U2E: PROTECTED COMMON DESCENT FOR TARGETED BEHAVIORAL FACTUAL UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Targeted factual unlearning seeks to suppress a small set of obsolete or sensitive associations in a deployed language model while preserving its remaining knowledge and capabilities. Knowledge editing offers a local intervention, but its formulation around a replacement value does not determine whether one protected update can suppress the old fact across several prompts and internal representations whose required directions may conflict. Building on the key–value view of transformer memory and covariance and null-space protection, we formulate factual unlearning as protected common descent. This formulation characterizes whether multiple access-specific decreases admit a shared protected update and, when feasible, the minimum cost of satisfying the local constraints. The associated quadratic problem has a unique solution with an active-support rank bound. Based on this formulation, U2E constructs low-rank candidate updates using local progress estimates and validates them on the full nonlinear model before commitment. Experiments demonstrate favorable tradeoffs between old-answer suppression and measured retention under the evaluated protocols, supporting protected common descent as a practical framework for selective factual unlearning.

1 INTRODUCTION

Language models deployed in practice may retain facts that later become incorrect, obsolete, sensitive, or subject to removal requests. Targeted factual unlearning seeks to suppress a small set of such associations while preserving the model's remaining knowledge and capabilities. Retraining without the relevant examples provides a clean conceptual reference, but is often infeasible when the original data, training state, or compute budget is unavailable (Bourtoule et al., 2021; Maini et al., 2024). Continued optimization methods such as NPO (Zhang et al., 2024) and RMU (Li et al., 2024) can produce strong forgetting, but may spread a narrow request into broad behavioral changes. For the small targeted removal regime considered here, a more local and selective intervention is therefore desirable.

Knowledge editing offers local parameter updates, but ordinarily writes a desired replacement value. Unlearning specifies what should be suppressed but does not prescribe a unique replacement. Writing "unknown," a refusal, or a neutral value may change one answer without weakening the old fact itself (Patil et al., 2024; Li et al., 2026; Lin et al., 2026). The fact may still be elicited by the original question, a paraphrase, or another linguistic form, and may remain recoverable from an internal representation. A factual removal request must therefore reduce the old answer across specified external and internal accesses while preserving other unrelated knowledge. These requirements share the same parameters, and knowledge protection removes some available update directions. Consequently, accesses that are editable separately need not remain jointly feasible under protection. This makes joint feasibility a prerequisite for constructing a protected update: whether the requested decreases admit a common solution under the preservation constraints.

We address this question by combining established knowledge editing mechanisms with convex analysis. The standard view of transformer feedforward modules as key–value memories represents a factual association by an input key and its corresponding output (Geva et al., 2021; Meng et al., 2022a). Linearization expresses how a small weight change affects the old answer at each access, while covariance weighting and null-space projection penalize changes on ordinary inputs and forbid changes along designated keys (Meng et al., 2022b; Gupta et al., 2024; Fang et al., 2025). Under these

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.

protections, each required decrease defines a linear halfspace in the remaining update space. Farkas' lemma states that either these halfspaces share a feasible update, or a nonnegative combination of their requirements cancels every permitted direction and certifies a conflict (Boyd & Vandenberghe, 2004). When feasible, strict convexity gives a unique update with minimum predicted disruption to retained inputs, whose optimality conditions express the solution as a sum of active rank-one terms and directly yield a low-rank update. We further impose trust constraints on value changes to quantify attainable common progress at finite steps, completing the formulation *protected common descent*.

This analysis yields U2E, which converts the requested old answer decreases and preservation requirement into a selective model edit. It measures complementary external and internal accesses, constructs their protected local directions, and tests whether they admit a common descent update. When feasible, U2E selects trusted progress and computes the unique minimum-protection-cost low-rank update. The candidate is then evaluated on the full nonlinear model. If its realized forgetting or preservation departs from the local prediction, U2E relinearizes and refines the feasible edit set, committing only updates that satisfy the specified full-model checks.

Across ten MultiCounterFact runs on Llama-3.2-3B-Instruct, U2E suppresses the old-answer access across linguistic forms: old-answer accuracy falls from 17.6% to 3.0% on original questions and from 18.1% to 6.3% on paraphrases relative to ZeroUnlearn. This stronger suppression remains selective rather than arising from broad model degradation: retained factual behavior stays similar, while the perplexity shift drops from 0.996 to 0.039, with 93.4% of requests handled.

The contributions of this work are as follows:

- We formulate targeted behavioral factual unlearning as the joint feasibility of multiple old-answer access decreases under explicit knowledge-preservation constraints.
- For the protected local response model, we derive a necessary and sufficient conflict certificate and the unique minimum-cost update using classical convex analysis, establish an active-support rank bound under this local model, and formulate a separate trust-constrained progress problem.
- We instantiate protected common descent as U2E, a multi-access behavioral factual unlearning method that jointly coordinates external and internal suppression requirements through minimum-protection-cost low-rank updates.

2 Related Work

Machine unlearning studies how to remove the influence of selected training examples without full retraining (Bourtoule et al., 2021). In language models, TOFU (Maini et al., 2024) and OpenUnlearning (Dorna et al., 2025) broaden this goal to include both targeted forgetting and retained utility. Gradient ascent directly lowers the likelihood of forget data, NPO (Zhang et al., 2024) and SimNPO (Fan et al., 2025) stabilize this objective through preference-based learning, and RMU (Li et al., 2024) intervenes in representation space. These methods learn a tradeoff between forgetting and retention through iterative optimization. U2E instead focuses on local behavioral factual access removal and determines, under a first-order protected response model, whether several forgetting requirements admit a common update that preserves specified knowledge.

Knowledge editing modifies selected associations through localized parameter updates. The key–value view of transformer feed-forward modules (Geva et al., 2021) motivates ROME (Meng et al., 2022a) for rank-one editing, MEMIT (Meng et al., 2022b) for multi-layer editing, and EMMET (Gupta et al., 2024) for covariance-weighted memorization–preservation. AlphaEdit (Fang et al., 2025) protects existing knowledge through null-space projection, while WISE (Wang et al., 2024) introduces a separate memory mechanism for continual editing; unlike these single-query or memory-based approaches, U2E jointly solves multiple external and internal access constraints and returns either a conflict certificate or a minimum-cost low-rank update. Building on these local writeback and protection mechanisms, U2E represents each prompt or internal readout of an old fact as a distinct access and determines whether these accesses remain compatible after protection, together with the unique minimum-protection-cost low-rank update when they are compatible.

