted answer targets. DPO/NPO reference ratios use summed shifted answer log probabilities; DPO does not length-normalize the pair. SimNPO uses its length-normalized forget component with $\delta = 0$. UnDIAL uses answer-masked soft cross-entropy from adjusted teacher logits; its retain forward may be evaluated even when its weight is zero. The explicit retain KL helper flattens all positions for batchmean $\mathrm{KL}(p_{\mathrm{ref}} \| p_{\mathrm{student}})$, including prompt and right padding.

Conditional closure and accounting. L denotes a conditional failure of the legacy path: access to a canonical model, round snapshot or private teacher means the executor sees more than the coded view, even if the object makes no gradient contribution. H denotes conditional same-view helper support: the cache is derived from the supplied view and allowed fixed inputs, with source/input binding and detached storage. The surrounding same-process audit runner is trusted; actual isolation is not verified. Frozen, no-gradient or detached storage does not make a private object public.

U means complete-transcript additional RDP is unknown or unaccounted. Z is only an object-specific conditional zero increment: public checkpoint-independent state or permitted same-view post-processing adds no release cost beyond its input. It does not erase the view's cost, pay for private canonical access or profile disclosure, or certify client-data privacy. Forget/retain inputs and tokenizer/templates require their declared conditioning; the checkpoint-dependent profile/base cache remains unaccounted. Metadata/log/error access must also be audited, and server-held random keys are not automatically public. The full 18-object inventory, including teacher labels, DPO refusal, coefficients and new LoRA state, is retained in the source JSON.

Reference replacement and controls. For a fixed noisy student $\widetilde{\theta}$, freeze canonical and same-view references $r_\theta, r_{\widetilde{\theta}}$. Comparison B changes only this source:

$$\Delta L_{\mathrm{ref}} = \left| L(\widetilde{\theta}; r_{\widetilde{\theta}}) - L(\widetilde{\theta}; r_\theta) \right|, \qquad r_g = \frac{\|g_{\mathrm{same}} - g_{\mathrm{can}}\|_2}{\|g_{\mathrm{can}}\|_2}, \tag{148}$$

where inverse-key alignment precedes gradient comparison and all measured denominators are nonzero. The recorded A criterion is $|\Delta L| \le 5 \cdot 10^{-4} + 10^{-4}|L_a|$ and either $r_g \le 10^{-4}$ or $\Delta_g \le 10^{-5}$. B flags use $|\Delta L| > 10^{-5} + 10^{-5}|L_a|$ or, for gradients, both $r_g > 10^{-4}$ and $\Delta_g > 10^{-6}$. Here L_a is the first-arm combined loss and Δ_g the absolute gradient-difference norm. These tolerances assess numerical agreement, not statistical significance or practical forgetting quality. NLL and KL are kept separate, and unsigned A/B/anchor differences cannot simply be added into a signed total change.

Table 27: Reference replacement at the same noisy student. NLL and explicit fixed-reference KL are separate branches. Δ_g is absolute gradient difference and r_g uses the canonical-reference norm. Loss/gradient flags denote numerical-tolerance exceedance, not statistical significance.

Trainer	Retain branch	$\lvert\Delta L_{\mathrm{ref}}\rvert$	Δ_g	r_g	L/G
GradDiff	NLL	0	0	0	no/no
GradDiff	fixed KL	1.02818×10^{-6}	0.33155	0.00966307	no/yes
DPO	NLL	0.00137424	0.291448	6.89473×10^{-5}	yes/no
NPO	NLL	0.000598907	0.0378785	2.98498×10^{-5}	yes/no
SimNPO	NLL	0	0	0	no/no
SimNPO	fixed KL	1.03191×10^{-6}	0.33155	0.108533	no/yes
UnDIAL	$\alpha_{\mathrm{r}} = 0$	0.000313282	0.116207	0.00062075	yes/yes
LoRA	NLL	0.000598907	0.000390556	2.98699×10^{-5}	yes/no

Comparison A tests clean-coordinate reparameterization with references, gradients and alias handling transformed consistently. The independent anchor comparison holds a canonical reference fixed while changing the student from clean to noisy; it is not the B comparison and does not include a training trajectory.

Table 28: Controls distinct from reference replacement. A compares clean and transformed coordinates with consistent reference/gradient mapping. Anchor changes the student from clean to noisy at a fixed canonical reference. Each relative difference uses its own first-arm gradient norm.

Trainer	A: $\lvert\Delta L\rvert$	A: r_g	Anchor: $\lvert\Delta L\rvert$	Anchor: r_g
GradDiff	1.2666×10^{-7}	6.79245×10^{-6}	5.48512×10^{-5}	0.00100771
DPO	0	4.05185×10^{-6}	0.0013361	0.000875045
NPO	0	4.80478×10^{-6}	0.000637054	0.000732279
SimNPO	7.45058×10^{-8}	8.9023×10^{-6}	4.0099×10^{-5}	0.00136156
UnDIAL	1.90735×10^{-6}	4.64746×10^{-6}	0.000221729	0.000761259
LoRA	0	4.76882×10^{-6}	0.000637054	0.000752124

All six A controls meet their recorded criteria; maximum loss/relative-gradient differences are 1.90735×10^{-6}/8.9023×10^{-6}. The independent identity repeat has differences 0/0. Full-parameter cases cover 146 unique tensors and 1,235,814,400 elements, with tied aliases counted once. Main B NLL is zero for GradDiff and SimNPO, but their fixed-KL relative gradient changes are 0.9663% and 10.8533%. DPO/NPO have loss changes with smaller relative gradient changes; UnDIAL crosses both thresholds. These are local reference effects at a single batch, not downstream performance differences or evidence of complete transcript closure.

LoRA Settings The LoRA diagnostic freezes its actual noisy base and adds rank-4, alpha-4 adapters to Q/V projections in all 16 layers: 32 modules, 64 trainable tensors and 425,984 elements. A uses fixed-seed Kaiming initialization, B is zero, and rank permutation $[2, 0, 3, 1]$ is aligned. B and anchor compare the 64 trainable factor gradients. The effective-direction record compares $s(B\,dA + dB\,A)$ over 32 weights.

J.2 Threshold factors and the limits of calibration

Table 29: Analytical coalition factors with separately reused historical NPO utility. These rows do not calibrate sensitivity, ϵ, a public profile, or an isolated deployment. t is the theoretical observation limit.

m	t	κ	Γ_t	Profile	FQ	MU
2	1	.01	1.000	legacy RMS	.919	.597
3	2	.01	2.815	legacy RMS	.766	.590
4	3	.01	5.584	legacy RMS	.766	.594

Enumeration of all nonempty strict coalitions gives $\Gamma_{m-1} = 1, 2.814815$ and 5.58375 for $m = 2, 3, 4$, respectively (exactly 1, $76/27$ and $4467/800$). Independent covariance solves agree. These analytical constants accompany reused historical NPO utility measurements. A numerical budget additionally requires the protected input, adjacency, conditional sensitivity, profile disclosure and complete-transcript accounting specified in Appendix B.4.

