STORAGE IS NOT STRATEGY: STATE-CONDITIONED SUPPORT CONTROL FOR LLM UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many localized large language model (LLM) unlearning methods select a small parameter subset from a localization signal and keep it fixed during optimization. The parameters most associated with a target, however, need not be the best ones to update, and candidate interventions can change value as optimization proceeds. In a controlled experiment, a storage-localization score reaches an area under the receiver operating characteristic curve (AUROC) of 0.981, yet storage identity agrees with the better intervention on only $17/36$ targets, while low-rank adaptation (LoRA) wins $35/36$. We introduce *Intervention Score*, which ranks editable groups by the predicted effect of the actual unlearning update while accounting for collateral damage, and use it to form the static intervention-value baseline (STATIC-IV). We then introduce selective dynamic intervention re-ranking (DIR-R), which revisits that subset only when a calibrated probe justifies the comparison. On the Natural-TOFU dataset, our method has positive descriptive margins in $19/20$ comparisons between methods and objectives, although several are near zero. On the LACUNA localization-precision benchmark, our mean terminal utility is higher in all six negative preference optimization (NPO) and SimNPO comparisons: NPO margins range from $+0.431$ to $+0.848$, and SimNPO margins range from $+0.503$ to $+0.571$. The gradient-difference (GradDiff) objective reveals substantial field dependence. Relative to STATIC-IV, the primary four-field GradDiff evaluation has six wins, six ties, and no losses, with mean and median paired gains of $+0.165$ and $+0.0025$. The evidence supports separating localization, initial intervention selection, and checkpoint-dependent support revision.

1 INTRODUCTION

Machine unlearning studies how to remove the effect of specified training data from a trained model while preserving utility on data that should be retained (Cao and Yang, 2015; Bourtoule et al., 2021). A natural reference is retraining without the data to be forgotten, but retraining is prohibitively expensive for LLMs (Maini et al., 2024; Shi et al., 2025). Practical LLM unlearning therefore modifies an existing model in place. Recent work develops direct gradient objectives and preference losses, including negative preference optimization (NPO) and its reference-free simplification SimNPO (Jang et al., 2023; Zhang et al., 2024; Fan et al., 2025b). A complementary line asks where the update should act. Localized LLM unlearning selects a small parameter subset from localization or attribution signals and restricts optimization to that subset (Jia et al., 2024; Tian et al., 2024; Guo et al., 2025; Lee et al., 2025).

Localization-first designs often rely on two operational assumptions. The first treats localization-derived importance as a proxy for where a particular unlearning objective should intervene, although localization quality does not necessarily predict intervention efficacy (Hase et al., 2023; Lee et al., 2025). The second keeps the selected parameter subset fixed as optimization proceeds, even though continual-unlearning work adapts masks dynamically and post-training analysis warns that static mechanistic localization can become stale as parameters evolve (Wuerkaixi et al., 2025; Chen et al., 2026). We therefore separate the initial intervention subset from the decision to revise it after optimization changes the model.

Our controlled experiment directly tests the first assumption by separating localization accuracy from intervention effectiveness. We construct synthetic facts for which the parameter interface used

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.

to inject each fact is known by design. A storage-localization score identifies this known interface with AUROC 0.981 and residualized AUROC 0.926, indicating accurate localization. Yet it predicts the better intervention on only 17 of 36 targets, while an alternative LoRA interface (Hu et al., 2022) achieves the better outcome on 35 of 36. Accurate localization does not by itself identify where a particular unlearning objective should intervene. This choice is also objective-dependent. On our Natural-TOFU setting, derived from TOFU (Maini et al., 2024), equal-size subsets selected independently under NPO and SimNPO (Zhang et al., 2024; Fan et al., 2025b) have a Jaccard overlap of about 0.143. Both observations motivate choosing the initial subset according to the expected effect of the actual update instead of localization alone. Running every candidate intervention to completion would be prohibitively expensive, so we introduce *Intervention Score* as an efficient proxy. The score estimates how the update induced by the chosen objective would act on each group. It rewards predicted forgetting, penalizes predicted damage to retained and protected behavior, and discounts nonspecific model changes. Ranking under a fixed parameter budget yields an objective-conditioned initial subset. Keeping it fixed defines STATIC-IV and isolates initial selection quality. Because even a strong initial subset can become less suitable as the model changes, selective DIR-R revisits it at a later checkpoint and runs the complete continuation comparison only when a calibrated probe indicates potential value.

We evaluate the two decisions in complementary settings. On the Natural-TOFU dataset, Intervention Score with DIR-R is compared with ten parameter-selection baselines under NPO and SimNPO; its descriptive mean margins are positive in $19/20$ comparisons, with several close to zero. The LACUNA localization-precision benchmark (Boglioni et al., 2026) provides direct comparisons with selected baselines. After selecting three objective-specific baselines from a ten-method panel, our method has higher mean terminal utility in all six NPO and SimNPO comparisons. NPO margins range from $+0.431$ to $+0.848$, and SimNPO margins range from $+0.503$ to $+0.571$. GradDiff exposes a boundary: static parameter-subset rankings change substantially across data fields, and several static methods remain ahead in the complete four-field primary. In the DIR-R ablation, the aggregate mean difference relative to STATIC-IV is positive in all six evaluation sets; the primary GradDiff evaluation has mean gain $+0.1646$, median paired gain $+0.0025$, and six wins, six ties, and no losses. The results treat initial selection and checkpoint-dependent revision as distinct empirical questions without claiming uniform dominance.

Our contributions are:

- **A controlled separation of localization and intervention.** We show that highly accurate storage localization need not identify the better unlearning intervention: despite storage AUROC 0.981, storage identity agrees with the better intervention on only $17/36$ targets.
- **Objective-conditioned initial parameter selection.** We introduce Intervention Score and its fixed-subset realization STATIC-IV. Against ten parameter-selection methods under NPO and SimNPO on the Natural-TOFU dataset, Intervention Score with DIR-R has positive descriptive margins in $19/20$ comparisons.
- **Checkpoint-dependent support revision.** We introduce selective DIR-R, which conditionally revisits the initial subset rather than recomputing alternatives throughout optimization. The LACUNA benchmark reports positive mean margins in all six NPO and SimNPO baseline comparisons, while the ablation of DIR-R against STATIC-IV has a positive aggregate mean difference in every reported evaluation set.

2 RELATED WORK

Unlearning. Classical machine unlearning studies deletion-efficient learning, certified removal, selective forgetting, and retraining-oriented guarantees (Ginart et al., 2019; Guo et al., 2020; Golatkar et al., 2020; Sekhari et al., 2021). LLM-specific work extends this setting with privacy-oriented dememorization, direct parameter updates, representation-level removal, parameter-efficient methods, causal formulations, and inference-time alternatives (Kassem et al., 2023; Chen and Yang, 2023; Ji et al., 2024; Yao et al., 2024; Pawelczyk et al., 2024; Liu et al., 2024; Cha et al., 2025; Ding et al., 2025; Gao et al., 2025). Evaluation work additionally stresses that apparent forgetting can be fragile under stronger metrics, obfuscation tests, or relearning (Hu et al., 2025; Wang et al., 2025; Fan et al., 2025a). Our work fixes the unlearning objective and studies a different decision: which parameters that objective should update, and whether that choice should later change.

Localization. Model editing uses direct parameter modification, learned editors, retrieval mechanisms, or localized weight updates to change factual behavior (De Cao et al., 2021; Mitchell et al., 2022a;b; Meng et al., 2022; 2023); later work further questions static fact-level localization assumptions (Chen et al., 2025). Localized unlearning inherits the same tension. WAGLE and MemFlex choose update regions from attribution or gradient evidence (Jia et al., 2024; Tian et al., 2024), while mechanistic and benchmark studies test how localization affects robustness and intervention efficacy (Guo et al., 2025; Lee et al., 2025; Boglioni et al., 2026). We treat localization evidence as an input to an intervention decision rather than as the decision target itself.

3 Preliminaries and Problem Formulation

Table 5 in Appendix A collects the notation used below.

3.1 Localized unlearning as support-constrained optimization

Let $f_{\boldsymbol{\theta}_0}$ be the model before unlearning, with forget set $\mathcal{D}_f$ and retain set $\mathcal{D}_r$. We fix an unlearning objective $\mathcal{L}_{\text{unl}}$, its optimizer and minibatch sequence, and a horizon of T optimization steps. Let $\mathbb{G}$ be a finite collection of editable parameter groups. For $g \in \mathbb{G}$, $\boldsymbol{\theta}_g$ denotes the coordinates in that group. A support $\mathbb{S} \subseteq \mathbb{G}$ specifies which groups may change during unlearning; all coordinates outside $\mathbb{S}$ remain frozen. If $\boldsymbol{P}_{\mathbb{S}}$ projects onto the active coordinates, a support-constrained update has the form

$$\boldsymbol{\theta}_{k+1}^{\mathbb{S}} = \boldsymbol{\theta}_k^{\mathbb{S}} - \eta_k \boldsymbol{P}_{\mathbb{S}} \mathbf{u}_k^{\mathbb{S}}, \tag{1}$$

where $\mathbf{u}_k^{\mathbb{S}}$ is the optimizer direction along the trajectory induced by support $\mathbb{S}$. Thus, the support is an intervention variable: it determines where the declared unlearning algorithm is allowed to act.