Recent work connects editing with factual unlearning. Editing as Unlearning (Li et al., 2026) studies knowledge editors with refusal and empty-set targets, while ZeroUnlearn (Lin et al., 2026) maps

sensitive inputs toward a neutral state. Mechanistic Unlearning (Guo et al., 2024) uses causal localization to improve robustness, CAE (Han et al., 2026) aggregates prompts for entity-level consistency, and KnowledgeSmith (Luo et al., 2026) evaluates editing and unlearning within a common framework. Studies of sensitive deletion further show that facts may remain recoverable through black-box or white-box extraction after a successful surface-level edit (Patil et al., 2024). U2E advances this direction with a feasibility-first formulation that maps external and internal old-answer responses into a protected metric space and derives compatibility, protection cost, low-rank realization, and fact-level abstention from one protected common-descent problem.

3 Problem Formulation

A factual association is commonly modeled as a mapping from a subject representation to an object prediction, enabling localized parameter editing (Geva et al., 2021; Meng et al., 2022a;b). Targeted behavioral factual unlearning, however, cannot be defined by a single query: the same old answer may remain accessible through paraphrases or internal representations, and entity-level removal requires consistency across expressions (Patil et al., 2024; Guo et al., 2024; Han et al., 2026). We therefore represent each removal request by a finite family of external and internal accesses that must be suppressed by the same parameter update.

Let W_ℓ denote the pretrained weight of editable layer ℓ and ΔW_ℓ its update, with W and ΔW collecting them across layers. We index the m accesses by $i = 1, \dots, m$ and associate each with an old-answer score $s_i(W)$. For an external access, s_i is the old answer's log-probability relative to a fixed competitor (Meng et al., 2022a; Patil et al., 2024); for an internal access, the same contrast is read from a fixed hidden state through the model's vocabulary head (Belrose et al., 2023; Guo et al., 2024). This common score representation reduces multi-access suppression to a single question: how does one update ΔW change each s_i?

To characterize this response, we use the key–value view of transformer feed-forward layers, where the input at the subject token acts as a factual key and the corresponding layer output as its value (Geva et al., 2021; Meng et al., 2022a). For access i at layer ℓ, let $k_{i\ell}$ denote this input and $u_{i\ell}$ the gradient of s_i with respect to the editable map's output. A first-order expansion (Conn et al., 2000) then gives

$$s_i(W + \Delta W) - s_i(W) = \sum_\ell u_{i\ell}^\top \Delta W_\ell k_{i\ell} + \varepsilon_i(\Delta W), \tag{1}$$

where $\varepsilon_i(\Delta W)$ collects finite-step and off-position effects. Thus $(k_{i\ell}, u_{i\ell})$ determines the local response of access i to the shared update ΔW. The remaining question is which such updates are compatible with preserving unrelated knowledge, which we encode through the protected geometry below.

3.1 Retention geometry

ROME (Meng et al., 2022a), MEMIT (Meng et al., 2022b), and EMMET (Gupta et al., 2024) measure collateral change using the distribution of pre-existing keys. Following this view, let $C_\ell \succ 0$ be a damped second-moment matrix of ordinary inputs at layer ℓ. We measure the cost of an update ΔW by

$$\langle X, Y \rangle_C = \sum_\ell \operatorname{tr}(X_\ell C_\ell Y_\ell^\top), \qquad \|X\|_C^2 = \langle X, X \rangle_C. \tag{2}$$

To additionally keep designated keys unchanged, let the columns of P_ℓ contain their layer-ℓ representations. Following null-space editing (Fang et al., 2025), admissible updates are restricted to

$$\mathcal{H}_P = \{\Delta W : \Delta W_\ell P_\ell = 0 \; \forall \ell\}. \tag{3}$$

We now express the access response from Eq. (1) within this protected geometry. For each access i, there is a unique $a_i \in \mathcal{H}_P$ such that

$$\langle \Delta W, a_i \rangle_C = \sum_\ell u_{i\ell}^\top \Delta W_\ell k_{i\ell}, \qquad \Delta W \in \mathcal{H}_P, \tag{4}$$

by the Riesz representation theorem (Kreyszig, 1991). We call a_i the *protected access direction*: it is the direction representing how an admissible update changes access i under the retention geometry.

Its layerwise form is $a_{i\ell} = u_{i\ell}\bar{k}_{i\ell}^\top$, where $\bar{k}_{i\ell}$ is the covariance-weighted key after projection away from P_ℓ; Appendix A gives the explicit expression.

Let $b_i > 0$ denote the required decrease of access score s_i. Using Eq. (4), the first-order requirement from Eq. (1) becomes

$$\langle \Delta W, a_i \rangle_C \le -b_i, \qquad i = 1, \dots, m. \tag{5}$$

Thus each access defines a halfspace in the same protected update space $\mathcal{H}_P$, and factual unlearning requires a single ΔW in their intersection. Section 4 characterizes when such a common protected update exists and constructs its minimum-cost solution.

4 Protected Common Descent

Projecting the zero update onto the intersection defined by (5) selects the least disruptive feasible update, following preservation–memorization formulations in constrained editing (Meng et al., 2022a;b; Gupta et al., 2024):

$$\min_{\Delta W \in \mathcal{H}_P} \ \tfrac{1}{2}\|\Delta W\|_C^2 \quad \text{s.t.} \quad \langle \Delta W, a_i \rangle_C \le -b_i,\ i = 1, \dots, m. \tag{6}$$

Positive-definite C_ℓ makes the objective strictly convex and hence the solution unique whenever the constraint set is nonempty (Boyd & Vandenberghe, 2004; Rockafellar, 1970). It remains to characterize when such a protected common update exists.

Theorem 1 (local feasibility and minimum-cost update). The theorem of alternatives for linear inequalities (Boyd & Vandenberghe, 2004; Rockafellar, 1970) states that problem (6) is feasible if and only if no nonnegative weights $\lambda = (\lambda_1, \dots, \lambda_m)$ satisfy

$$\sum_i \lambda_i a_i = 0, \qquad \sum_i \lambda_i b_i > 0. \tag{7}$$

Thus, λ is an explicit certificate of directional conflict: a nonnegative combination of the required access directions cancels in the permitted space, so no protected update can decrease them all together.

To construct that update, let $b = (b_1, \dots, b_m)^\top$ and let the Gram matrix $Q_{ij} = \langle a_i, a_j \rangle_C$ encode agreement between protected directions (Kreyszig, 1991). Convex duality (Boyd & Vandenberghe, 2004) yields

$$\lambda^* \in \arg\max_{\lambda \ge 0} \{\lambda^\top b - \tfrac{1}{2}\lambda^\top Q \lambda\}, \qquad \Delta W_\ell^* = -\sum_i \lambda_i^* a_{i\ell}. \tag{8}$$

Proof sketch. Weighting the feasible constraints by a certificate in (7) would imply $0 < 0$. Conversely, Farkas' lemma supplies a common direction when no certificate exists; Lagrange duality gives (8), and strict convexity gives uniqueness. Appendix B provides the proof.