J.3 Copy-scale schedules and local-reference objectives

Table 30: Analytical copy-scale comparison. Heterogeneous-schedule utility reuses historical NPO. Line 2 and 3 are theoretically computed so no FQ/MU reported.

α schedule	Γ_t	Curvature factor	Product	FQ	MU
$(1, 1.5, 2)$ (supplied)	2.815	2.077	5.846	.766	.590
$(1, 1, 1)$	4.000	1.000	4.000	–	–
(c, c, c), matched marginal RMS	$4.000/c^2$	c^2	4.000	–	–

For heterogeneous, unit-equal and matched-marginal-RMS scales, the recalculated $(\Gamma_t, C, \Gamma_t C)$ are $(2.814815, 2.076923, 5.846154)$, $(4, 1, 4)$ and $(1.655172, 2.416667, 4)$. Here $C = (\sum_k \alpha_k)/(\sum_k \alpha_k^{-1})$ and $c = \sqrt{29/12} = 1.554563$. The product $\Gamma_t C$ is the factor optimized at a fixed RDP-order budget (Proposition 4).

GradDiff training. The reported schedule settings use master seed 505001, training seed 303001, $\kappa = .01$ and three FFN-transformed copies each. Every copy takes 40 updates of $-L_{\text{forget}} + L_{\text{retain}}$, with Forget01 rows 0–39 paired with Retain99 rows 0–39. Each loss averages valid shifted answer-plus-EOT targets. AdamW uses learning rate 10^{-5}, moments $(.9, .999)$, $\epsilon = 10^{-8}$, weight decay .01 and fresh state. Parameters, gradients and state are FP32, with eager execution and no additional masks, scheduler, clipping, AMP, TF32, fused or foreach optimizer. The layer profile has no newly imposed global floor. The reported endpoints use the completed training arms and subsequent streaming aggregation.

Explicit endpoint evaluation. Complete evaluations contain 40 forget, 400 retain, 100 real-author and 117 world-fact rows. The 400-row retain evaluation file differs from the 3,960-row training split. Forget/retain ratios use a paraphrased correct answer and five wrong alternatives; real-author/world-fact ratios use the answer and three alternatives. For correct mean answer NLL ℓ_+ and mean wrong-answer NLL $\bar{\ell}_-$, the stabilized ratio is

$$r = \frac{\exp(-\bar{\ell}_-)}{\exp(-\ell_+) + 10^{-10}}.$$

This is based on a geometric mean of wrong-answer probabilities. FQ is the nominal two-sided SciPy KS p-value comparing 40 target ratios with those from the pinned OpenUnlearning Llama-3.2-1B-Instruct Retain99 checkpoint on the same questions. The same reference observations are reused for FQ only. The paired-sample limitation and descriptive interpretation in Appendix H.2.2 apply here. The forget aggregate $\min(r, 1/(r + 10^{-10}))$ is a separate statistic.

MU is the equal-weight harmonic mean of nine component means: probability, ROUGE-L recall and $\max(0, 1 - r)$ on retain, real-author and world-fact data. It is zero when any component is zero. Retain probability is $\exp(-\ell)$; real-author/world-fact probability divides the correct-option mass by correct-plus-wrong mass plus 10^{-10}. ROUGE-L recall uses stemming. The historical metrics retain their separate definitions.

Evaluation uses FP32/eager, seed 303001 and teacher/generation batches of eight. Inputs longer than 512 tokens are rejected without truncation. Teacher/generation padding is right/left. The system prompt is "You are a helpful assistant.", the fixed date is "10 Apr 2025", and answer EOT is included. Generation permits 200 new tokens, disables sampling, uses one beam/group/return sequence, null temperature/top-p/top-k, unit repetition/length penalties and no early stopping. EOS IDs are 128001/128008/128009; pad ID is 128009. Generation cache is enabled, teacher cache disabled. The wrapper disables compilation, removes padded prompt width, skips special tokens, cleans tokenization spaces, applies EOS-text trimming and strips whitespace, without configured stopwords. Effective settings are captured before the first forward and checked on all 78 generation calls for each complete 617-sample utility evaluation. Saved FP32 displacements are combined by an FP64 streaming sum with one FP32 cast per parameter. Independent checks recomputed the exported scalars and aggregates; private generated text and GPU inference remained outside those checks.

Table 31: GradDiff/NLL copy-scale comparison at one master seed, three FFN-transformed copies, 40 updates per copy and $\kappa = .01$, with $c = \sqrt{29/12}$. Explicit endpoint evaluation is used. Historical and local-reference NPO protocols are reported separately.

Copy scales	FQ (KS p)	MU
Heterogeneous $(1, 1.5, 2)$	1.86034×10^{-23}	0.26473
Equal matched RMS (c, c, c)	1.86034×10^{-23}	0.282028

Saved updates. The following comparisons use a true clean DIRECT displacement with matched checkpoint, objective, 40-pair order, optimizer, precision and training seed. Scale rows use master seed 505001; profile rows use 707001. Two CPU passes use different chunk lengths/sums but share a parser; all 146 unique parameter tensors are included. Per-tensor sufficient statistics support the norm and cosine calculations, while complete tensors remain server-side. No local training is repeated.

Table 32: Saved GradDiff update statistics. Each three-copy inverse-scale update is compared with its matched clean DIRECT displacement. Scale and profile rows form separate blocks; this is offline arithmetic without new training or evaluation.

Saved-update setting	Relative update error	Update cosine
Scale: heterogeneous	0.325808	0.945463
Scale: matched RMS	0.308709	0.951161
Profile: task vector	0.367431	0.930073
Profile: public base	0.648083	0.763513

Independently specified local-reference NPO. Every arm freezes its own actual initialization's reference cache before any optimizer update, binding the 40 exact forget batches, masks and positions. It is never refreshed. The forget term is $-2\log\mathrm{sigmoid}(\beta(s_\theta - s_{\mathrm{ref}}))/\beta$, with $\beta = .1$ and *summed* valid shifted answer-plus-EOT negative log-likelihoods $s_\theta, s_{\mathrm{ref}}$. Retain NLL instead averages valid shifted targets. Both coefficients are one. Each local reference is private and specific to its noisy initialization.

Each arm takes 40 AdamW updates with the same settings as above, one example per role and a 512-token input cap. The master seed permutes the 40 paired rows identically within each comparison; the training RNG remains 303001. FFN keys and the inherited layer profile are used with $\kappa = .01$. This local-reference protocol is distinct from historical NPO. The clean seed-808001 run is reused for heterogeneous, matched-RMS and public-base settings, rather than counted three times.

Table 33: NPO local-reference protocol at master seed 808001 and $\kappa = .01$. The same clean control is reused for heterogeneous and matched-RMS experiments; $c = \sqrt{29/12}$. Update errors use that control, not a historical seed-0 NPO endpoint.