The editable-parameter cost of a support is

$$c(\mathbb{S}) = \sum_{g \in \mathbb{S}} \dim(\boldsymbol{\theta}_g), \tag{2}$$

and the feasible family under budget B is

$$\mathcal{F}_B = \{\mathbb{S} \subseteq \mathbb{G} : c(\mathbb{S}) \leq B\}. \tag{3}$$

For support-comparison experiments, a terminal model is scored by $J(\boldsymbol{\theta}) = G_F(\boldsymbol{\theta}) - D_{\text{coll}}(\boldsymbol{\theta})$, where G_F measures forgetting gain and $D_{\text{coll}} \geq 0$ measures collateral degradation. Larger J is better. The objective, optimizer, minibatches, budget, and evaluator are held fixed when supports are compared.

3.2 From localization to intervention selection

A localization method asks which parameter groups are associated with the information to be forgotten. That is a descriptive question. Choosing a support asks a different question: if the same unlearning procedure is allowed to update only a particular set of groups, which choice leads to the best terminal outcome? A group may therefore receive a strong localization signal without being the best place to intervene under the actual objective, budget, and optimization trajectory.

For a fixed support $\mathbb{S} \in \mathcal{F}_B$, let $\boldsymbol{\theta}_T^{\mathbb{S}}$ be the endpoint obtained by running the fixed unlearning procedure from $\boldsymbol{\theta}_0$ for all T steps while updating only $\mathbb{S}$. Its *initial intervention value* is

$$V_0(\mathbb{S}) = J(\boldsymbol{\theta}_T^{\mathbb{S}}), \tag{4}$$

and the ideal initial support is

$$\mathbb{S}_0^\star \in \arg\max_{\mathbb{S} \in \mathcal{F}_B} V_0(\mathbb{S}). \tag{5}$$

Computing Eq. 5 directly would require full unlearning runs for many feasible supports. The first problem is therefore to select a practical $\mathbb{S}_0 \in \mathcal{F}_B$ that is informed by the unlearning objective without exhaustively evaluating terminal outcomes.

3.3 CHECKPOINT-CONDITIONED SUPPORT REVISION

The initial support is chosen at $\boldsymbol{\theta}_0$, but optimization changes the model and its optimizer state. After following $\mathbb{S}_0$ to step t, let $\mathcal{C}_t$ denote a restorable checkpoint. Restoring $\mathcal{C}_t$ reinstates $\boldsymbol{\theta}_t$, the optimizer and scheduler state, the position in the minibatch sequence, and the random-number-generator state. These quantities are restored identically for every continuation, so the editable support is the only intervention that changes.

For a continuation support $\mathbb{S} \in \mathcal{F}_B$, let $\boldsymbol{\theta}_T(\mathbb{S}; \mathcal{C}_t)$ be the endpoint obtained by restoring $\mathcal{C}_t$ and running the remaining steps with updates restricted to $\mathbb{S}$. Its *continuation value* is

$$V_t(\mathbb{S} \mid \mathcal{C}_t) = J(\boldsymbol{\theta}_T(\mathbb{S}; \mathcal{C}_t)), \tag{6}$$

with ideal checkpoint-dependent continuation

$$\mathbb{S}_t^\star(\mathcal{C}_t) \in \arg\max_{\mathbb{S} \in \mathcal{F}_B} V_t(\mathbb{S} \mid \mathcal{C}_t). \tag{7}$$

There is no requirement that $\mathbb{S}_t^\star(\mathcal{C}_t)$ coincide with the support preferred at initialization. The second problem is therefore computational: after observing $\mathcal{C}_t$, determine whether the current support is still adequate or whether a comparison of alternative continuations is worth its cost. Our method addresses these two decisions: initial intervention selection and checkpoint-conditioned revision.

4 METHODOLOGY

4.1 METHOD OVERVIEW

The method separates support selection from support revision. At initialization, *Intervention Score* ranks parameter groups by the predicted effect of the actual unlearning update and produces support $\mathbb{S}_0$. STATIC-IV uses this support for the entire trajectory. Selective DIR-R adds a decision at checkpoint $\mathcal{C}_t$: a short comparison determines whether the frozen continuation candidates should be evaluated to completion. Algorithm 1 gives the full procedure.

This construction is narrower than solving Eqs. 5 and 7 globally. Intervention Score is a local surrogate for terminal intervention value, and DIR-R searches a small, prespecified family around the initial support. The method makes both decisions testable under a fixed training procedure; it does not claim global support optimality.

4.2 INTERVENTION SCORE: OBJECTIVE-CONDITIONED INITIAL PARAMETER SELECTION

A localization score is typically derived from how the forget data are represented or attributed in the current model. Intervention Score instead asks how a candidate group would participate in the update induced by $\mathcal{L}_{\text{unl}}$. Let $\mathbf{u}_0$ be the preconditioner-free direction induced by the unlearning objective at $\boldsymbol{\theta}_0$, defined so that a small objective step is $\boldsymbol{\theta}_0 - \eta \mathbf{u}_0$. Let $\boldsymbol{P}_g$ project onto group g.

We use four outcome-blind diagnostic roles $q \in \{F, R, P, N\}$: forget, retain, protected behavior, and neutral behavior. The forget diagnostic measures whether the restricted update moves in a direction favorable to forgetting. The retain and protected diagnostics flag predicted collateral degradation, with the protected role covering behavior that is not represented by the ordinary retain set. The neutral diagnostic measures whether the same update produces broad, nonspecific movement. These diagnostics are used only to construct the parameter-selection rule and do not read the terminal outcomes used for evaluation. Let ℓ_q be a diagnostic loss and $\mathbf{g}_q = \nabla_{\boldsymbol{\theta}} \ell_q(\boldsymbol{\theta}_0)$.

The score follows from the first-order change produced by a restricted step. For small η,

$$\ell_q(\boldsymbol{\theta}_0 - \eta \boldsymbol{P}_g \mathbf{u}_0) = \ell_q(\boldsymbol{\theta}_0) - \eta \langle \boldsymbol{P}_g \mathbf{g}_q, \boldsymbol{P}_g \mathbf{u}_0 \rangle + O(\eta^2). \tag{8}$$

We therefore define the signed first-order effect

$$e_{g,q} = -\langle \boldsymbol{P}_g \mathbf{g}_q, \boldsymbol{P}_g \mathbf{u}_0 \rangle. \tag{9}$$

The diagnostics are oriented so that $e_{g,F} > 0$ predicts useful forgetting, whereas $e_{g,R} > 0$ and $e_{g,P} > 0$ predict collateral degradation; $|e_{g,N}|$ measures nonspecific movement. Intervention Score is

$$s_{\text{int}}(g) = \frac{e_{g,F} - \max\{e_{g,R}, e_{g,P}, 0\}}{|e_{g,N}| + \epsilon}, \tag{10}$$

where $\epsilon > 0$ is a fixed numerical stabilizer. The numerator rewards predicted forgetting while subtracting the larger predicted collateral effect. The denominator downweights groups whose predicted effect is broad rather than forget-directed. This score is used as a ranking surrogate for V_0; it is not assumed to equal the terminal value of a support.

Groups are ranked by s_{int} and added while respecting $c(\mathbb{S}) \leq B$, producing $\mathbb{S}_0$. In Natural-TOFU and LACUNA, an atomic group is one input column of an MLP down-projection matrix. Groups within a model therefore have equal parameter count, so ranking and packing reduce to taking the highest-scoring groups up to the budget. The candidate geometry and editable-parameter budget are the same for all parameter-selection methods in a comparison.

4.3 STATIC-IV: FIXING THE INITIAL SUPPORT

STATIC-IV keeps $\mathbb{S}_0$ active for all T steps. At step k, the objective and optimizer produce the support-conditioned direction $\mathbf{u}_k^{\mathbb{S}_0}$ and step size η_k, giving

$$\boldsymbol{\theta}_{k+1}^{\mathbb{S}_0} = \boldsymbol{\theta}_k^{\mathbb{S}_0} - \eta_k \boldsymbol{P}_{\mathbb{S}_0} \mathbf{u}_k^{\mathbb{S}_0}. \tag{11}$$

This variant answers a specific ablation question: what is obtained from the objective-conditioned initial parameter selection if the support is never reconsidered? It also supplies the current STATIC-IV continuation against which DIR-R evaluates possible revisions.