Because each $a_{i\ell}$ has the outer-product structure of ROME and MEMIT (Meng et al., 2022a;b),

$$\operatorname{rank}(\Delta W_\ell^*) \le \big|\{i : \lambda_i^* > 0\}\big|. \tag{9}$$

The number of active constraints bounds the rank of the optimal update. The theorem and its dual therefore provide a complete characterization of the local geometric problem: they determine whether a common protected direction exists and, when it does, construct its least-cost realization. The trust-region stage below turns this local solution into a finite candidate update; fresh forward passes determine the committed edit.

4.1 Finite trusted progress

The common-descent solution is derived from a local linear model, so a feasible decrease need not remain reliable for a finite update. Following trust-region methods (Conn et al., 2000), we therefore bound each value change $\Delta W_\ell k_{i\ell}$ by a layer radius $\rho_\ell > 0$ and maximize the fraction t of the requested decreases achievable within these bounds:

$$\begin{aligned} t^* = \max_{0 \le t \le 1,\ \Delta W \in \operatorname{span}\{a_1, \dots, a_m\}} \ & t \\ \text{s.t.} \quad & \langle \Delta W, a_i \rangle_C \le -t b_i, \quad \|\Delta W_\ell k_{i\ell}\|_2 \le \rho_\ell, \quad \forall i, \ell. \end{aligned} \tag{10}$$

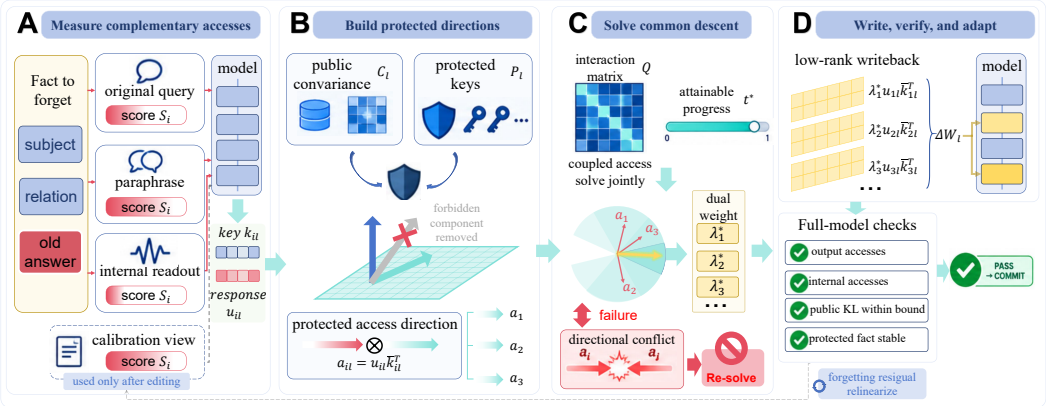

Figure 1: U2E overview. U2E maps complementary output and internal accesses into a protected geometry, determines whether they admit a common descent, and realizes the feasible solution as a minimum-protection-cost low-rank update. Full-model measurements guide relinearization and refinement when the local prediction is insufficient.

Here t^* is the largest common fraction of the requested decreases supported by the local model within the trusted region: $t^* = 1$ permits the full step, whereas $0 < t^* < 1$ yields a correspondingly smaller update. We use this trusted progress to construct the next candidate, and remeasure its effect on the full model before any further refinement.

4.2 Accounting for Nonlinearity

The trust region limits the step size but does not eliminate error in the local linear model. Following trust-region methodology (Conn et al., 2000), U2E therefore evaluates each candidate on the full nonlinear model and compares the realized forgetting and preservation with their local predictions. When the candidate falls short, U2E remeasures the access responses at the updated weights and resolves for the remaining decrease. This prediction-measurement refinement connects the local common-descent analysis to finite model edits (details in Appendix C).

5 U2E

We now instantiate protected common descent as U2E for factual removal in transformer language models. Figure 1 summarizes the pipeline, and Algorithm 1 gives the complete procedure.

5.1 Constructing Accesses and Selecting Layers

For each factual removal request, U2E constructs multiple prompt views and measures the old answer relative to a fixed comparison answer. It additionally selects internal readouts on the unedited model and fixes their locations before editing, yielding complementary output and internal accesses as defined in Section 3. One prompt view is held out from construction to assess transfer during layer selection. The concrete prompt and readout choices are given in Appendix E.

For each access i and candidate MLP `down_proj` layer ℓ, U2E measures the key $k_{i\ell}$ and response $u_{i\ell}$ from Eq. (1). It selects the layer whose predicted update transfers most strongly to the held-out view, breaking ties by lower protection cost and earlier depth. Requests already satisfying the suppression criterion are treated as no-ops, while those without an editable response are excluded from the common-descent solve.

5.2 Solving and Writing the Update

U2E estimates the retention geometry C_ℓ from public inputs and constructs the protected-key matrix P_ℓ from representative public and relation-matched keys; all protection data are disjoint from the

Algorithm 1 U2E: Protected Multi-Access Factual Unlearning

Require: weights W; factual removal requests S; candidate layers; protection examples
Ensure: edited weights and a status for every requested fact

while $S \neq \varnothing$ **do**
  construct output and internal accesses for each request in S
  measure $s_i(W)$, $k_{i\ell}$, and $u_{i\ell}$ at each candidate layer ℓ
  select one layer per request by transfer to the held-out prompt view
  record no-ops and exclude requests without editable support
  estimate C_ℓ and construct P_ℓ from protection examples
  compute $\bar{k}_{i\ell}$ by Eq. (12) and $a_{i\ell} = u_{i\ell}\bar{k}_{i\ell}^\top$
  form $Q_{ij} = \langle a_i, a_j \rangle_C$
  solve Eq. (10) for t^*
  solve $\lambda^* \in \arg\max_{\lambda \geq 0}\{\lambda^\top (t^* b) - \frac{1}{2}\lambda^\top Q\lambda\}$
  form $\Delta W_\ell^* = -\sum_i \lambda_i^* u_{i\ell}\bar{k}_{i\ell}^\top$
  evaluate $W + \Delta W^*$ on the specified forgetting and preservation checks
  if forgetting is insufficient and preservation holds **then**
    remeasure $k_{i\ell}$ and $u_{i\ell}$ at $W + \Delta W^*$
    update b to the remaining decrease and re-solve for ΔW^*
  end if
  if all specified checks are satisfied **then**
    $W \leftarrow W + \Delta W^*$ and record accepted requests
  else
    reduce S using failed checks and predicted protection cost
  end if
end while
return W and request statuses

evaluation retain set. Detailed sampling choices are given in Appendix E. Projecting the measured keys by Eq. (12) gives the protected directions and their Gram matrix

$$Q_{ij} = \sum_\ell (u_{i\ell}^\top u_{j\ell})(\bar{k}_{i\ell}^\top C_\ell \bar{k}_{j\ell}), \qquad \Delta W_\ell^* = -\sum_i \lambda_i^* u_{i\ell}\bar{k}_{i\ell}^\top. \tag{11}$$