Endpoint	FQ	MU	Relative update error	Cosine
Clean DIRECT	0.265687	0.462547	–	–
Heterogeneous $(1, 1.5, 2)$	0.265687	0.462766	0.00517068	0.999987
Matched RMS (c, c, c)	0.265687	0.462802	0.00547754	0.999985

Table 34: NPO local-reference public-base endpoint at master seed 808001. It uses three 40-step training arms. The clean DIRECT MU is 0.4625473, reused rather than re-evaluated. Public-base scaling changes both profile shape and noise energy; historical NPO uses a separate protocol.

Scale / profile	FQ (KS p)	MU	ΔMU vs. DIRECT	Relative update error	Update cosine
Heterogeneous / public-base RMS	0.2656871	0.501869	0.03932172	0.5243668	0.8515173

Public-base uses $\alpha = (1, 1.5, 2)$ and $(6, 4, 3)/13$ weights. It trains three new arms at the same master seed 808001. Public-base changes both shape and total noise energy (Appendix J.8). The endpoint was serialized and reloaded, and its captured generation settings match the reused clean control.

J.4 Paired repeated runs

Table 35: **Mechanism controls with matched definitions.** Top: the inverse-scale decoder uses saved aligned GradDiff updates, with clean DIRECT as the displacement reference; its FQ value is 1.86034×10^{-23}. Bottom: aggregate minus paired clean MU over three seeds. Intervals are posthoc Student t over the three paired seeds.

Saved-update decoder	Rel. update error	Cosine	MU
Inverse scale	0.389034	0.921465	0.247535

Three-seed paired comparison	Mean ΔMU	Posthoc 95% CI
GradDiff, $\kappa = .01$	0.0307531	$[0.025269, 0.036237]$
GradDiff, $\kappa = .10$	-0.0575939	$[-0.32064, 0.20546]$
NPO local reference, $\kappa = .01$	0.00026873	$[-0.00032217, 0.00085963]$

Table 36: GradDiff repeated runs: mean $\pm$ sample SD over three paired seeds. MU intervals are posthoc paired Student-t intervals under the assumptions in the text; FQ averages are not pooled significance tests. FTR is not reconstructed from the new truth-ratio protocol.

Objective	Setting	FQ	MU	Paired ΔMU 95% CI
GradDiff	DIRECT	$5.085 \times 10^{-22} \pm 8.49 \times 10^{-22}$	0.2526 ± 0.176	–
GradDiff	$m = 3, \kappa = .01$	$1.86 \times 10^{-23} \pm 0$	0.2834 ± 0.174	$[0.025269, 0.036237]$
GradDiff	$m = 3, \kappa = .10$	$5.085 \times 10^{-22} \pm 8.49 \times 10^{-22}$	0.195 ± 0.239	$[-0.32064, 0.20546]$

GradDiff master seeds 606001–606003 determine paired data permutations, noise and keys; internal training seed stays 303001. Within a seed, DIRECT and the three noisy copies use the same permutation.

Table 37: All three paired GradDiff/NLL master seeds. The same three DIRECT results serve both noise levels and are counted once. At seed 606002, $\kappa = .10$ gives zero MU.

Master seed	Endpoint	FQ (KS p)	MU
606001	DIRECT (reused)	1.48827×10^{-21}	0.410622
606001	$\kappa = .01$	1.86034×10^{-23}	0.439129
606001	$\kappa = .10$	1.48827×10^{-21}	0.461654
606002	DIRECT (reused)	1.86034×10^{-23}	0.0632927
606002	$\kappa = .01$	1.86034×10^{-23}	0.0962127
606002	$\kappa = .10$	1.86034×10^{-23}	0
606003	DIRECT (reused)	1.86034×10^{-23}	0.284003
606003	$\kappa = .01$	1.86034×10^{-23}	0.314835
606003	$\kappa = .10$	1.86034×10^{-23}	0.123482

At $\kappa = .01$ and .10, mean paired MU changes are 0.0307531 and -0.0575939, with sample SD 0.00220755 and 0.105892. Relative decoded-update errors have mean/sample-SD 0.257403/0.232421 and 0.427838/0.208907. At high noise, seed 606002 has zero mean real-author ROUGE-L recall over 100 samples and thus zero MU.

For NPO, master seeds 808001–808003 use the local-reference protocol above; the first pair is reused, and only the two remaining pairs add training/evaluation.

Table 38: All six local-reference NPO endpoints for master seeds 808001–808003. The first pair is reused; only the two remaining pairs add new training and evaluation. Copies and repeated display are not independent seeds.

Master seed	Endpoint	FQ	MU
808001	DIRECT	0.265687	0.462547
808001	Three-copy aggregate	0.265687	0.462766
808002	DIRECT	0.265687	0.512321
808002	Three-copy aggregate	0.265687	0.51238
808003	DIRECT	0.0970748	0.471307
808003	Three-copy aggregate	0.0970748	0.471835

Table 39: Mean $\pm$ sample SD over three paired NPO seeds. The paired MU interval is posthoc. Student t with two degrees of freedom.

Endpoint / paired difference	FQ mean $\pm$ SD	MU mean $\pm$ SD	Paired MU 95% CI
DIRECT	0.209483 ± 0.0973484	0.482058 ± 0.0265715	–
Three-copy aggregate	0.209483 ± 0.0973484	0.482327 ± 0.0264189	–
Paired aggregate minus DIRECT	0 ± 0	$0.00026873 \pm 0.00023787$	$[-0.00032217, 0.00085963]$

All three NPO paired MU differences are positive; the mean 0.00026873 has interval $[-0.00032217, 0.00085963]$, which contains zero. The estimates are small and uncertain across three seeds. All reported paired intervals are posthoc Student t with two degrees of freedom; exact coverage assumes independent, normally distributed seed differences. The NPO interval method was recorded after the first seed and before the other two. FQ summaries describe the individual nominal p-values.

J.5 Saved-update decoding and reference endpoints

Table 40: Inverse-scale decoding of saved aligned updates. Update metrics use the matched clean DIRECT displacement; FQ/MU use explicitly configured endpoint evaluations.

Estimator	FQ	MU	Relative update error	Cosine
Inverse-scale decoder, $w_k \propto \alpha_k^{-1}$	1.86034×10^{-23}	0.247535	0.389034	0.921465

Let Δ_{direct} be the actual noise-free GradDiff displacement and $\widehat{\Delta}_k$ the aligned copy updates. Update error is $\| \sum_k w_k \widehat{\Delta}_k - \Delta_{\mathrm{direct}} \|_2 / \| \Delta_{\mathrm{direct}} \|_2$, accompanied by cosine. The clean run uses the same fixed 40 row pairs, optimizer and training seed as the paired FFN experiment. The clean displacement comes from a completed export. Canonical-noisy updates and truly inverse-mapped updates remain separate collections. For $\alpha = (1, 1.5, 2)$, weights are $(6, 4, 3)/13$.

Table 41: Saved-update comparison against a true clean DIRECT displacement. Canonical-noisy and actually inverse-mapped updates are distinct collections. The inverse-scale decoder and all three individual-copy estimates use the same clean reference.