4.4 SELECTIVE DIR-R: REVISING THE SUPPORT WHEN USEFUL

The score that produced $\mathbb{S}_0$ was computed at initialization. By step t, the parameters, optimizer state, and remaining horizon have changed, so the initial ranking need not remain best for the continuation problem in Eq. 7. Resolving that problem at every checkpoint would defeat the purpose of localized unlearning. DIR-R therefore separates whether to run a continuation comparison from which support to choose after a trigger.

Local candidate family. Order the groups in $\mathbb{S}_0$ from lowest to highest Intervention Score as $g_1^-, \ldots, g_M^-$, and unselected groups from highest to lowest as $g_1^+, g_2^+, \ldots$. For replacement fraction ρ, let $k(\rho) = \lfloor \rho M \rfloor$ and define

$$\mathbb{S}^{(\rho)} = \Big(\mathbb{S}_0 \setminus \{g_1^-, \ldots, g_{k(\rho)}^-\} \Big) \cup \{g_1^+, \ldots, g_{k(\rho)}^+\}. \tag{12}$$

We freeze $\mathbb{A} = \{0, 0.01, 0.025, 0.05, 0.10\}$, where $\rho = 0$ is the unchanged support. In the primary equal-cost geometry, every exchange preserves the editable-parameter count exactly. With heterogeneous group costs, the implementation must instead enforce $c(\mathbb{S}^{(\rho)}) = c(\mathbb{S}_0)$ explicitly.

Short probe and gate. Running every candidate to step T is expensive, so the gate first compares the current support with one fixed probe action $\rho_p \in \mathbb{A} \setminus \{0\}$. For support $\mathbb{S}$, let $\boldsymbol{\theta}_{t+d}(\mathbb{S}; \mathcal{C}_t)$ denote the state reached after exactly d additional steps from the same checkpoint, using the same minibatches and randomness. Define

$$\widetilde{V}_t^{(d)}(\rho \mid \mathcal{C}_t) = J\Big(\boldsymbol{\theta}_{t+d}(\mathbb{S}^{(\rho)}; \mathcal{C}_t) \Big), \tag{13}$$

and the nonnegative probe signal

$$s_t = \max\{0, \widetilde{V}_t^{(d)}(\rho_p \mid \mathcal{C}_t) - \widetilde{V}_t^{(d)}(0 \mid \mathcal{C}_t)\}. \tag{14}$$

The probe does not choose the final support. It only asks whether the observed checkpoint provides enough evidence to justify a more expensive comparison. If $s_t \leq \tau^\star$, the run continues with $\mathbb{S}^{(0)}$ and no full continuation comparison is performed.

Exhaustive continuation comparison after a trigger. If $s_t > \tau^\star$, every candidate is restored from the same $\mathcal{C}_t$ and run to T under the same future minibatches and randomness. DIR-R chooses

$$\widehat{\rho}_t = \arg\max_{\rho \in \mathbb{A}} V_t(\mathbb{S}^{(\rho)} \mid \mathcal{C}_t), \tag{15}$$

with ties broken toward smaller ρ. Including $\rho = 0$ means the exhaustive comparison can retain the STATIC-IV continuation whenever every proposed exchange is worse. This guarantee is only relative to the frozen candidate family; it is not a claim of global optimality over $\mathcal{F}_B$.

Algorithm 1 Objective-conditioned support selection with selective DIR-R

Require: initial model $\boldsymbol{\theta}_0$, forget and retain data, objective $\mathcal{L}_{\mathrm{unl}}$, diagnostics $\{\ell_q\}_{q\in\{F,R,P,N\}}$, groups $\mathbb{G}$, budget B, horizon T
Require: checkpoint t, probe length d, actions $\mathbb{A}$, probe action ρ_p, calibrated threshold $\tau^\star$

Compute $\mathbf{u}_0$ and $\mathbf{g}_q = \nabla_{\boldsymbol{\theta}} \ell_q(\boldsymbol{\theta}_0)$
for each $g \in \mathbb{G}$ **do**
  $e_{g,q} \leftarrow -\langle \boldsymbol{P}_g \mathbf{g}_q, \boldsymbol{P}_g \mathbf{u}_0 \rangle$ for $q \in \{F, R, P, N\}$
  $s_{\mathrm{int}}(g) \leftarrow \big(e_{g,F} - \max\{e_{g,R}, e_{g,P}, 0\}\big)/(|e_{g,N}| + \epsilon)$
end for
Select $\mathbb{S}_0$ under B; construct $\{\mathbb{S}^{(\rho)} : \rho \in \mathbb{A}\}$ with $\mathbb{S}^{(0)} = \mathbb{S}_0$
Run with $\mathbb{S}_0$ to t; save $\mathcal{C}_t$; probe $\rho = 0$ and ρ_p from independent restorations
$s_t \leftarrow \max\{0, \widetilde{V}_t^{(d)}(\rho_p \mid \mathcal{C}_t) - \widetilde{V}_t^{(d)}(0 \mid \mathcal{C}_t)\}$
if $s_t \le \tau^\star$ **then**
  return $\boldsymbol{\theta}_T(\mathbb{S}^{(0)}; \mathcal{C}_t)$
else
  Evaluate every $\rho \in \mathbb{A}$ from independent restorations; set $\widehat{\rho}_t \leftarrow \arg\max_{\rho\in\mathbb{A}} V_t(\mathbb{S}^{(\rho)} \mid \mathcal{C}_t)$
  return $\boldsymbol{\theta}_T(\mathbb{S}^{(\widehat{\rho}_t)}; \mathcal{C}_t)$
end if

Threshold calibration and compute reference. The threshold $\tau^\star$ is selected using development examples only. At each development checkpoint, the probe signal is paired with the positive available gain inside the frozen candidate family. Among thresholds in a prespecified trigger-rate range, calibration favors the threshold that captures the most positive available gain per unit of additional step-equivalent compute. Final outcomes do not alter $\tau^\star$, ρ_p, d, $\mathbb{A}$, the budget, or the unlearning objective. Natural-TOFU uses $T = 200$, $t = 160$, $d = 20$, and $\rho_p = 0.01$; the calibration equations are given in Appendix A.

For context, the $T = 200$ exhaustive multi-checkpoint reference evaluates all five actions from checkpoints $\{0, 40, 80, 120, 160\}$. It requires $5(200 + 160 + 120 + 80 + 40) = 3000$ continuation steps in addition to the committed 200-step trajectory, or $16\times$ normalized step-equivalent compute. This exhaustive construction is used only as a compute-heavy reference and is distinct from the single-checkpoint exhaustive comparison invoked by deployed DIR-R.

5 Experiments

Overview. We test localization versus intervention in a controlled setting, breadth on Natural-TOFU, comparisons and heterogeneity on LACUNA, and checkpoint-dependent revision in the final ablation.

Metrics. The terminal utility is $J = G_F - D_{\mathrm{coll}}$, where larger values mean more forgetting gain after benchmark-specific nonnegative collateral damage. For paired runs, $\Delta J(A, B) = n^{-1} \sum_i [J_i(A) - J_i(B)]$, so positive values favor A. We also report paired medians, wins/ties/losses (W/T/L), and leave-one-out (LOO) ranges for GradDiff. DIR-R metrics are defined in Appendix B.2.

5.1 Experimental Protocol

Setup. Our Natural-TOFU dataset, derived from TOFU (Maini et al., 2024), uses Llama 3 8B (Grattafiori et al., 2024) and Gemma 2 2B (Gemma Team et al., 2024) under NPO and SimNPO (Zhang et al., 2024; Fan et al., 2025b), with approximately six million editable multilayer-perceptron (MLP) down-projection scalars. The LACUNA localization-precision benchmark (Boglioni et al., 2026) uses frozen OLMo 3 7B (Team Olmo et al., 2025) on Email Address, Driver's License, Birth City, and Phone Number under NPO, SimNPO, and gradient difference (GradDiff) (Yao et al., 2024). We compare weight-attribution-guided LLM unlearning (WAGLE) (Jia et al., 2024), MemFlex (Tian et al., 2024), Fisher-initialized low-rank adaptation (FILA) (Cha et al., 2025), activation-patching MLP (AP-MLP) and Activation-Down (Lee et al., 2025), gradient-based adaptive unlearning (GRAIL) (Kim et al., 2025), and Projected Causal (Liu et al., 2024; Guo et al., 2025). Random, Activation, and Storage are outcome-blind controls defined in this study. All methods use the same editable geometry when available.