Using the trusted progress t^*, U2E solves Eq. (8) with target decrease $t^* b$ and writes ΔW_ℓ^* to the selected MLP layers. Its outer-product form directly realizes the protected common-descent solution through the standard low-rank ROME/MEMIT interface (Meng et al., 2022a;b).

5.3 Adaptive refinement

After each candidate update, U2E remeasures the constructed and held-out accesses together with the designated preservation criteria. If suppression remains insufficient while preservation holds, it relinearizes at the candidate weights and solves for the residual decrease. If the joint requirements remain unsatisfied after refinement, U2E reduces the active request set according to the observed failures and predicted protection cost and resolves the problem. Requests for which no admissible update is found are reported as abstentions; the exact acceptance criteria and refinement budget are given in Appendix E.

5.4 Complexity

Let m be the number of active accesses, L the edited layers, d their maximum width, and P the number of model parameters. After covariance factorization, projection and construction of Q cost $O(Lmd^2 + Lm^2 d)$, while the dense dual solve requires $O(m^3)$ time and $O(m^2)$ memory (Boyd & Vandenberghe, 2004). Relative to direct model-editing approaches, U2E therefore adds an access-level optimization whose dimension grows with m, not P. This adds a bounded access-level cost beyond direct low-rank editing, with the additional optimization confined to the typically small space $m \ll P$. At the same time, its optimization scale remains far below that of optimization-based unlearning, which requires repeated parameter updates over multiple training steps in a parameter space that can scale with P. The resulting dual coefficients write back directly as low-rank factors.

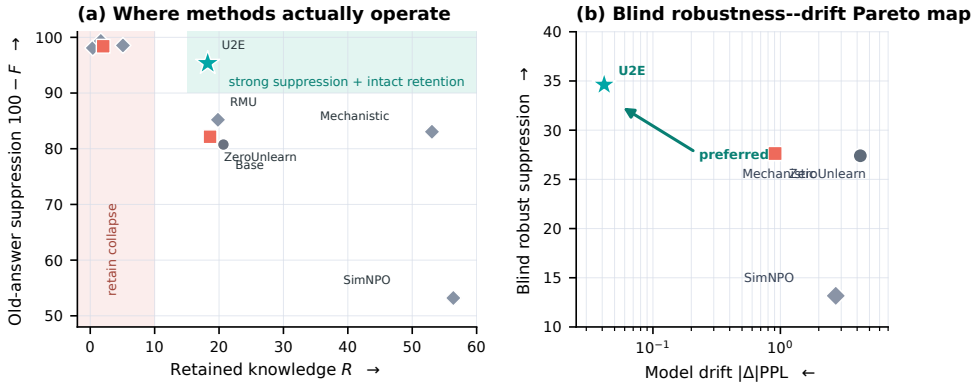

Figure 2: Forgetting-retention maps on MultiCounterFact and unseen accesses. U2E combines strong old-answer suppression with retained knowledge and low model drift, whereas strong suppression alone does not imply preservation.

Each unsuccessful outer round reduces the active request set, so at most the initial number of requests can trigger reduced-batch resolves (Appendix D).

6 Experiments

Our experiments aim to answer six questions: whether U2E achieves selective forgetting while preserving retained knowledge, which components are responsible for this behavior, whether suppression generalizes to unseen accesses, how the method responds to its principal budgets, whether the predicted geometry reflects realized model changes, and what computational cost this entails.

Datasets and models. Our main study evaluates Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on MultiCounterFact (Meng et al., 2022b). Each of ten seeds contains 50 forget facts and 1,000 disjoint retain facts under the case-balanced ZeroUnlearn (Lin et al., 2026) protocol. A separate six-seed evaluation holds out nine access families, including natural paraphrases, aliases, contextual and reverse queries, same-relation multiple choice, Arabic, French, Japanese, and black-box attacks.

Baselines. We compare with loss-based unlearning methods GA, NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025), and UNDIAL (Dong et al., 2025); representation-level methods RMU (Li et al., 2024) and Mechanistic Unlearning (Guo et al., 2024); and neutral remapping with ZeroUnlearn (Lin et al., 2026). We use each method's released implementation and the same case-balanced evaluator. The preservation-calibrated comparison additionally includes AlphaEdit as a strong local-update reference.

Metrics and statistics. Forget efficacy (FE) and generalization (FG) are old-answer accuracies and are lower-is-better. Forget specificity (FS) and retain efficacy/generalization/specificity (RE/RG/RS) measure preservation and are higher-is-better (Dorna et al., 2025). All are absolute post-edit accuracies, so preservation scores may exceed Base; higher FE/FG, however, means easier old-answer retrieval. We define $F = (\mathrm{FE} + \mathrm{FG})/2$ and $R = (\mathrm{RE} + \mathrm{RG} + \mathrm{RS})/3$ only for visualization. Behavioral metrics use all requested facts as the denominator, so accepted, no-op, and abstained requests all contribute; coverage is the fraction accepted or already satisfied as no-ops and is reported separately. Seeds are replication units; confidence intervals use 20,000 paired seed bootstraps, and primary tests use exact sign flips with Holm correction. Worst-access residual averages prompts within each fact-access family and then takes the largest family value for that fact; Extraction@20 is the fraction of facts whose old answer is recovered in at least one of 20 fixed target-free attempts.

Implementation. U2E edits the MLP `down_proj` modules at layers $4, 6, \ldots, 26$. Output and internal forgetting ratios are 0.25 and 0.35; the public mean/max KL limits are $0.002/0.01$; and the neighboring-fact log-probability drop is limited to 0.1. The solver uses FP64 and performs at most three relinearization rounds. Appendix E gives the remaining construction and protection settings.

Table 1: Ten-seed MultiCounterFact mean (%) across model scales. FE/FG are lower-is-better; the remaining metrics are higher-is-better.