Saved collection	Estimator	Relative update error	Update cosine
Canonical noisy	Inverse scale	0.388005	0.921897
Canonical noisy	Copy 1 only	0.484896	0.880746
Canonical noisy	Copy 2 only	0.516765	0.86309
Canonical noisy	Copy 3 only	0.385181	0.924932
Actual aligned	Inverse scale	0.389034	0.921465
Actual aligned	Copy 1 only	0.491701	0.877365
Actual aligned	Copy 2 only	0.51696	0.862979
Actual aligned	Copy 3 only	0.372592	0.929707

Table 42: Explicitly configured clean DIRECT endpoint evaluation. The clean endpoint uses the explicit-generation protocol; earlier aggregate evaluations have incompletely recorded defaults.

Endpoint	FQ (KS p)	MU
Clean DIRECT	1.86034×10^{-23}	0.254884

The inverse-scale error is 0.389034, compared with 0.491701, 0.51696 and 0.372592 for the three aligned copies. It improves on two copies and lies slightly above the lowest individual-copy error. These total displacement errors combine the residual sources in Theorem 2; their individual magnitudes remain unmeasured.

Table 43: Five preselected endpoints from one GradDiff saved-update collection. Each adds the indicated canonical-coordinate displacement to the clean checkpoint. The clean evaluation is reused; inverse-scale and individual-copy rows use explicit generation settings. Each individual-copy update is committed onto the clean state.

Endpoint	FQ (KS p)	MU
Clean DIRECT	1.86034×10^{-23}	0.254884
Inverse-scale aggregate	1.86034×10^{-23}	0.247535
Aligned copy 1	1.86034×10^{-23}	0.129661
Aligned copy 2	1.86034×10^{-23}	0.279671
Aligned copy 3	1.86034×10^{-23}	0.2492

The five endpoints add the stated displacement to the clean checkpoint, distinguishing the individual-copy controls from historical SINGLE-NOISED commitment. The inverse endpoint uses a new explicit-generation evaluation; its FQ/MU happen to match an earlier run with missing defaults. All five endpoints share extremely small nominal FQ, including clean DIRECT, so preserving the trainer's behavior leaves the forgetting limitations of this GradDiff protocol unresolved. Equal-compute clean ensembles and verified historical noisy-state controls remain unmeasured.

J.6 Optimizer compatibility, state and fixed history

The initial paired experiment uses three stored noisy views, $\alpha = (1, 1.5, 2)$, $\kappa = .01$, noise/key seeds 101001/202001 and training seed 303001. Each copy runs canonical and transformed arms from the same initialization, using Forget01/Retain99 rows 0–39 in order. This covers the complete Forget01 split and the first 40 Retain99 rows. The AdamW settings are those above, with one parameter group, no AMSGrad/foreach/fused implementation and no gradient checkpointing. All six arms completed 40 updates: 240 backward/update calls and 480 loss forwards. There are 146 unique parameter objects and 1,235,814,400 coordinates.

Error definitions. With $\Delta_{k,h} = \theta_{k,h} - x_{0,k}$ and $\widehat{\Delta}_{k,h} = T_k^{-1}(\theta'_{k,h} - T_k x_{0,k})$, first/final update discrepancies divide $\|\widehat{\Delta}_{k,h} - \Delta_{k,h}\|_2$ by $\|\Delta_{k,h}\|_2$ at $h = 1, 40$. In contrast, trajectory error is

$$\max_{1 \le h \le 40} \frac{\|T_k^{-1}\theta'_{k,h} - \theta_{k,h}\|_2}{\|\theta_{k,h}\|_2}.$$

The summary averages per-copy quantities and, for trajectory, per-copy maxima at one training seed. All observed denominators are nonzero. Numerical tolerances are .005 at step 1 and .05 at step 40.

Table 44: Three-copy optimizer/key summaries on one fixed 40-pair traversal. Mechanism columns average per-copy quantities, not seeds or aggregated displacements. Dashes are unavailable measurements.

Optimizer	Key family	Coupling	Step defect	Traj. error	Update error	Cosine	FQ	MU
SGD	continuous RoPE rotations	coupled	0.0223	1.29×10^{-7}	0.0219	1	–	–
AdamW	continuous RoPE rotations	coupled	0.251	–	0.687	0.762	–	–
AdamW	RoPE quarter turns	coupled	0.000752	9.7×10^{-5}	0.0246	1	–	–
AdamW	FFN permutations	no added masks	6.98×10^{-4}	2.12×10^{-4}	0.0538	0.998	1.86×10^{-23}	0.248
RMSProp	FFN permutations	coupled	0.000204	–	0.123	0.992	–	–

Table 45: All three FFN-permutation/AdamW copies, each with 40 updates per arm. Trajectory error is the within-copy maximum parameter-state-normalized error; update error is displacement-normalized. Copies 1 and 3 exceed the final tolerance .05. These are coded initializations at one training seed.

Copy	One-step defect	Max trajectory error	Final update error	Update cosine	Final criterion
1	6.553212×10^{-4}	2.565012×10^{-4}	0.06499222	0.9978879	exceeds .05
2	6.620816×10^{-4}	4.250445×10^{-5}	0.01090644	0.9999405	within .05
3	7.76739×10^{-4}	3.377592×10^{-4}	0.08538667	0.9963514	exceeds .05

Parameter-state and displacement errors use different normalizations, so close full parameter states can coexist with appreciable relative update differences. Theorem 3 describes the ideal coupled update rule.

Other keys and optimizers. SGD uses zero momentum. RMSProp uses $\alpha = .99$, $\epsilon = 10^{-8}$, zero momentum and no centering. Learning rate 10^{-5}, weight decay .01, one group, fresh state and native optimizer weight-decay semantics are retained. The earlier continuous-rotation AdamW and RMSProp slice includes copy 1 only; later endpoint controls add the stated remaining copies and a distinct FFN/RMSProp configuration.

Table 46: Paired 40-step optimizer/key controls. The final column compares first and final defects with .005 and .05, respectively. The criteria assess numerical agreement. All rows contain 40 updates per arm; denominators are specified in the text.

Optimizer/key	Copy	Step 1	Max trajectory	Final update	Cosine	First/final criterion
SGD/continuous	1	0.0222417	1.28688×10^{-7}	0.0219298	0.99976	exceeds/within
AdamW/quarter-turn	1	0.000690042	5.57865×10^{-5}	0.0141352	0.9999	within/within
AdamW/quarter-turn	2	0.000683839	3.62335×10^{-5}	0.00929734	0.999957	within/within
AdamW/quarter-turn	3	0.000882534	0.000198922	0.0502882	0.998735	within/exceeds
SGD/continuous	2	0.0222864	1.28948×10^{-7}	0.0219764	0.999759	exceeds/within
SGD/continuous	3	0.0222543	1.28437×10^{-7}	0.0218948	0.99976	exceeds/within
AdamW/continuous	1	0.251112	0.0026251	0.665148	0.775979	exceeds/exceeds
RMSProp/continuous	1	0.227483	0.00956637	0.344471	0.940987	exceeds/exceeds

All continuous-SGD first-step discrepancies exceed .005, while final discrepancies remain below .05. Quarter-turn AdamW copy 3 is marginally above the endpoint threshold at .0502882. These are theoretically compatible combinations, whose numerical agreement is assessed separately from exact arithmetic. Continuous-rotation AdamW and RMSProp instead have a structural basis-dependent

defect. Shared-gradient controls isolate optimizer arithmetic at a common history; full-model trajectory discrepancies require additional diagnosis.