Protocol. Natural-TOFU combines a completed parameter-selection panel with later DIR-R evaluation runs, so its reported margins are descriptive summaries rather than direct paired estimates.

For LACUNA, baseline identities are fixed before outcome evaluation and compared methods use the same field, objective, seed, parameter budget, and evaluator. GradDiff's four-field aggregate is primary; the three-field exclusion of Birth City is a sensitivity analysis defined after observing the field effect. Missing runs are not imputed, and DIR-R settings are frozen before final outcomes are read; Appendix C gives the audit.

5.2 Controlled Separation Between Localization And Intervention Selection

Localization. Synthetic facts are injected through known parameter interfaces in Qwen2.5-7B (Qwen Team et al., 2025). Without using intervention outcomes, the locator reaches AUROC 0.98148, residualized AUROC 0.92593, and the correct direction on $17/18$ blind pairs.

Table 1: Controlled separation between localization and intervention.

Quantity	Result
Storage AUROC	0.98148
Residualized storage AUROC	0.92593
Correct within-pair storage direction	$17/18$
LoRA intervention wins	$35/36$
Storage/best-intervention agreement	$17/36$ (47.2%)
LoRA−backbone paired difference	$+0.1905$ $[0.1570, 0.2242]$
Storage-score / paired-difference corr.	-0.2302

Intervention. Across 36 NPO targets, LoRA wins $35/36$, while storage identity agrees with the better intervention on only $17/36$ (47.2%). The mean effect is $+0.1905$ (95% interval $[0.1570, 0.2242]$), and localization score correlates negatively with the LoRA-minus-backbone paired difference (-0.2302). ***Accurate localization does not reliably identify the better intervention in this controlled setting.***

5.3 Natural-TOFU: Broad Method Comparison

Scope. The Natural-TOFU dataset compares Intervention Score with DIR-R against ten parameter-selection methods under NPO and SimNPO. Because the parameter-selection and DIR-R components come from different final evaluation sets, Table 2 is descriptive rather than a paired state-of-the-art estimate.

Table 2: Natural-TOFU margins (ΔJ); positive values favor ours.

Baseline	NPO ΔJ	SimNPO ΔJ
Random	$+0.004799$	$+0.006974$
Activation	$+0.041006$	$+0.154743$
WAGLE	$+0.136450$	$+0.297768$
Storage	$+0.032802$	$+0.162268$
MemFlex	$+0.094780$	$+0.210191$
FILA-relative Fisher	$+0.003390$	-0.000799
AP-MLP	$+0.029772$	$+0.098141$
GRAIL	$+0.002172$	$+0.000719$
Activation-Down	$+0.039595$	$+0.149278$
Projected Causal	$+0.110948$	$+0.131786$

Results. ***Margins are positive in*** $19/20$ ***comparisons, but several are close to zero.*** The largest is $+0.297768$ against WAGLE under SimNPO; the only negative is the -0.000799 near-tie with FILA-relative Fisher.

5.4 LACUNA: Comparisons And GradDiff Heterogeneity

Comparisons. Table 3 reports the six NPO/SimNPO comparisons and the GradDiff sensitivity.

Results. ***Our method has higher mean terminal utility and more run-level wins than losses in all six direct comparisons.*** NPO margins range from $+0.431$ to $+0.848$, and SimNPO margins range from $+0.503$ to $+0.571$.

Three-Field Sensitivity. In the three-field sensitivity excluding Birth City, DIR-R has the highest mean terminal utility, but its $+0.014444$ margin over Projected Causal is unstable: median -0.94, W/T/L $4/0/5$, and LOO $[-1.6425, +1.0056]$.

Birth City. The four-field primary remains in Appendix C.3. Projected Causal has a mean terminal utility 7.104583 higher than DIR-R, and all eight external parameter-selection methods have higher

Table 3: LACUNA results. Panel A gives NPO/SimNPO comparisons. In Panel B, DIR-R is the reference with mean $J = 106.591877$; all remaining columns report paired DIR-R-minus-baseline statistics, so no self-comparison row is shown for DIR-R. The four-field primary is in Appendix C.3.

Panel A: NPO/SimNPO comparison

Objective / baseline	Ours	Baseline	ΔJ	Wins/Losses
NPO / Projected Causal	6.193908	5.763156	+0.430752	8/4
NPO / MemFlex	6.193908	5.502749	+0.691159	10/2
NPO / Activation exact	6.193908	5.346050	+0.847858	9/3
SimNPO / MemFlex	1.776698	1.271154	+0.505544	11/1
SimNPO / FILA	1.776698	1.205905	+0.570793	8/4
SimNPO / Activation-Down	1.776698	1.273278	+0.503420	12/0

Panel B: GradDiff three-field sensitivity, Birth City excluded

Baseline	Mean J	DIR-R−baseline	Median gap	W/T/L	LOO range
Projected Causal	106.577433	+0.014444	−0.940000	4/0/5	[−1.6425, +1.0056]
Activation-Down	104.534099	+2.057778	+0.030000	5/0/4	[+0.5869, +2.8506]
Storage	103.033544	+3.558333	+2.150000	7/0/2	[+2.2525, +4.7450]
FILA	102.956044	+3.635833	+1.430000	7/0/2	[+1.9897, +4.2047]
GRAIL	102.482155	+4.109722	+3.150000	8/0/1	[+2.3316, +4.6578]
MemFlex	102.189655	+4.402222	+1.795000	6/0/3	[+3.2206, +5.4519]
WAGLE	101.389099	+5.202778	+4.070000	7/0/2	[+4.1113, +6.2088]
AP-MLP	99.490766	+7.101111	+5.315000	9/0/0	[+5.5400, +7.8356]

Table 4: Selective revision relative to STATIC-IV. Available gain is the improvement from exhaustive candidate evaluation at the frozen decision checkpoint; it is distinct from the 16× exhaustive multi-checkpoint reference.

Benchmark	Objective / setting	n	ΔJ_{DIRR}	Trigger	Avail. gain	Captured
Natural-TOFU	NPO	8	+0.002010	50.0%	+0.005667	35.5%
Natural-TOFU	SimNPO	8	+0.005130	25.0%	+0.013162	39.0%
Natural-TOFU‡	GradDiff	20	+0.004298	40.0%	+0.014793	29.1%
LACUNA	NPO	12	+0.002433	33.3%	+0.008024	30.3%
LACUNA	SimNPO	12	+0.008228	58.3%	+0.014868	55.3%
LACUNA	GradDiff, four-field	12	+0.164583	50.0%	+0.282500	58.3%

‡Natural GradDiff uses held-out trace replay. The GradDiff candidate-set denominator was completed post hoc without changing the frozen DIR-R decisions or threshold.

Birth City means, with gaps from -28.461667 to -5.872500 (Appendix Tables 8 and 9). ***Most gaps outside Birth City reverse sign, so this field drives the aggregate reversal.***

Replication. A three-field rerun on new seeds reverses the near-tie: Projected Causal leads DIR-R by 0.861389 with the same $4/0/5$ W/T/L (Appendix Table 12). We therefore treat the external ordering as seed-sensitive rather than as a stable three-field ranking.

5.5 ABLATION: SELECTIVE REVISION

Question. The DIR-R ablation asks whether the gate can preserve the fixed-support solution while using substantially less step-equivalent compute than the $16\times$ exhaustive multi-checkpoint reference. STATIC-IV is the $1\times$ fixed-support baseline; DIR-R changes only whether support is reconsidered at the checkpoint. Table 4 collects the main Natural-TOFU and LACUNA DIR-R evaluation sets.

Safety. The $\rho = 0$ candidate is exactly the STATIC-IV continuation. Non-triggered runs retain it, and triggered exhaustive selection includes it among the candidates; when all branches complete, terminal utility cannot decrease relative to STATIC-IV. ***The primary four-field GradDiff evaluation has mean gain*** $+0.164583$***, median paired gain*** $+0.002500$***, and six wins, six ties, and no losses.*** The difference between the mean and median shows that gain magnitude is heterogeneous across runs.

Efficiency. The main question is how much useful revision is retained for far less computation, not whether every absolute gain is large. Under the GradDiff cost model, non-trigger and trigger paths cost $1.10\times$ and $1.80\times$; with 6/12 triggers, the primary four-field evaluation averages $1.45\times$, below one tenth of the $16\times$ exhaustive multi-checkpoint reference in normalized step-equivalent compute. Across the six evaluation sets with complete candidate-set denominators, DIR-R captures from 29.1% to 58.3% of the gain available from exhaustive candidate evaluation at the frozen checkpoint. The primary four-field evaluation captures 58.3% ($+0.164583$ of $+0.282500$). Its denominator was completed only after the DIR-R decisions were frozen and is used for diagnosis, not retuning. Appendix C.4 retains the three-field replication as a boundary analysis rather than a headline result.