	Llama-3.2-3B-Instruct						Llama-3.1-8B-Instruct					
Method	FE↓	FG↓	FS↑	RE↑	RG↑	RS↑	FE↓	FG↓	FS↑	RE↑	RG↑	RS↑
Base	18.20	20.30	19.60	21.67	20.12	20.25	24.80	22.50	21.80	25.71	23.63	22.68
GA (Maini et al., 2024)	2.00	1.80	1.06	0.40	0.39	0.41	1.20	1.00	0.26	0.24	0.25	0.27
NPO (Zhang et al., 2024)	0.80	2.10	2.50	5.32	5.03	4.89	22.80	21.90	21.28	25.83	23.51	22.55
SimNPO (Fan et al., 2025)	44.60	49.00	55.88	59.76	53.71	55.67	57.00	59.10	61.78	71.00	66.36	62.86
RMU (Li et al., 2024)	14.40	15.20	17.82	20.70	19.11	19.69	19.00	18.50	21.00	24.82	22.97	21.91
UNDIAL (Dong et al., 2025)	0.00	1.30	1.08	1.85	1.39	1.68	0.00	0.90	0.28	0.63	0.50	0.55
Mechanistic (Guo et al., 2024)	0.00	33.90	49.26	56.69	50.47	52.03	0.00	31.60	51.48	65.09	61.53	56.27
ZeroUnlearn (Lin et al., 2026)	17.60	18.10	17.46	19.26	18.65	17.91	23.20	22.40	20.60	25.25	23.22	21.71
U2E	3.00	6.30	15.82	19.39	17.18	18.15	7.60	13.70	19.44	24.57	22.19	21.59

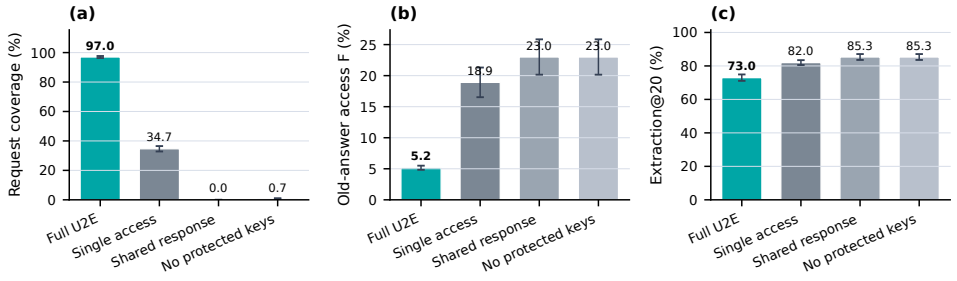

Figure 3: Six-seed component ablation (mean ± standard error).

Main comparison. Table 1 shows that U2E suppresses both direct and paraphrased access while largely preserving retained behavior. Against ZeroUnlearn, it lowers FE/FG from $17.60/18.10$ to $3.00/6.30$; paired gains are 14.6 (95% CI $[12.0, 17.4]$) and 11.8 points ($[8.7, 15.0]$), with Holm-adjusted $p = 0.00586$, similar aggregate retention, and only $1/25.3$ as much perplexity shift. Mechanistic Unlearning reaches FE $= 0$ but raises FG to $33.90/31.60$ on 3B/8B, whereas U2E reaches $6.30/13.70$ with near-base retention. Figure 2 makes this tradeoff explicit: strong suppression alone does not imply preservation, while U2E combines cross-form suppression with retained knowledge and low model drift.

Transfer across model scale. The 8B results in Table 1 confirm the same pattern: relative to ZeroUnlearn, U2E lowers FE/FG from $23.2/22.4$ to $7.6/13.7$, with paired gains of 15.6 points (95% CI $[8.6, 22.2]$) and 8.7 points ($[4.7, 12.5]$; adjusted $p = 0.0352$). Aggregate retention remains within 0.61 points of ZeroUnlearn, and perplexity shift falls from 0.8944 to 0.0546 at 67.0% request coverage. Together, these results show that the low-drift suppression–preservation tradeoff persists as model scale increases.

Component ablation. Figure 3 examines three ingredients used to instantiate protected common descent in U2E. Restricting construction to the dataset query reduces coverage from 97.0% to 34.7%. Replacing the fact-specific response with a shared response, or removing protected keys, leaves only 0.0% or 0.7% coverage and reverts toward the base old-answer accuracy and extractability. These results support complementary roles for these ingredients: access diversity captures alternative retrieval routes, fact-specific responses determine what must decrease, and protected projection aligns these decreases with the preservation constraints.

Hyperparameter sensitivity. We vary forgetting strength, the public KL budget, and the value trust-region radius one factor at a time over $\{0.5, 1, 2\}\times$ on three development seeds; the canonical setting is fixed before this analysis. Across the seven configurations, retained knowledge R stays between 18.61% and 20.13% and mean absolute perplexity shift is at most 0.0294, while coverage ranges from 42.0% to 83.3%. Tightening the KL budget and doubling the value radius reduce committability, exposing the role of the full-model acceptance checks rather than a uniformly insensitive solver. Appendix G reports the descriptive curves and seed variation.

Generalization to unseen accesses. For each held-out route, prompts are averaged within fact before aggregating over facts. Against ZeroUnlearn, U2E reduces worst-route residual by 0.0699 (95% CI $[0.0374, 0.0957]$) and Extraction@20 by 0.1100 ($[0.0633, 0.1567]$), with Holm-adjusted

Table 2: Six-seed robustness on nine unseen routes. Worst is the per-fact maximum route residual; Ext.@20 measures black-box recovery. Lower is better.

Method	Worst↓	Ext.@20↓	PPL↓
U2E	**.654**	**.730**	**.041**
ZeroUnlearn	.724	.840	.908
SimNPO	.868	.813	2.719
Mechanistic	.726	.817	4.237

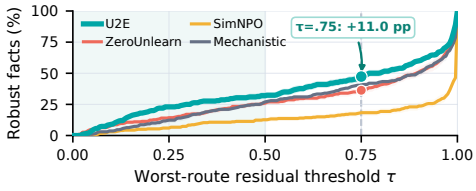

Figure 4: Fraction of facts simultaneously below residual threshold τ on all nine held-out routes, averaged over six seeds.