Table 47: Forty-step endpoint controls under the original .005 first-step and .05 final tolerances. Continuous-AdamW copy 1 is reused from the earlier control. Saved first/final canonical states do not supply missing intermediate trajectory maxima. Both mask rows use copy 1 and mask seed 909001.

Optimizer/key/masks	Copy	Step 1	Max trajectory	Final update	Cosine	First/final criterion
AdamW/rotation	1	0.251112	0.0026251	0.665148	0.775979	exceeds/exceeds
AdamW/rotation	2	0.252036	-	0.71745	0.74235	exceeds/exceeds
AdamW/rotation	3	0.250527	-	0.677003	0.768273	exceeds/exceeds
RMSProp/FFN	1	0.000188972	-	0.120542	0.992722	within/exceeds
RMSProp/FFN	2	0.000190969	0.00225695	0.0814605	0.99668	within/exceeds
RMSProp/FFN	3	0.000232661	0.00464839	0.166949	0.986035	within/exceeds
AdamW/FFN coupled	1	0.0020156	2.55567×10^{-6}	0.000612863	1	within/within
AdamW/FFN independent	1	1.37804	-	1.26149	0.1677	exceeds/exceeds

Continuous-AdamW copies 2/3 reuse saved canonical endpoints while running new transformed arms. FFN/RMSProp copy 1 reuses canonical endpoints; copies 2/3 run both arms. All three FFN/RMSProp first-step errors meet tolerance, and all final errors exceed it. Trajectory maxima remain unavailable where intermediate histories are missing. The additional mask experiment uses copy 1, dropout .1 and seed 909001: coupled masks meet both tolerances, whereas independent masks exceed them. This single paired realization illustrates the role of randomness coupling; Proposition 6 concerns the distribution across random realizations.

Early aggregate evaluations. The canonical-noisy and aligned endpoints are $\theta_{\mathrm{src}} + \sum_k w_k \Delta_{k,40}$ and $\theta_{\mathrm{src}} + \sum_k w_k \widehat{\Delta}_{k,40}$, with $w = (6, 4, 3)/13$. Both aggregates use noisy-copy displacements, with the completed aligned evaluation counted once.

Table 48: Early aggregate-endpoint evaluation after the three-copy FFN/AdamW experiment. Both endpoints add inverse-scale-weighted updates to the source checkpoint. FQ uses the pinned retain-only reference and stabilized-ratio KS test.

Endpoint	FQ (KS p)	MU
Canonical matched aggregate	1.86034×10^{-23}	0.2489758
Aligned aggregate	1.86034×10^{-23}	0.2475354

Table 49: The nine component means used for the two early aggregate-endpoint MU values. Retain, real-author and world-fact counts are 400, 100 and 117. Their harmonic mean gives MU.

Subset	Component	Canonical	Aligned
Retain	Probability	0.07601443	0.07678595
Retain	ROUGE-L recall	0.1406141	0.1423907
Retain	TR_V2 utility	0.6185419	0.627271
Real authors	Probability	0.4349007	0.4343789
Real authors	ROUGE-L recall	0.1961667	0.1811667
Real authors	TR_V2 utility	0.6383092	0.640523
World facts	Probability	0.4687555	0.4670832
World facts	ROUGE-L recall	0.6512821	0.6413105
World facts	TR_V2 utility	0.6127982	0.6117383

These early evaluations share the stabilized ratios, row counts and component definitions above. They explicitly set 200 new tokens, no sampling, EOS IDs 128001/128008/128009 and pad 128009, with FP32/eager, seed 303001 and batches of eight. Both KS statistics are 1.0 and the MU difference is -0.001440438. The tiny FQ values indicate poor agreement with the retain-only reference under this descriptive metric.

Table 50: Conditional FP64 AdamW state controls from real saved m_{39}, v_{39}, g_{40}. The calculation reuses a single saved history. Moment differences are bias-corrected, and relative defect uses the whole-model direction norm. A dash in the ratio is undefined $0/0$, not zero.

State source	Transform rule	Formal defect	$\|\delta_m\|_2$	$\|\delta_v\|_\infty$	Rel. defect ↓	Bound ratio ↓
Fresh zero	zero in both coordinates	0	0	0	0	–
Nonzero accumulated	correctly mapped	0	0	0	0	–
Nonzero accumulated	deliberately unmapped	Eq. (86)	150.6069	452.026	1.150945	2.546875×10^{-16}

The source is the real copy-1 canonical state after step 39 and its step-40 gradient. FP32 saved moments/gradient are updated in FP64. Fresh-zero uses this same g_{40} with zero moments and first-step correction, rather than the original trajectory's first gradient. Nonzero branches use step-40 correction and correctly mapped or deliberately unmapped moments; the transformed gradient is mapped in every branch. Weight decay cancels in this same-state direction comparison. The first two controls have zero discrepancy and undefined $0/0$ ratios. The unmapped negative control has absolute discrepancy 7899.63, learning-rate-scaled discrepancy 0.0789963, floor 4.538788×10^{-19}, floor bound 4.295863×10^{29} and smaller Hölder bound 3.101695×10^{19}. The large bounds are non-informative at the reporting tolerance. Correct state mapping gives zero discrepancy in this conditional arithmetic check, while deliberately unmapped state produces a substantial defect. The calculation reuses saved tensors without model forward/backward execution.

Fixed-history rotation diagnostics. A clean GradDiff history supplies native FP32 moments and cross moments at $h = 10, 40$. Each node is evaluated in FP64 over 64 Q/K/V/O tensors with 32 architecture-compatible SO(2) keys, shared GQA rotations and uniform plane angles. NumPy PCG64 uses seed sequences $(20260921, k)$, $k = 0, \ldots, 31$. Unchanged-coordinate directions enter the whole-model normalization; the 32 keys index conditional arithmetic evaluations of each saved history.

Table 51: One clean GradDiff history at each node, evaluated with 32 architecture-compatible planar keys. Mean relative defect uses the whole-model canonical adaptive-direction norm. The 95th percentile uses linear interpolation across keys; keys are not training seeds.

h	Mean $\|q\|$	Mean $\|q\|^2$	q p95	Mean relative defect	η mean $\|q\|$
10	1448.48	2.09816×10^6	1457.41	0.13352	0.0144848
40	793.372	629466	800.355	0.101188	0.00793372

Table 52: Fixed-history moment and bound diagnostics. U_h is the deterministic squared Hölder envelope of Proposition 8, distinct from the Haar-specific expectation bound B_h. Both reported envelopes are non-informative at the reporting tolerance.

h	Observed plane spectral minimum	U_h	$\sqrt{U_h/.05}$
10	2.8183×10^{-20}	4.44101×10^{33}	2.98027×10^{17}
40	1.34477×10^{-18}	7.08019×10^{34}	1.18997×10^{18}

Identity and exact quarter-turn controls are zero at both nodes. The observed plane eigenvalues and rotated diagonal entries are nonnegative. Spectral minima concern rounded saved histories; the additional diagonal samples are canonical. The deterministic envelope $U_h = 4\epsilon^{-4} \sum_b \|A_b\|_{\mathrm{op}} \|\widehat{m}_b\|_2^2$ and the Haar-specific conditional expectation bound are distinct quantities. The measured rotation defects are consistent with basis dependence, but the evaluated envelopes are too large to certify negligible mismatch. Full-trajectory control additionally requires the stability conditions in Proposition 10.