6 Discussion and Limitations

Our results distinguish parameter localization from the downstream decision of where an unlearning objective should act. The controlled experiment makes this distinction explicit: the injected interface can be identified accurately, yet this information does not predict the better intervention. The parameter-selection comparisons further show that intervention quality depends on the objective and data field. These results do not imply that localization is uninformative; rather, they indicate that localization evidence alone is not a sufficient decision criterion for selecting an editable subset.

The evidence also has several important boundaries. Natural-TOFU margins for Intervention Score + DIR-R combine results from different final evaluation sets and therefore are not direct paired estimates. Natural GradDiff uses held-out trace replay rather than the confirmation evaluation set. For LACUNA GradDiff, the four-field result is the primary endpoint: several static parameter-selection methods remain ahead of DIR-R, and the three-field analysis excluding Birth City is a sensitivity analysis. We therefore interpret it as evidence of field heterogeneity rather than as a replacement superiority result.

Finally, DIR-R searches only within a frozen family of equal-budget continuation candidates. Its exhaustive comparisons identify the best continuation within that family, not the globally optimal editable subset. The gate trades additional computation for continuation gain: the $16\times$ exhaustive multi-checkpoint construction is a normalized step-equivalent reference rather than a wall-clock measurement, and the appropriate balance between compute and gain may vary across objectives and deployment settings.

7 Conclusion

Localized LLM unlearning involves two decisions that are often conflated: which parameters should be updated initially, and whether that choice should remain fixed as optimization changes the model. We formalized these decisions separately. Intervention Score supplies an objective-conditioned initial support, while selective DIR-R tests whether a checkpoint justifies revising that support.

The evaluations clarify both the value and the limits of this decomposition. The controlled study shows that accurate storage localization can fail to identify the better intervention. Natural-TOFU provides broad descriptive evidence across ten parameter-selection baselines, while the direct LACUNA comparisons show positive mean margins under NPO and SimNPO. The GradDiff results add an important qualification: external static rankings vary by data field and seed, and several static methods remain ahead in the four-field primary. Against its own STATIC-IV reference, however, DIR-R has a positive aggregate mean difference in every reported evaluation set while invoking the complete continuation comparison only after a trigger. Localized unlearning should therefore treat support selection as a decision about the effect of an update, report where that decision is unstable, and evaluate checkpoint-dependent revision separately from initial localization quality. This framing permits stronger claims where comparisons are direct without hiding boundary cases where method orderings change. Future work should test learned candidate families and multi-checkpoint gates under stronger forgetting attacks and measured wall-clock constraints. The central question is not only where target information is found, but whether updating those parameters improves forgetting without unacceptable collateral damage at the current model state. This makes support control a distinct design and evaluation problem.

AI USE STATEMENT

Generative AI tools were used for literature search and related-work discovery, feedback on experimental design and methodology, experimental implementation including code generation, analysis and interpretation of experimental results, and drafting and polishing parts of the manuscript. All AI-assisted content, code, references, and experimental results were reviewed and verified by the authors, who take responsibility for the final paper.

REPRODUCIBILITY STATEMENT

Section 4 and Algorithm 1 specify Intervention Score, the fixed-subset STATIC-IV reference, and the calibration and deployment procedures for selective DIR-R. Appendix A documents support-constrained optimization, checkpoint restoration, candidate construction, and threshold calibration. Appendix B records the comparison protocols, evaluation metrics, aggregation and missingness rules, and the controlled separation between storage and intervention. Appendix C reports the complete audited result tables, including the GradDiff DIR-R analysis, candidate-set coverage, and replication analyses. An anonymized supplementary code package is organized by the table it supports.

REFERENCES

Matteo Boglioni, Thibault Rousset, Siva Reddy, Marius Mosbach, and Verna Dankers. Lacuna: A testbed for evaluating localization precision for llm unlearning. *arXiv preprint arXiv:2607.02513*, 2026.

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*, pages 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480. IEEE, 2015. doi: 10.1109/SP.2015.35.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. Towards robust and parameter-efficient knowledge unlearning for llms. In *International Conference on Learning Representations*, 2025.

Hang Chen, Jiaying Zhu, Hongyang Chen, Hongxu Liu, Xinyu Yang, and Wenya Wang. Navigating by old maps: The pitfalls of static mechanistic localization in llm post-training. *arXiv preprint arXiv:2605.06076*, 2026.

Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.738.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Knowledge localization: Mission not accomplished? enter query localization! In *International Conference on Learning Representations*, 2025.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, 2021. doi: 10.18653/v1/2021.emnlp-main.522.

Chenlu Ding, Jiancan Wu, Yancheng Yuan, Jinda Lu, Kai Zhang, Alex Su, Xiang Wang, and Xiangnan He. Unified parameter-efficient unlearning for llms. In *International Conference on Learning Representations*, 2025.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 15762–15778. PMLR, 2025a.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning. In *International Conference on Learning Representations*, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y. Zou. Making ai forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3832–3842, 2020.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 20964–20992. PMLR, 2025.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *International Conference on Learning Representations*, 2025.

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)*, pages 14389–14408. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.805.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Wagle: Strategic weight attribution for effective and modular unlearning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Aly Kassem, Omar Mahmoud, and Sherif Saad. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.265.

Kun-Woo Kim, Ji-Hoon Park, Ju-Min Han, and Seong-Whan Lee. Grail: Gradient-based adaptive unlearning for privacy and copyright in llms. In *International Joint Conference on Neural Networks*, 2025. arXiv:2504.12681.

Hwiyeong Lee, Uiji Hwang, Hyelim Lim, and Taeuk Kim. Does localization inform unlearning? a rigorous examination of local parameter attribution for knowledge unlearning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21857–21869. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1109.

Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8708–8731, 2024. doi: 10.18653/v1/2024.emnlp-main.495.

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

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

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, pages 15817–15831, 2022b.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 40034–40050. PMLR, 2024.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. To forget or not? towards practical knowledge unlearning for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1524–1537. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.82. URL https://aclanthology.org/2024.findings-emnlp.82/.

Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. Towards effective evaluations and comparisons for llm unlearning methods. In *International Conference on Learning Representations*, 2025.

Abudukelimu Wuerkaixi, Qizhou Wang, Sen Cui, Wutong Xu, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Adaptive localization of knowledge negation for continual llm unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 68094–68117. PMLR, 2025.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, 2024. doi: 10.18653/v1/2024.acl-long.457.

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

A Methodological Details

A.1 Core Notation

Notation. Table 5 summarizes symbols already defined in Sections 3 and 4.

Table 5: Core notation used in the problem formulation and method.

Symbol	Meaning	Symbol	Meaning
$f_{\boldsymbol{\theta}}, \boldsymbol{\theta}_0$	Model and pre-unlearning parameters.	T, t	Total steps and DIR-R decision checkpoint.
$\mathcal{D}_f, \mathcal{D}_r$	Forget and retain sets.	$\mathcal{L}_{\text{unl}}$	Fixed unlearning objective.
$\mathbb{G}, g$	Editable groups and one group.	$\boldsymbol{\theta}_g$	Parameter coordinates in group g.
$\mathbb{S}$	Active editable support.	$c(\mathbb{S}), B$	Editable-parameter cost and budget.
$\mathcal{F}_B$	Supports satisfying $c(\mathbb{S}) \le B$.	$J(\boldsymbol{\theta})$	Forgetting and retention evaluation utility.
$V_0(\mathbb{S})$	Value of using $\mathbb{S}$ from the start.	$\mathbb{S}_0^\star, \mathbb{S}_0$	Ideal and actually selected initial supports.
$\mathcal{C}_t$	Restorable training checkpoint saved at step t.	$\boldsymbol{\theta}_T(\mathbb{S}; \mathcal{C}_t)$	Terminal parameters after continuing from $\mathcal{C}_t$ with support $\mathbb{S}$.
$\boldsymbol{\theta}_{t+d}(\mathbb{S}; \mathcal{C}_t)$	Parameters after d continuation steps from $\mathcal{C}_t$.	$V_t(\mathbb{S} \mid \mathcal{C}_t)$	Terminal continuation value from $\mathcal{C}_t$.
$\mathbb{S}_t^\star(\mathcal{C}_t)$	Ideal support after observing $\mathcal{C}_t$.	$\boldsymbol{P}_{\mathbb{S}}$	Projector onto support coordinates.
$\mathbf{u}_k^{\mathbb{S}}, \eta_k$	Support-conditioned optimizer direction and step size.	$e_{g,q}$	First-order diagnostic effect, $q \in \{F, R, P, N\}$.
$\rho, \mathbb{A}$	Replacement fraction and frozen action set.	ρ_p, d	Probe action and probe length.
$\widetilde{V}_t^{(d)}(\rho \mid \mathcal{C}_t)$	Short-horizon probe score.	$s_t, \tau^\star$	Gate signal and frozen threshold.