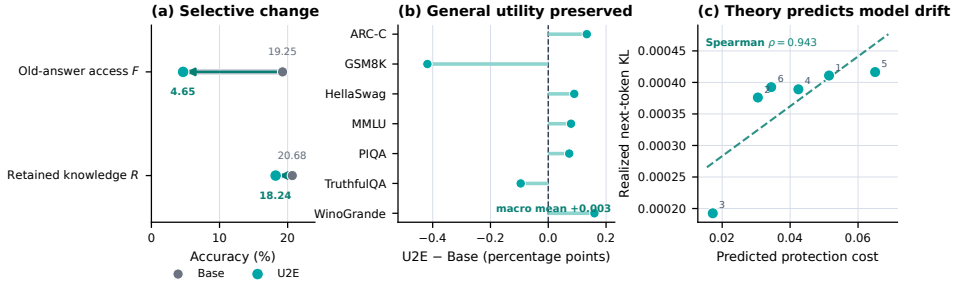

Figure 5: Mechanism validation: (a) selective target suppression, (b) stable general capabilities, and (c) agreement between predicted protection cost and realized KL.

$p = 0.03125$ for both. Figure 4 shows the full curve: at the prespecified $\tau = 0.75$, 47.3% of U2E facts remain below the threshold on every unseen route, compared with 36.3% for ZeroUnlearn. Against Mechanistic Unlearning, U2E also lowers these metrics from $.726/.817$ to $.654/.730$, while its perplexity shift is only about $1/103$ as large ($.041$ versus 4.237). These results show improved suppression on held-out accesses with low model drift, although 73.0% of facts remain recoverable under Extraction@20. At preservation-calibrated operating points, U2E also achieves the lowest worst-route residual, Extraction@20, and perplexity shift among five methods at 97.0% coverage (Appendix K).

Mechanism and theory in practice. Figure 5(a–b) shows that the edit remains selective beyond the primary benchmark metrics. Target old-answer access F falls from 19.25% to 4.65%, whereas retained knowledge R changes from 20.68% to 18.24%. Across three fixed 3,500-example panels and seven utility tasks, the macro average is 62.087% after editing versus 62.084% before editing, with every task changing by only -0.42 to $+0.16$ points.

Figure 5(c) further connects the predicted geometry to realized model change: protection cost strongly tracks next-token KL (Spearman $\rho = 0.943$, 95% CI $[0.50, 1.00]$, adjusted $p = 0.0167$). Under a controlled trust-radius intervention, 272 of 274 interior-progress candidates pass full-model validation; across 4,704 exact node solves, every update satisfies the strictly-positive active-support rank bound in (9) (Appendix H). A 50-fact warm-start edit takes 191.54 seconds, mostly for joint constraint solving; Appendix I gives the full profile.

7 Conclusion and Limitations

We introduced protected common descent, a feasibility-first formulation of factual unlearning that determines whether multiple access-specific decreases admit a shared update under explicit knowledge protection. When feasible, it yields a unique minimum-protection-cost solution with an active-support low-rank structure, which U2E realizes as a selective model edit. Experiments on the evaluated Llama models demonstrate strong cross-form suppression with limited changes to retained knowledge and model behavior, supported by component and geometry analyses. Our current formulation characterizes specified accesses within a local response model and is designed for targeted removal requests. Extending this framework to broader access discovery, stronger cross-architecture transfer, and more efficient large-request solving provides natural directions for future work.

AI USE STATEMENT

Generative AI tools were used solely for language editing. The authors verified the final manuscript and take full responsibility for the submission.

ETHICS STATEMENT

We evaluate factual unlearning on language-model benchmarks. Targeted editing may support legitimate deletion requests but may also facilitate information manipulation. The reported behavioral suppression and validation results do not certify removal of training-data influence; broader implications are discussed in Appendix M.

REPRODUCIBILITY STATEMENT

The supplementary artifact provides machine-readable numerical summaries and maps every reported table and figure to its source measurements. The appendix specifies environments, task contexts, training schedules, evaluation units, statistical procedures, and complete proofs.

REFERENCES

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

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.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust Region Methods*. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2000. doi: 10.1137/1.9780898719857.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramón Huerta, and Ivan Vulić. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

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

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

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

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.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *arXiv preprint arXiv:2410.12949*, 2024.

Akshat Gupta, Dev Sajnani, and Gopala Anumanchipalli. A unified framework for model editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15403–15418, 2024. doi: 10.18653/v1/2024.findings-emnlp.903.

Xiaoqi Han, Víctor Gutiérrez-Basulto, Ru Li, Xiaoli Li, Jiye Liang, and Jeff Z. Pan. Consistency-aware editing for entity-level unlearning in language models. *arXiv preprint arXiv:2601.08840*, 2026.

Erwin Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, 1991.

Nathaniel Li, Alexander Pan, Anjali Gopal, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning*, 2024.

Zexi Li, Xiangzhu Wang, William F Shen, Meghdad Kurmanji, Xinchi Qiu, Dongqi Cai, Chao Wu, and Nicholas D Lane. Editing as unlearning: Are knowledge editing methods strong baselines for large language model unlearning? In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pp. 37627–37635, 2026.

Yujie Lin, Chengyi Yang, Zhishang Xiang, Yiping Song, and Jinsong Su. ZeroUnlearn: Few-shot knowledge unlearning in large language models. In *International Conference on Machine Learning*, 2026.

Yinyi Luo, Zhexian Zhou, Hao Chen, Kai Qiu, Marios Savvides, Sharon Li, and Jindong Wang. Knowledgesmith: Uncovering knowledge updating in llms with model editing and unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 117241–117279, 2026.

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

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Vaidehi Ramesh Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. In *International Conference on Learning Representations*, volume 2024, pp. 45497–45514, 2024.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems*, 2024.

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

A Protected Projection

The protected access direction in Section 3 has the following explicit layerwise form:

$$\bar{k}_{i\ell} = \left[C_\ell^{-1} - C_\ell^{-1} P_\ell \big(P_\ell^\top C_\ell^{-1} P_\ell\big)^\dagger P_\ell^\top C_\ell^{-1}\right] k_{i\ell}, \qquad a_{i\ell} = u_{i\ell} \bar{k}_{i\ell}^\top. \tag{12}$$

Here $\bar{k}_{i\ell}$ is the covariance-weighted key after removing components along the protected keys; $\dagger$ denotes the Moore–Penrose pseudoinverse. This construction gives $P_\ell^\top \bar{k}_{i\ell} = 0$ and the representation identity in (4).

B Proof of Theorem 1

Work in the finite-dimensional Hilbert space $\mathcal{H}_P$ with inner product $\langle \cdot, \cdot \rangle_C$. For necessity, assume that ΔW satisfies all constraints in (6) and that a certificate (7) exists. Then

$$0 = \left\langle \Delta W, \sum_i \lambda_i a_i \right\rangle_C = \sum_i \lambda_i \langle \Delta W, a_i \rangle_C \leq -\sum_i \lambda_i b_i < 0, \tag{13}$$

a contradiction.