J.7 Initialization arithmetic and stochastic coupling

The stored three-copy views use the same full/public checkpoints, $\alpha = (1, 1.5, 2)$ and $\kappa = .01$. The scan covers 146 unique tensors and 1 235 814 400 coordinates. Stored $x_{0,k}$ agrees with the actual FP32 addition of θ and serialized scaled noise $\varepsilon_k^{\mathrm{ser}}$; subsequent subtractions/norms use FP64.

Original pre-scaling e_k were not saved. Recovered $u_k = \varepsilon_k^{\text{ser}}/\alpha_k$, realized $v_k = (x_{0,k} - \theta)/\alpha_k$ and unweighted $\varepsilon_k^{\text{ser}}$ are therefore distinct quantities.

Table 53: Serialization diagnostics for the shared three-copy initializations. The relative residual divides the norm of the vector sum by the sum of the individual norms. Recovered base vectors are not the unsaved pre-scaling noise; the unweighted scaled sum need not vanish.

Vectors summed	Absolute residual	Relative residual
Recovered $\varepsilon_k^{\text{ser}}/\alpha_k$	5.22414×10^{-9}	3.90213×10^{-8}
Actual $(x_{0,k} - \theta)/\alpha_k$	2.01812×10^{-5}	1.50741×10^{-4}
Scaled $\varepsilon_k^{\text{ser}}$, unweighted	0.03864818	0.1924525

Table 54: Per-copy initialization norms for the shared stored views. Rounding discrepancy is $\|x_{0,k} - \theta - \varepsilon_k^{\text{ser}}\|_2$, with subtraction and norm accumulation in FP64. These are not trained-displacement errors.

Copy	$\|\varepsilon_k^{\text{ser}}\|_2$	$\|x_{0,k} - \theta\|_2$	Rounding discrepancy
1	0.04462662	0.04462662	1.55035×10^{-5}
2	0.06693885	0.06693885	1.55037×10^{-5}
3	0.08925389	0.08925389	1.55013×10^{-5}

The actual-initialization residual includes floating-point addition; an unweighted scaled sum need not cancel. These initialization-level quantities are checked from per-tensor statistics and selected slices. Local Hessians, Gaussian escape probabilities, conditional-history sensitivity and key-uniform stability constants remain unmeasured.

At the first stored noisy initialization, mask seeds 909001–909003 each drive coupled and independent three-step FFN/AdamW controls. Injected hidden-channel dropout is .1, with zero native dropout. Coupled masks are permuted consistently; independent masks are separate. Both arms use the same first three fixed row pairs and the paired optimizer settings.

Table 55: Three-step FFN-mask controls at one noisy initialization. Gradient, parameter and moment discrepancies each use their own canonical norm. Parameter-state discrepancy is not displacement error or decoded-ensemble error.

Mask seed	Condition	Gradient	Parameters	First moment	Second moment
909001	coupled	3.14732×10^{-5}	9.90299×10^{-7}	2.6975×10^{-5}	3.30574×10^{-5}
909001	independent	1.31335	0.00132408	1.25313	0.928108
909002	coupled	3.1642×10^{-5}	8.68717×10^{-7}	3.4685×10^{-5}	3.68925×10^{-5}
909002	independent	1.16121	0.00126678	1.17028	0.843169
909003	coupled	3.68911×10^{-5}	9.8419×10^{-7}	2.70985×10^{-5}	2.874×10^{-5}
909003	independent	1.99565	0.00127839	1.50216	1.22383

These six observations quantify local gradient, parameter and moment discrepancies. Decoded-ensemble variance and endpoint FQ/MU require separate measurements, and the historical Single-Noised commit rule remains a separate protocol question.

J.8 Layer profiles and perturbation energy

The shared profile is constructed from the canonical initialization and public base. Across 16 layers, $b_\ell = \text{RMS}(\theta_\ell - \theta_{\text{base},\ell})$ and $\sigma_\ell = .01 b_\ell$. The minimum observed b_ℓ is 9.59801×10^{-5}; zero and near-zero ($0 < |\sigma| \leq 10^{-12}$) layer counts are 0 and 0.

Table 56: Offline distributions of $\sigma_\ell = \kappa b_\ell$ over 16 shared layer blocks, $\kappa = .01$. Medians use linear interpolation.

Scale rule	Minimum	Median	Maximum
Shared raw task-vector	9.59801×10^{-7}	1.46634×10^{-6}	1.75193×10^{-6}
σ floor 10^{-6}	1×10^{-6}	1.46634×10^{-6}	1.75193×10^{-6}
Public-base layer RMS	1.87872×10^{-4}	1.9545×10^{-4}	2.11137×10^{-4}
Public scalar	1.96667×10^{-4}	1.96667×10^{-4}	1.96667×10^{-4}

The offline floor $\max(\sigma_\ell, 10^{-6})$ changes 3 blocks. Public alternatives use $.01\,\mathrm{RMS}(\theta_{\mathrm{base},\ell})$ or one scale from that layer subspace. Applying whole-checkpoint RDP additionally requires those coordinates to be public or fixed across adjacent inputs.

Trained GradDiff profiles. Master seed 707001 couples noise/key construction for the two profiles. Both use the same row order, three FFN-transformed copies and 40 updates per copy. Noise covers layer tensors only, excluding tied embedding/head and final normalization. Table 57 uses the explicit generation protocol; Table 32 separately gives true clean-DIRECT update comparisons.

Table 57: GradDiff/NLL layer-profile comparison at one master seed, $m = 3$, $\alpha = (1, 1.5, 2)$ and $\kappa = .01$. Changing the profile changes both its shape and energy; noise is applied to layer tensors.

Layer profile	FQ (KS p)	MU
Task-vector layer RMS	1.86034×10^{-23}	0.29979
Public-base layer RMS	1.86034×10^{-23}	0.40008

Public-base minus task-vector MU is 0.10029, with equally tiny FQ. Public-base clean-relative update error is 0.648083, compared with 0.367431 for task-vector scaling. Thus higher utility accompanies a less faithful update in this single-seed comparison. Both profile shape and perturbation energy change, leaving their individual contributions unresolved.