A.2 Intervention Score Details

Intervention Score uses gradients only at the initial model. The objective direction $\mathbf{u}_0$ is computed before optimizer preconditioning so that the score reflects the declared objective rather than optimizer-specific moment estimates. Each diagnostic gradient $\mathbf{g}_q$ is evaluated on its own frozen diagnostic records at the same $\boldsymbol{\theta}_0$. No terminal model score, test outcome, or later checkpoint is used to construct s_{int}.

Equation 8 gives the local interpretation of $e_{g,q}$. The four roles separate the intended effect from two forms of collateral damage and from nonspecific movement: F measures the forget-directed effect, R ordinary retention, P protected behavior outside the ordinary retain diagnostic, and N a neutral control. The score in Eq. 10 is used only for ordering groups. Once $\mathbb{S}_0$ is selected, the actual training update is the original unlearning objective restricted to the active coordinates; Intervention Score is not added to the training loss.

For the equal-cost column geometry used in the primary experiments, selecting $\mathbb{S}_0$ amounts to sorting groups by s_{int} and taking the largest prefix whose editable-parameter count does not exceed B. If group costs differ, the same definition of $c(\mathbb{S})$ applies, but the packing routine must enforce the budget explicitly rather than assuming a fixed number of groups.

A.3 Support-Constrained Optimization

Updates. For support $\mathbb{S}$, $\boldsymbol{P}_{\mathbb{S}}$ projects onto the coordinates in its groups. If the fixed objective and optimizer produce direction $\mathbf{u}_k^{\mathbb{S}}$ and step size η_k at step k, support-constrained optimization applies

$$\boldsymbol{\theta}_{k+1}^{\mathbb{S}} = \boldsymbol{\theta}_k^{\mathbb{S}} - \eta_k \boldsymbol{P}_{\mathbb{S}} \mathbf{u}_k^{\mathbb{S}}. \tag{16}$$

All coordinates outside $\mathbb{S}$ are held fixed. Restoring checkpoint $\mathcal{C}_t$ also reinstates the optimizer and scheduler state, the remaining minibatch position, and the random-number-generator state, so continuation candidates differ only in editable support.

A.4 DIR-R Candidate Construction And Calibration

Candidates. Let $M = |\mathbb{S}_0|$. DIR-R orders selected groups from lowest to highest Intervention Score and unselected groups in the reverse order. For $k(\rho) = \lfloor \rho M \rfloor$, it replaces the first $k(\rho)$ selected groups by the first $k(\rho)$ replacement groups. In the primary equal-cost column geometry this preserves the editable-parameter count exactly. The frozen family is $\mathbb{A} = \{0, 0.01, 0.025, 0.05, 0.10\}$ and $\rho = 0$ is the unchanged STATIC-IV continuation.

Compute. For the $T = 200$ exhaustive multi-checkpoint reference, all five candidates are continued from $t \in \{0, 40, 80, 120, 160\}$. Counting all partial rollouts gives $5(200+160+120+80+40) = 3000$ branch steps; with the committed 200-step trajectory, normalized compute is $16\times$. This is algorithmic step-equivalent accounting, not measured wall-clock time.

Calibration. For development checkpoint $\mathcal{C}_{i,t}$, define the best candidate continuation score and nonnegative available gain

$$V_{i,t}^{\text{best}} = \max_{\rho \in \mathbb{A}} V_{i,t}(\mathbb{S}^{(\rho)} \mid \mathcal{C}_{i,t}), \tag{17}$$

$$\gamma_i = \max\left\{0, V_{i,t}^{\text{best}} - V_{i,t}(\mathbb{S}^{(0)} \mid \mathcal{C}_{i,t})\right\}. \tag{18}$$

For threshold τ, let $\mathcal{I}_\tau = \{i : s_i > \tau\}$. Among thresholds whose trigger fraction lies in a prespecified admissible interval, $\tau^\star$ maximizes $\sum_{i \in \mathcal{I}_\tau} \gamma_i$ divided by the corresponding additional step-equivalent continuation cost. Natural-TOFU uses an admissible trigger range from 20% to 80%. Final outcomes never alter this threshold or the action family.

The complete procedure now appears in Algorithm 1 in the main paper. Its probe branches and triggered continuation candidates are restored independently from the same checkpoint, preventing optimizer or random-number-generator state from leaking across candidates.

B Experimental Details

B.1 Comparison Protocols

Natural-TOFU. The Natural-TOFU dataset uses Llama 3 8B and Gemma 2 2B under NPO and SimNPO for the method comparison. All parameter-selection methods act in the same MLP down-projection column space and use approximately six million trainable scalar parameters in the primary setting. Because the parameter-selection panel and DIR-R evaluation set are distinct, the reported Natural-TOFU margins combine separately estimated components and are not interpreted as direct paired estimates.

LACUNA. LACUNA uses the frozen OLMo 3 7B artifact, four personally identifiable information (PII) fields, and NPO, SimNPO, and GradDiff. NPO and SimNPO baseline identities are fixed before the outcomes are opened and are evaluated on the same field, objective, seed, parameter budget, and evaluator as DIR-R. The GradDiff closure uses $-g_{\text{forget}} + g_{\text{retain}}$, $T = 200$, a $456{,}862{,}557$-parameter budget, final seeds $4099/8191/16381$, calibration seeds $73/311/997$, decision checkpoint 160, and a 20-step probe. STATIC-IV, DIR-R, and the external parameter-selection methods are compared from completed outcomes without retraining or cross-run substitution.

B.2 Metric Definitions

Terminal Score. For evaluation run i, the evaluator returns forgetting gain $G_{F,i}$ and benchmark-specific nonnegative collateral damage $D_{\text{coll},i}$, with

$$J_i = G_{F,i} - D_{\text{coll},i}. \tag{19}$$

Larger J indicates a better balance between forgetting and utility. For methods A and B, $\Delta J(A, B) = n^{-1} \sum_i [J_i(A) - J_i(B)]$; positive values favor A. Median gap and W/T/L use the same paired differences. The LOO range is the minimum and maximum mean gap after deleting one run at a time; a range entirely above zero is less dependent on any single run.

DIR-R Metrics. Trigger rate is $n^{-1}\sum_i \mathbf{1}[s_i > \tau^\star]$ and measures how often exhaustive continuation evaluation is invoked; lower is cheaper but is not intrinsically better. If every final run has outcomes for all $\rho \in \mathbb{A}$, the available candidate-set gain is

$$G_{\text{avail}} = \frac{1}{n}\sum_i \left[\max_{\rho\in\mathbb{A}} V_{i,t}(\mathbb{S}^{(\rho)} \mid \mathcal{C}_{i,t}) - V_{i,t}(\mathbb{S}^{(0)} \mid \mathcal{C}_{i,t})\right]. \tag{20}$$

The fraction of available gain captured is

$$C_{\text{gain}} = \frac{\sum_i [J_i(\text{DIRR}) - J_i(\text{StaticIV})]}{\sum_i \left[\max_{\rho\in\mathbb{A}} V_{i,t}(\mathbb{S}^{(\rho)} \mid \mathcal{C}_{i,t}) - V_{i,t}(\mathbb{S}^{(0)} \mid \mathcal{C}_{i,t})\right]}. \tag{21}$$

Larger C_{gain} means that selective revision captures more opportunity within the frozen candidate family. It is undefined when non-trigger final runs lack exhaustive candidate outcomes.

B.3 Aggregation And Missingness

Missingness. Operationally missing runs are excluded rather than replaced. The only missing run affecting the Natural-TOFU static confirmation is the Gemma 2 2B, NPO, Storage configuration, for which the relevant aggregation uses $n = 19$ units. Engineering failures are not treated as negative scientific outcomes. Baseline identities, candidate fractions, calibration and final assignments, and gate thresholds are frozen before final outcomes are inspected.

B.4 Controlled Storage and Intervention Construction

Independence. The controlled experiment uses synthetic facts injected through two parameter interfaces in Qwen2.5-7B. Localization is evaluated without access to intervention outcomes. Intervention quality is subsequently measured with NPO runs that share the objective, batches, optimization steps, seeds, and approximately 6M editable scalar parameters. This protocol prevents direct use of downstream intervention outcomes when constructing the localization score.

C Complete Audited Results

C.1 Natural-TOFU Complete Composed Comparison

Decomposition. For baseline C, the Natural-TOFU margin for Intervention Score + DIR-R is

$$\Delta J_{\text{full},C} = \Delta J_{\text{Static},C} + \Delta J_{\text{DIRR,Static}}. \tag{22}$$

Because the two terms come from different final evaluation sets, this equation combines separately estimated components rather than defining a paired estimator. Table 6 exposes both terms instead of reporting only the final margin.