For sufficiency, let x be the coordinate vector of ΔW in any basis of $\mathcal{H}_P$, and stack the linear functionals $\langle \Delta W, a_i \rangle_C$ as the rows of a matrix A. Farkas' lemma states that exactly one of the following holds: (i) $Ax \leq -b$ has a solution; or (ii) there exists $\lambda \geq 0$ with $A^\top \lambda = 0$ and $b^\top \lambda > 0$. The adjoint equation $A^\top \lambda = 0$ is exactly $\sum_i \lambda_i a_i = 0$ because a_i is the Riesz representative of row i. Excluding (ii) therefore gives feasibility.

For optimality, introduce multipliers $\lambda_i \geq 0$:

$$\mathcal{L}(\Delta W, \lambda) = \frac{1}{2} \|\Delta W\|_C^2 + \sum_i \lambda_i \big(\langle \Delta W, a_i \rangle_C + b_i \big). \tag{14}$$

Stationarity in $\mathcal{H}_P$ gives $\Delta W(\lambda) = -\sum_i \lambda_i a_i$. Substitution yields

$$g(\lambda) = \lambda^\top b - \frac{1}{2} \lambda^\top Q \lambda, \qquad \lambda \geq 0, \tag{15}$$

which is (8). Because every $b_i > 0$, multiplying any feasible point by $1 + \epsilon$ for any $\epsilon > 0$ makes every inequality strict; Slater's condition therefore gives strong duality. Since $C_\ell \succ 0$, the primal objective is strictly convex on $\mathcal{H}_P$, so its optimizer is unique. Finally, each term $a_{i\ell} = u_{i\ell} k_{i\ell}^\top$ has rank at most one, proving (9).

C Transfer to an Unmeasured Access

Let a_{new} denote the protected direction of an access not used to construct the update. Suppose it can be decomposed as

$$a_{\text{new}} = \sum_i w_i a_i + e, \qquad w_i \geq 0, \tag{16}$$

where w_i measures its coverage by the constructed access directions and e is the uncovered residual. If an update achieves a fraction t of every constructed decrease, then Cauchy–Schwarz gives

$$\langle \Delta W, a_{\text{new}} \rangle_C \leq -t \sum_i w_i b_i + \|\Delta W\|_C \|e\|_C. \tag{17}$$

The bound predicts transfer when descent along the covered directions exceeds the contribution of the residual. It also makes coverage a geometric property of model responses rather than a consequence of surface-level textual similarity.

D Termination of Adaptive Scope Reduction

Let S_0 be the initial candidate set in Algorithm 1. Each verification round either identifies facts that fail their individual checks or, when a shared protection check fails, removes the fact with maximum individual protection cost. Hence every shrink round produces $S_{r+1} \subsetneq S_r$, and at most $|S_0|$ such rounds can occur. Because each proposal is recomputed from the same base model and committed only after all full-model forgetting and retention checks pass, any returned edit satisfies the stated verification criteria.

E Implementation Details

F Protocol and Coverage Accounting

Every method receives the same 50-forget/1,000-retain split and case-balanced evaluator. Prompt scores are first averaged within each fact, and seeds are the replication units. U2E fixes internal

Table 3: Configuration used in the MultiCounterFact experiments.

Component	Setting
Editable modules	MLP `down_proj`, layers $4, 6, \ldots, 26$
Construction prompts	dataset prompt; "Complete the following factual statement: " + dataset prompt; fixed Chinese completion prefix + dataset prompt
Output score	old-answer minus comparison-answer teacher-forced mean log probability
Comparison answer	protocol-provided neutral target <|eot_id|>, fixed before scoring
Pre-edit eligibility	output contrast ≥ 0 on at least one construction prompt
Internal positions	top 2 old-answer first-token logit-lens scores from hidden-state index 9 onward; fixed before editing
Layer choice	maximize the weakest leave-one-prompt transfer; then lower protection cost and earlier layer
Acceptance prompt	"Completa la siguiente afirmación factual en inglés: " + dataset prompt; excluded from solving
Construction ratios	output 0.25; internal 0.35
Acceptance ratios	output/Spanish prompt 0.35; internal 0.50
Public protection	32 Wikipedia texts; at most 512 keys per layer; 8 shared hard-protection examples
Fact protection	4 relation-matched facts per request
Covariance	low-rank second moment; damping 0.001
Value radius	$0.5\times$ public-output RMS
KL gates	mean ≤ 0.002; maximum ≤ 0.01
Relation-matched fact gate	maximum log-probability drop ≤ 0.1
Relinearization	at most 3 relinearization refinement rounds
Numerics	FP64 solver/checks; model-native forward dtype

readouts before editing, and its public and relation-matched protection examples are disjoint from the evaluation retain set.

Across the ten 3B runs, the mean 50-request batch contains 44.7 accepted edits, 2.0 no-ops that already meet the output criterion, and 3.3 abstentions, giving $(44.7 + 2.0)/50 = 93.4\%$ coverage. The main behavioral metrics evaluate all 50 requested facts after the final batch update; coverage therefore records the fraction of requests handled, while FE/FG/FS score the complete request set. The same accounting is used for the 8B evaluations.

G Hyperparameter Sensitivity

We perform a one-factor-at-a-time descriptive analysis on three fixed development seeds. Forgetting strength scales all four construction and audit probability-ratio constraints together; this avoids the invalid constraint ordering that would result from scaling construction thresholds alone. The KL mean/maximum budgets and the value trust-region radius are independently scaled by $0.5\times$ and $2\times$ around the canonical setting. These runs do not select or alter that setting.

The mean old-answer access F remains between 7.67% and 12.00%, while retained knowledge R spans only 18.61%–20.13% and absolute perplexity shift remains below 0.030. Committability is more sensitive: the strict $0.5\times$ KL budget lowers coverage from 66.7% to 48.0%, whereas loosening it to $2\times$ leaves the measured behavioral outputs unchanged. Halving the value radius raises coverage to 83.3%, but doubling it lowers coverage to 42.0% and triggers additional full-model audit removals or insufficient-progress exits. A $2\times$ forgetting requirement improves F from 8.67% to 7.67% at the cost of lowering R from 19.57% to 18.61%. Because $n = 3$ and seed variation is substantial, these curves establish operating behavior rather than a confirmatory robustness claim.