NPO public-base scaling. The public-base endpoint in Table 34 uses $\sigma_\ell = .01\,\mathrm{RMS}(\theta_{\mathrm{base},\ell})$ without normalization to task-vector energy. Its marginal layer scales range from 0.0001878724 to 0.0002111372. The expected pre-addition noise-energy ratio is 18899.98, computed from $\sum_\ell d_\ell \sigma_\ell^2$ at equal copy scale; this expectation differs from actual FP32 norms.

Table 58: Public-base perturbations before FFN reparameterization, with $\alpha = (1, 1.5, 2)$. Sampled ϵ_k and realized FP32 $x_{0,k} - \theta$ are distinct vectors. Norms cover 146 unique parameter objects: 144 perturbed layer tensors and two unnoised inherited objects.

Copy	α_k	$\|\epsilon_k\|_2$	$\|x_{0,k} - \theta\|_2$
1	1	6.134956	6.134956
2	1.5	9.202847	9.202847
3	2	12.2701	12.2701

MU is 0.501869, or 0.03932172 above reused clean, while relative update error is 0.5243668. Shape and magnitude vary together. Only 144 layer tensors are noised; the tied embedding/head object and final normalization remain unnoised. Transcript closure, private-reference accounting and full protected-coordinate coverage remain additional requirements for privacy calibration.

J.9 Bounded recovery and returned-update privacy

Four-prompt initial suite. The pinned full/public checkpoints are processed sequentially with FP32/eager, seed 1234, no gradients or TF32. The initial records use PyTorch 2.12.1+cu132 and Transformers 5.14.1. The prompts are "The capital of France is", "Two plus two equals", "Water freezes at", and "A short sentence about science:". Tokenization adds special tokens, applies no chat template and uses a cap of 16; lengths $(6, 5, 4, 7)$ give 22 positions including BOS, not independent

examples. KL is the full-vocabulary $D_{\mathrm{KL}}(p_{\mathrm{canonical}} \| p_{\mathrm{endpoint}})$, token-weighted over these positions, with FP64 stable-log-softmax reduction. The two-view case uses noise seed 2026092001, while the three-view case reuses the original stored initializations; scales are $(1, 2)$ and $(1, 1.5, 2)$, respectively, at $\kappa = .01$. They are not paired copy-count estimates. The true-key oracle inverse-transforms and averages parameters in FP64 before an FP32 cast. Parameter RelErr uses the unique parameter vector; TaskCos compares recovered-minus-public with canonical-minus-public. No text generation or unlearning is performed. Harmonic logit averaging combines raw logits before softmax and is not parameter averaging.

Table 59: Initial four-public-prompt recovery observations (22 positions). Logit averaging and parameter recovery are separate controls. True-key errors are duplicated in the following oracle table for completeness, not additional runs. General alignment attacks are outside this initial suite.

m	Views s	κ	Control	RelErr	KL	TaskCos
2	2	.01	Harmonic logit average (no alignment)	–	3.66242×10^{-11}	–
2	2	.01	True-key oracle harmonic recovery	2.41081×10^{-8}	8.16048×10^{-12}	0.999999999999
3	3	.01	Harmonic logit average (no alignment)	–	3.34787×10^{-11}	–

Table 60: True-key recovery on four public prompts (22 input positions), without unlearning. RelErr and TaskCos use the complete unique parameter vector. The $m = 2$ row repeats the completed subcase in Table 59

m	Absolute parameter error	RelErr	KL	TaskCos
2	1.70338×10^{-5}	2.41081×10^{-8}	8.16048×10^{-12}	0.999999999999
3	1.36803×10^{-5}	1.93618×10^{-8}	6.70233×10^{-12}	0.999999999999

The two-view subcase completed before a memory stop in the three-view branch. For $m = 4$, noise/key seeds are 2026092104/2026092105; $\alpha = (1, 4/3, 5/3, 2)$ and weights are $(20, 15, 12, 10)/57$. Newly generated noise prevents a controlled copy-count comparison.

Table 61: Four-view recovery diagnostic at $\kappa = .01$ on four public prompts (22 positions). New noise/key draws prevent a paired comparison with $m = 2, 3$. The oracle uses true keys; no GPM or learned alignment is performed in this case.

m/s	Recovery rule	Oracle ℓ_2 error	RelErr	KL	TaskCos
4 / 4	Harmonic logits, no alignment	–	–	2.912208×10^{-11}	–
4 / 4	True-key oracle	1.173316×10^{-5}	1.660602×10^{-8}	6.693997×10^{-12}	0.999999999999292

Small KL demonstrates functional proximity on these prompts. True-key oracles supply an authorized reference for the learned-alignment tests below. Exported full-vocabulary and parameter slices support limited checks; complete checkpoints and all-position logits remain outside the independent reconstruction.

Public-anchor alignment on a distinct suite. The four prompts are "What is the capital of France?", "Explain why water freezes.", "Write a short greeting.", and "What is two plus two?". Special tokens without a chat template give 28 positions, not pooled with the initial suite. Noise/permutation/rotation roots are 1115001/1116001/1117001 for $m = 2$ and 1118001/1119001/1120001 for $m = 3$. At $\kappa = .01$ the scales are $(1, 2)$ and $(1, 1.5, 2)$. These draws are not coupled across copy counts.

FFN alignment applies Hungarian assignment to normalized gate/up/down descriptors over all 16 layers. Attention alignment fits shared-GQA, RoPE-compatible SO(2) planes from Q/K/V/O using constrained Procrustes. Only published and public-base tensors enter fitting; true keys and the private model are separate scoring inputs. The scoring inputs are separated logically within the same-process evaluation.

Table 62: Parameter-space recovery on four public prompts (28 positions). Public-anchor alignment combines all-16-layer FFN Hungarian assignment with shared-GQA SO(2) fitting; it is not general GPM. Unaligned parameters are averaged before inference; true keys are separate scorer controls. s is the number of observed views.

m/s	Candidate	Relative parameter error	Task cosine	Token-mean KL
2/2	Unaligned parameters	1.06245	0.01226577419309	7.20964
2/2	True-key oracle	2.39892×10^{-8}	0.9999999999985	9.71553×10^{-12}
2/2	Public-anchor alignment	1.25631×10^{-5}	0.9999995951267	9.29272×10^{-12}
3/3	Unaligned parameters	0.994633	0.01310600084609	8.50271
3/3	True-key oracle	1.92122×10^{-8}	0.9999999999991	5.77483×10^{-12}
3/3	Public-anchor alignment	1.25631×10^{-5}	0.9999995951272	7.11686×10^{-12}
3/2	Unaligned parameters	1.05092	0.01239705416927	6.74602
3/2	Two-view true-key oracle	3.7897×10^{-5}	0.999996315861	3.45658×10^{-8}
3/2	Sequential pairwise + anchor	3.99251×10^{-5}	0.9999959110032	3.45279×10^{-8}
2/1	One-view public-anchor alignment	6.43975×10^{-5}	0.9999893620253	8.57031×10^{-8}

The one-view result uses the preselected first two-view copy, not the best copy. The two-of-three-view result aligns copy 2 to copy 1 and copy 1 to the public base, carrying the map to copy 2; weights renormalize to $(.6, .4)$ without a third view. Its partial true-key oracle retains noise.