Table 6: Natural-TOFU audited calculation of the Intervention Score + DIR-R margin. Positive values favor our method. Totals are computed from unrounded values; displayed components are rounded.

Objective	Baseline	STATIC-IV $- C$	DIR-R$-$STATIC-IV	Full method$-C$
NPO	Random	+0.002789	+0.002010	+0.004799
NPO	Activation	+0.038996	+0.002010	+0.041006
NPO	WAGLE	+0.134441	+0.002010	+0.136450
NPO	Storage	+0.030793	+0.002010	+0.032802
NPO	MemFlex	+0.092770	+0.002010	+0.094780
NPO	FILA-relative Fisher	+0.001380	+0.002010	+0.003390
NPO	AP-MLP	+0.027762	+0.002010	+0.029772
NPO	GRAIL	+0.000163	+0.002010	+0.002172
NPO	Activation-Down	+0.037585	+0.002010	+0.039595
NPO	Projected Causal	+0.108938	+0.002010	+0.110948
SimNPO	Random	+0.001844	+0.005130	+0.006974
SimNPO	Activation	+0.149612	+0.005130	+0.154743
SimNPO	WAGLE	+0.292638	+0.005130	+0.297768
SimNPO	Storage	+0.157138	+0.005130	+0.162268
SimNPO	MemFlex	+0.205061	+0.005130	+0.210191
SimNPO	FILA-relative Fisher	-0.005930	+0.005130	-0.000799
SimNPO	AP-MLP	+0.093010	+0.005130	+0.098141
SimNPO	GRAIL	-0.004412	+0.005130	+0.000719
SimNPO	Activation-Down	+0.144147	+0.005130	+0.149278
SimNPO	Projected Causal	+0.126655	+0.005130	+0.131786

Boundary Cases. Under NPO, STATIC-IV already has a positive margin over every baseline, so the $+0.002010$ DIR-R increment increases rather than reverses the ordering. SimNPO is more informative. STATIC-IV trails both FILA-relative Fisher by 0.005930 and GRAIL by 0.004412. Adding the $+0.005130$ DIR-R increment is sufficient to move GRAIL to a small positive margin for Intervention Score + DIR-R of $+0.000719$, but not sufficient to overturn FILA-relative Fisher, which remains ahead by 0.000799. The dynamic component therefore contributes a positive descriptive increment to the reported margin while leaving a visible boundary on the method-level claim.

C.2 LACUNA NPO/SIMNPO PARAMETER-SELECTION RESULTS

Objective Sensitivity. Table 7 reports the ten-method parameter-selection comparison. Under NPO, Projected Causal has the highest mean J at 6.538359, followed by MemFlex at 6.374923 and WAGLE at 6.253223. Under SimNPO, MemFlex becomes the best-performing parameter-selection method at 1.745046, FILA rises to 1.605025, while Projected Causal falls to 1.331232. In rank terms, Projected Causal moves from first under NPO to eighth under SimNPO, whereas FILA moves from eighth to second. The same editable geometry therefore produces materially different parameter-selection rankings when only the unlearning objective changes.

Table 7: LACUNA NPO/SimNPO parameter-selection comparison. Each mean for a method and objective uses $n = 12$ unique runs after deduplication.

Baseline	NPO mean J	SimNPO mean J
Random	5.105164	1.035321
Activation exact	6.174900	1.586493
WAGLE	6.253223	1.470818
Storage	5.839900	1.489109
MemFlex	6.374923	1.745046
FILA	5.362761	1.605025
AP-MLP	5.088650	1.316195
GRAIL	5.614737	1.580461
Activation-Down	6.169626	1.585372
Projected Causal	6.538359	1.331232

Static Ordering. The objective-dependent rank reversals are larger than a simple exchange between two near-tied methods. Projected Causal is the best-performing NPO parameter-selection method but is below seven alternatives under SimNPO, while FILA exhibits the opposite pattern. MemFlex is comparatively stable, ranking second under NPO and first under SimNPO. These differences support conditioning initial parameter selection on the actual unlearning objective rather than transferring one static ranking across objectives.

C.3 LACUNA GRADDIFF FOUR-FIELD PRIMARY AND FIELD HETEROGENEITY

Four-Field Primary. The GradDiff closure uses $-g_{\text{forget}} + g_{\text{retain}}$ and contains 12 outcomes covering Birth City, Driver's License, Email Address, and Phone Number with three final seeds per field. In the four-field aggregate, DIR-R has mean $J = 102.628354$. Projected Causal, Activation-Down, FILA, GRAIL, and MemFlex have higher means, while Storage, WAGLE, and AP-MLP have lower means. The largest negative mean gap is -7.104583 against Projected Causal, whereas the largest positive gap is $+3.857708$ against AP-MLP. The primary result therefore establishes neither universal superiority nor universal inferiority.

Table 8: GradDiff four-field primary result. Positive DIR-R gaps favor DIR-R.

Method	Mean J	DIR-R$-$method	Median paired gap	DIR-R W/T/L
Projected Causal	109.732938	-7.104583	-1.412500	4/0/8
Activation-Down	105.319396	-2.691042	-0.030000	6/0/6
FILA	104.562938	-1.934583	+0.620000	7/0/5
GRAIL	103.815438	-1.187083	+1.072500	8/0/4
MemFlex	103.104396	-0.476042	+1.105000	7/0/5
DIR-R	102.628354	N/A	N/A	N/A
STATIC-IV	102.463771	+0.164583	N/R	6/6/0
Storage	102.227313	+0.401042	+2.020000	8/0/4
WAGLE	100.976063	+1.652292	+2.880000	8/0/4
AP-MLP	98.770646	+3.857708	+3.997500	10/0/2

Birth City. The field decomposition in Table 9 shows a common pattern across all eight external parameter-selection methods: every Birth City gap is negative, ranging from -5.872500 against AP-MLP to -28.461667 against Projected Causal. Outside Birth City, most gaps reverse sign. For example, DIR-R trails FILA by 18.645833 on Birth City but leads it by 5.211667, 3.740000, and 1.955833 on Driver's License, Email Address, and Phone Number. Similar sign reversals occur for GRAIL, Storage, WAGLE, and AP-MLP. The four-field aggregate therefore hides substantial variation across data fields.

Table 9: GradDiff field-level DIR-R gaps. Positive values favor DIR-R.

Baseline	Birth City	Driver's License	Email Address	Phone Number
Projected Causal	-28.461667	-0.651667	+0.855000	-0.160000
Activation-Down	-16.937500	+1.838333	+4.823333	-0.488333
FILA	-18.645833	+5.211667	+3.740000	+1.955833
GRAIL	-17.077500	+7.231667	+3.265000	+1.832500
MemFlex	-15.110833	+7.693333	+5.926667	-0.413333
Storage	-9.070833	+6.756667	+3.368333	+0.550000
WAGLE	-8.999167	+8.260000	+6.793333	+0.555000
AP-MLP	-5.872500	+9.175000	+10.338333	+1.790000

Three-Field Sensitivity. Section 5.4 reports the three-field sensitivity obtained by removing the already-completed Birth City runs. That sensitivity changes the aggregate ordering but does not replace the four-field primary above. The replication below further tests whether the apparent three-field near-tie with Projected Causal persists.

C.4 GradDiff Raw DIR-R Audit And Replication

Field Decomposition. Table 10 decomposes the primary four-field GradDiff DIR-R result by field. DIR-R has a higher mean J than STATIC-IV in each field, including Birth City. The large Birth City discrepancy in the external parameter-selection comparison therefore does not arise because DIR-R underperforms its own STATIC-IV baseline on that field; it arises because several alternative static parameter-selection methods attain substantially higher Birth City scores.

Table 10: GradDiff primary four-field DIR-R decomposition. The final column gives triggered runs among the three final seeds.

Field	STATIC-IV J	DIR-R J	DIR-R$-$STATIC-IV	Triggered
Birth City	90.536120	90.737786	$+0.201667$	1/3
Driver's License	111.466042	111.839375	$+0.373333$	2/3
Email Address	118.499479	118.559479	$+0.060000$	1/3
Phone Number	89.353444	89.376777	$+0.023333$	2/3
Overall	102.463771	102.628354	$+0.164583$	6/12

Gate Behavior. The primary GradDiff threshold is frozen at $\tau = 0.02125$ from 12 calibration instances spanning four fields and seeds $73/311/997$. Table 11 summarizes the calibration outcome. All five triggered calibration decisions have positive available gain, but the gate does not identify every beneficial revision: for example, Birth City seed 311 and Driver's License seed 997 have zero probe signal while their best candidate-set gains are $+0.52$ and $+0.22$, respectively. The gate is therefore a conservative screening rule rather than a predictor of every continuation opportunity.