H Structural Consistency under Trusted Progress

We apply a common $0.5\times$ trust-radius multiplier on six fixed seeds, selected from pre-update quantities, to expose non-saturated trusted progress. Among 312 fact attempts, 274 have $0 < t^* < 1$; 272 of these satisfy the specified full-model checks, giving a 99.3% validation rate. Across 4,704 exact node solves, every numerical writeback satisfies the bound in (9) when active support is defined by strictly positive dual multipliers. The controlled intervention therefore confirms high-validity

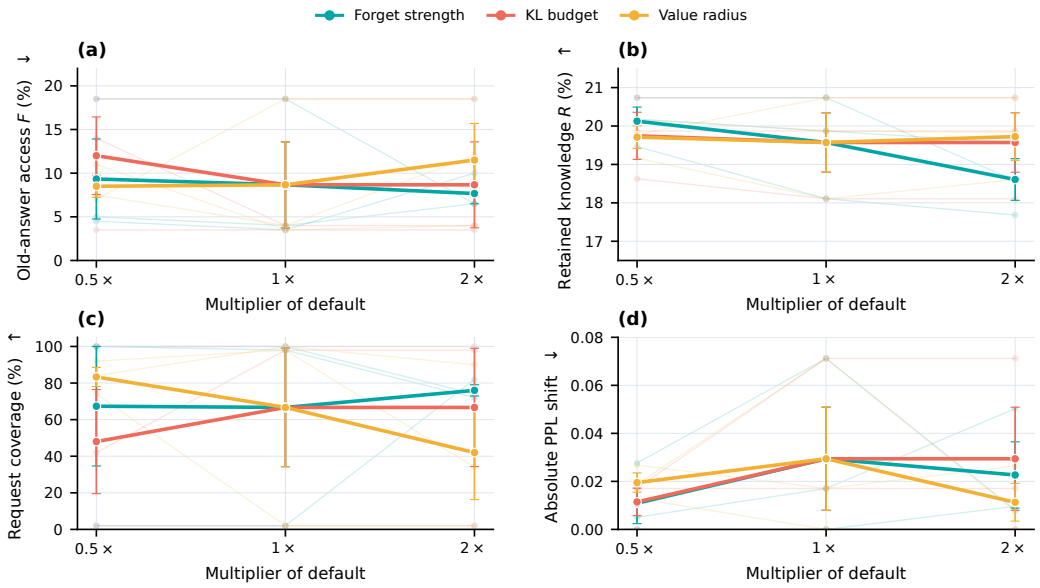

Figure 6: One-factor-at-a-time sensitivity on three development seeds. Faint lines show individual seeds; solid lines show mean ± standard error. The $1\times$ point is the same frozen default in all three curves.

candidate construction across interior trusted progress and the predicted active-support low-rank structure.

I Efficiency Profile

Table 4 reports the same-model, same-GPU profile. All methods use the same 50-fact request, and warm-start time is averaged over three runs. For U2E, the constrained solve and access construction account for 164.15 and 25.29 seconds of the 191.54-second total, respectively.

Table 4: Same-model, same-GPU warm-start profile. Success counts committed non-no-op edits.

Method	Success	Time (s)↓	s/success↓	GPU (GiB)↓
U2E	45/50	191.54	4.256	43.45
ZeroUnlearn	50/50	15.70	0.314	17.41
RMU	50/50	9.96	0.199	42.95
Mechanistic	50/50	171.80	3.436	17.54

J Paired MultiCounterFact Comparisons

Table 5: Selected paired differences for U2E versus ZeroUnlearn on MultiCounterFact. Differences follow each metric's preferred direction, so positive values favor U2E. Intervals are 20,000-replicate paired seed bootstrap CIs.

Metric	U2E	ZeroUnlearn	U2E gain [95% CI]
Forget efficacy ↓	3.000	17.600	14.600 [12.000, 17.400]
Forget generalization ↓	6.300	18.100	11.800 [8.700, 15.000]
Absolute PPL shift ↓	0.03945	0.99639	0.95694 [0.5821, 1.4210]

K PRESERVATION-CALIBRATED ROBUSTNESS

We additionally compare methods at preservation-calibrated operating points. Each comparator's setting is selected on development seeds by minimizing its maximum deviation from the original model over forget specificity and the three retention metrics; the selected setting is then fixed for six confirmation seeds. This aligns the methods by retained behavior before comparing suppression across access paths.

Table 6: Preservation-calibrated robustness over six confirmation seeds. Worst is the largest residual over the measured access families, Ext.@20 is black-box extraction after 20 attempts, and Cov. is request-level coverage. Lower is better except for coverage.

Method	Worst↓	Ext.@20↓	$\lvert\Delta\text{PPL}\rvert\downarrow$	Cov.↑
U2E	**0.654**	**0.730**	**0.041**	0.970
ZeroUnlearn	0.724	0.840	0.908	1.000
AlphaEdit	0.738	0.863	0.053	1.000
SimNPO	0.868	0.813	2.719	1.000
Mechanistic	0.726	0.817	4.237	1.000

Across the confirmation seeds, U2E obtains the lowest worst-access residual, Extraction@20, and perplexity shift. Its paired gains over the four comparators range from 0.070 to 0.214 on worst-access residual and from 0.083 to 0.133 on Extraction@20. Every paired bootstrap interval is strictly positive, with Holm-adjusted $p \leq 0.0469$. Together with 97.0% coverage, these results show that U2E combines consistent multi-route suppression, low recoverability, and small model drift at a preservation-calibrated operating point.

L SECONDARY CROSS-MODEL ACCESS AUDIT

We further evaluate six seeds on DeepSeek-R1-Distill-Qwen-1.5B, a reasoning checkpoint built on the Qwen2 architecture. Its tokenizer maps the neutral target to seven ordinary tokens, so this architecture-compatible comparison uses nine access families, black-box extraction, and base-normalized endpoints.

For U2E, all nine access-family mean log-survival values decrease, but the route-balanced change is shallow (-0.0226), Extraction@20 remains 0.4967, and only 8.67% of facts decrease on every path. ZeroUnlearn is much stronger on average (route-balanced -3.4983, Extraction@20 0.1700, and 43.33% all-path decrease), but its MCQ family increases by 0.3606 $\log_2$ units and its worst-path change reaches $+0.7669$, versus $+0.0303$ for U2E. The paired U2E advantage on worst-path amplification is 0.7366 (95% CI $[0.6017, 0.8921]$, one-sided exact $p = 0.015625$). This audit identifies an architecture-dependent tradeoff: U2E produces less worst-path amplification, while ZeroUnlearn achieves stronger average suppression and lower extraction success.

M BROADER IMPACT

Targeted factual unlearning could help correct obsolete information and respond to legitimate deletion requests without retraining a full model. The same capability can also be used to conceal facts, manipulate historical records, or create misleading claims of compliance. The method's abstention records and validation results should therefore be retained, and deployment claims should distinguish behavioral suppression from certified removal of training-data influence.