Table 63: Output-space controls using the same published views and 28 input positions as the parameter-space recovery test. Raw logits are inverse-scale averaged before softmax; these are neither parameter nor probability averages. The two-view subset reuses saved logits, not an extra model run.

m/s	Combination	Token-mean KL
2/2	No alignment, weighted logits	4.80292×10^{-11}
3/3	No alignment, weighted logits	3.92627×10^{-11}
3/2	No alignment, weighted logits	3.46377×10^{-8}

Unaligned parameter averages have large KL, whereas learned alignment and subset-view controls recover closely matching outputs. Recovery succeeds on these prompts with the available public base and unperturbed outside-layer tensors. The tests measure functional accessibility; Theorem 1 instead quantifies divergence for neighboring checkpoints under a declared metric and budget. The earlier noncontiguous-serialization failure is recorded separately.

Thirty-two-prompt constructed endpoints. The expanded suite contains 32 fixed public prompts and 457 input positions, without generation or unlearning. Micro KL weights positions, macro KL weights prompts, and signed tiny numerical KL values are not clipped. Argmax discrepancies are counts out of 457. Logit combinations reuse model outputs rather than adding forwards.

Table 64: Constructed-endpoint observations on 32 public prompts and 457 positions. Micro/macro KL weight positions/prompts equally.

m/s	Actual endpoint / combination	Micro KL	Macro KL	Argmax disagreements
2/1	Published view 1	2.0272×10^{-7}	2.11003×10^{-7}	0
2/1	Published view 2	8.10108×10^{-7}	8.43355×10^{-7}	0
3/1	Published view 1	1.84766×10^{-7}	1.92875×10^{-7}	0
3/1	Published view 2	4.27491×10^{-7}	4.49558×10^{-7}	0
3/1	Published view 3	7.21074×10^{-7}	7.31456×10^{-7}	0
2/1	Public-anchor aligned view 1	2.02632×10^{-7}	2.10903×10^{-7}	0
3/1	Public-anchor aligned view 1	1.84873×10^{-7}	1.92992×10^{-7}	0
2/2	Serialized unaligned parameters	7.25498	7.25936	450
3/2	Serialized unaligned parameters	6.4462	6.45454	456
3/3	Serialized unaligned parameters	8.74694	8.73364	457
2/2	Inverse-scale raw logits	1.47438×10^{-11}	1.33631×10^{-11}	0
3/2	Inverse-scale raw logits	6.48136×10^{-8}	6.57142×10^{-8}	0
3/3	Inverse-scale raw logits	1.15643×10^{-11}	1.04382×10^{-11}	0

Table 65: Correctly serialized endpoints on 32 public prompts and 457 input positions. Three rows use repaired true-key oracles; the saved learned endpoint is evaluated without refitting. Micro KL weights positions, macro KL weights prompts, and argmax entries count disagreements out of 457.

m/s	Endpoint	Micro KL	Macro KL	Argmax disagreements
$2/2$	Corrected true-key oracle	5.222728×10^{-12}	4.760344×10^{-12}	0
$3/2$	Corrected partial true-key oracle	6.49403×10^{-8}	6.586741×10^{-8}	0
$3/3$	Corrected true-key oracle	4.927644×10^{-12}	4.458855×10^{-12}	0
$2/2$	Saved learned endpoint	5.591727×10^{-12}	5.244834×10^{-12}	0

The corrected oracles require 96 new forwards. With 352 previously valid observations they form a corrected 448-observation analysis set; the original 96 faulty observations remain in the failure record. The full grid history contains 544 actual forwards across the original and corrected constructions.

J.10 Component costs and deployment limits

The following observations concern sequential same-process execution, where one process or UID can access all views. Enforced container, enclave or provider isolation remains a deployment requirement. Reported file sizes and loaded-to-last-forward intervals measure storage and component residency, respectively; network traffic and full exposure lifetimes require separate instrumentation.

Table 66: Component timing sums over 17 recorded copy lifecycles, including validation overhead.

Measured component	Sum (s)
Model/optimizer cleanup	29.4236
Key, zero-update gate and optimizer setup	165.577
Measured stage total	1653.36
Model transfer to GPU	16.5254
CPU noise generation	244.822
Return transfer and inverse displacement	60.2313
CPU source-model load	74.6983
Statistics and identity checks	194.145
Local update region	581.113
Update serialization	154.973
View serialization	131.696

Table 67: Same-process timing for the NPO experiment. The nested three-copy region includes DIRECT-comparison statistics but excludes final endpoint commitment.

Measured boundary	Elapsed seconds
Comparison block	3109.49
Extra clean control	170.529
Nested three-copy region	2937.96

The NPO timing record separates the whole comparison block 3109.49, extra clean control 170.529 and nested three-copy region 2937.96. Sampled process RSS is 11.3994 GiB, with sampled CUDA allocated/reserved peaks 18.7213/23.873 GiB.

K Optional Client-Update DP Extension

A formal client-data guarantee is orthogonal. The client must clip per-example or per-user gradients, add calibrated noise, and account for sampling, local steps, copies, and rounds. Fix a client-dataset adjacency $D \simeq D'$ and hold the server's initial information fixed across neighboring worlds. Order the server-observed releases (k, r) by their release sequence. For every common feasible server-observed history h, including prior updates and feedback, require the conditional release kernel $K_{k,r}$ to satisfy

$$D_\lambda(K_{k,r}(D, h) \| K_{k,r}(D', h)) \le \rho_{k,r}(\lambda), \qquad \lambda > 1,$$

with deterministic bounds uniform over h and neighboring datasets. As in Appendix B.4, marginal guarantees alone do not suffice. Conditional adaptive composition then bounds the complete server-visible transcript $\mathcal{V}$ by

$$D_\lambda(\mathcal{V}(D)\|\mathcal{V}(D')) \le \rho_{\text{client,total}}(\lambda) := \sum_{r=1}^{R}\sum_{k=1}^{m}\rho_{k,r}(\lambda). \tag{149}$$

The right-hand side is an accounting upper bound, not an equality for the true divergence; every copy-wise release must be charged, including repeated use of the same client records. Authorized aggregation is post-processing only after the server has already observed all copy-wise releases and cannot retroactively reduce this cost. To account only for an aggregate, secure aggregation, MPC, or a trusted mechanism must prevent the server from seeing individual updates. This extension is not part of the reported experiments.

L Broader Impacts and Responsible Use

Threshold-private publication can reduce checkpoint exposure in legitimate deletion workflows, but the guarantee can be misrepresented. View isolation may fail, a checkpoint-dependent scale profile may leak information, behavioral forgetting metrics are not deletion certificates, and returned updates may reveal the records being removed. Conversely, unlearning may be misused to obscure provenance or suppress audits. Deployments should authenticate requests, record executor assignments and cumulative privacy accounting, test full-view recovery and update inversion, and avoid describing the protocol as end-to-end secure beyond its stated channel and threshold.