Table 11: GradDiff primary calibration summary. Gain captured is the fraction of total positive available gain over the frozen continuation candidate set.

Quantity	Value
Frozen threshold τ	0.021250
Calibration instances	12
Triggered instances	5/12
Trigger rate	41.67%
Trigger precision	1.000
Captured positive gain	0.9025
Fraction of positive available gain captured	0.476882
Captured gain / relative extra compute	0.192021

Candidate-Set Coverage. The primary four-field evaluation and the three-field replication stored exhaustive five-candidate outcomes only for triggered runs (6/12 and 2/9). A post-hoc supplement evaluated the missing non-trigger branches without changing the frozen threshold, support choices, or original DIR-R actions, yielding complete 12/12 and 9/9 candidate-set denominators. The primary evaluation has mean available candidate-set gain $+0.282500$ and captures 58.26% of it; the replication has mean available gain $+0.183333$ and captures 5.45%. These are single-checkpoint candidate-set quantities, not quantities normalized to the $16\times$ exhaustive multi-checkpoint reference.

Replication. The three-field replication reruns Driver's License, Email Address, and Phone Number on seeds $32771/65537/131071$, keeps the primary threshold $\tau = 0.02125$, and performs no recalibration. All nine DIR-R/STATIC-IV runs and all nine Projected Causal runs complete.

Table 12: GradDiff three-field sensitivity and replication. The first row is derived from the completed primary runs; the second uses new seeds.

Setting	Seeds	STATIC-IV	DIR-R	DIR-R$-$Static	Projected	DIR-R$-$Projected	Median gap	W/T/L	Trigger
Original three-field	4099/8191/16381	106.439655	106.591877	$+0.152222$	106.577433	$+0.014444$	-0.940000	4/0/5	5/9
Replication	32771/65537/131071	107.729099	107.739099	$+0.010000$	108.600488	-0.861389	-1.440000	4/0/5	2/9

Replication Outcome. The original three-field sensitivity places DIR-R only $+0.014444$ above Projected Causal in mean score, with a negative paired median and 4/0/5 W/T/L. In the replication, Projected Causal instead has the higher mean by 0.861389, while W/T/L remains 4/0/5. The paired gaps are highly variable across seeds; exposed examples include $+16.975$ and -10.310 on different Driver's License seeds, -6.725 and -7.205 on two Email Address seeds, and $+2.8425$ on one Phone Number seed. Thus the external ordering is not stable enough to support a three-field superiority claim.

DIR-R. The primary four-field evaluation has mean paired gain $+0.164583$, median $+0.002500$, and 6/6/0 W/T/L across its 12 final runs. The completed denominator shows that this corresponds to 58.26% of the $+0.282500$ mean gain available from exhaustive candidate evaluation at the frozen checkpoint. This quantifies the fraction of candidate-set revision retained by DIR-R. In the three-field replication, only Email Address seed 32771 and Phone Number seed 32771 trigger, with gains of $+0.05$ and $+0.04$; the remaining seven runs return the STATIC-IV continuation. The aggregate gain is $+0.010000$ against $+0.183333$ available candidate-set gain, or 5.45%, so we retain the replication as a near-tie boundary result rather than evidence of a substantial DIR-R gain.

C.5 LACUNA FIXED-SUBSET ANALYSES

Static-IV. On the canonical NPO/SimNPO parent panel, STATIC-IV has combined mean $J = 4.187641$. The corresponding means are 3.070243 for Random, 3.862021 for WAGLE, 3.664504 for Storage, and 4.059985 for MemFlex, producing paired gaps of $+1.117398$, $+0.325620$, $+0.523136$, and $+0.127656$. The MemFlex descriptive interval crosses zero, so the smallest of these four gaps should not be interpreted as a robust separation.

Table 13: LACUNA analysis for modern parameter-selection methods.

Baseline	Baseline mean J	STATIC-IV $- C$	Run wins	Setting wins / 95% CI
FILA-relative Fisher	3.4839	+0.7037	36/48	14/16 $[+0.3765, +1.0244]$
AP-MLP	3.2024	+0.9852	44/48	16/16 $[+0.7097, +1.2726]$
GRAIL	3.5976	+0.5900	39/48	14/16 $[+0.3185, +0.8568]$
Activation-Down	3.8775	+0.3101	37/48	13/16 $[+0.0491, +0.5454]$
Projected Causal	3.9348	+0.2528	35/48	11/16 $[-0.1109, +0.6112]$

Modern Parameter-Selection Methods. STATIC-IV has positive mean gaps against all five modern parameter-selection methods in Table 13. The largest is $+0.9852$ against AP-MLP, accompanied by 44/48 run wins and 16/16 setting wins. The intervals against FILA-relative Fisher, AP-MLP, GRAIL, and Activation-Down remain positive. Projected Causal is closer: the mean gap is $+0.2528$, but its 95% interval $[-0.1109, +0.6112]$ crosses zero. The static evidence therefore supports using Intervention Score as a competitive fixed-subset baseline without establishing separation from every parameter-selection method near the frontier.

C.6 LACUNA DIR-R COMPARISON

Comparison. Table 14 contains the six comparisons used for the main direct method comparison. Every mean gap is positive. NPO improvements range from $+0.430752$ to $+0.847858$, while SimNPO improvements range from $+0.503420$ to $+0.570793$. Every baseline also loses a majority of its 12 runs, with the strongest run-level result occurring against Activation-Down under SimNPO at 12/0.

Table 14: LACUNA comparison of Intervention Score + DIR-R with the selected baselines.

Objective	Baseline	Baseline J	DIR-R J	DIR-R$-C$	Wins/Losses
NPO	Projected Causal	5.763156	6.193908	+0.430752	8/4
NPO	MemFlex	5.502749	6.193908	+0.691159	10/2
NPO	Activation exact	5.346050	6.193908	+0.847858	9/3
SimNPO	MemFlex	1.271154	1.776698	+0.505544	11/1
SimNPO	FILA-relative Fisher	1.205905	1.776698	+0.570793	8/4
SimNPO	Activation-Down	1.273278	1.776698	+0.503420	12/0

Evidence Strength. The LACUNA results provide a direct comparison between Intervention Score + DIR-R and the selected baselines because both sides are evaluated within the same evaluation settings. Natural-TOFU has a different evidential status because its static and dynamic components come from different final evaluation sets.

C.7 LACUNA DIR-R Ablation

Efficiency. Under NPO, DIR-R increases J from 1.325934 to 1.328368, a gain of $+0.002433$, while triggering on 33.3% of runs and producing a wall-time ratio of $1.33\times$. SimNPO increases from 1.680125 to 1.688353, a gain of $+0.008228$, with a 58.3% trigger rate and a $1.52\times$ wall-time ratio. These rows capture 30.32% and 55.34% of the corresponding available candidate-set gain. The post-hoc GradDiff supplement completes the same denominator for the primary evaluation and replication without altering their frozen DIR-R decisions.

Table 15: LACUNA DIR-R versus the fixed-subset STATIC-IV continuation. Available gain and capture use exhaustive candidates at the frozen decision checkpoint.

Objective / setting	STATIC-IV J	DIR-R J	Difference	Trigger	Avail. gain	Captured
NPO	1.325934	1.328368	+0.002433	33.3%	+0.008024	30.32%
SimNPO	1.680125	1.688353	+0.008228	58.3%	+0.014868	55.34%
GradDiff, four-field	102.463771	102.628354	+0.164583	50.0%	+0.282500	58.26%
GradDiff, three-field replication	107.729099	107.739099	+0.010000	22.2%	+0.183333	5.45%

GradDiff Replication. The primary four-field evaluation triggers on 50.0% of its 12 runs. All six triggered runs improve, giving six wins, six ties, and no losses over all 12 runs; the mean triggered gain is $+0.329167$ and the median is $+0.365$. After the frozen decisions, the missing non-trigger continuations were evaluated only to complete the denominator: the primary evaluation captures 58.26% of the available candidate-set gain. The three-field replication triggers on 2/9 runs and captures only 5.45%, consistent with a more conservative gate in that evaluation.

Audit Summary. The audited results support three distinct empirical conclusions. First, Intervention Score has positive mean gaps against several parameter-selection methods, but neither Natural-TOFU nor GradDiff supports universal static dominance. Second, LACUNA NPO/SimNPO has positive mean margins in all six reported baseline comparisons. Third, the aggregate DIR-R-minus-STATIC-IV mean is positive in each reported DIR-R evaluation set, including both GradDiff evaluation sets in which the external static parameter-selection ordering is field- and seed-sensitive. The data therefore support a decomposition between initial intervention selection and checkpoint-dependent revision rather than a claim that one parameter-selection method or DIR-R dominates every setting.