REXMAP: COUNTERFACTUAL RECOVERY FOR MODEL-EDIT ROLLBACK AFTER COMPOSITION AND SERVING DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

A revoked edit can disappear from the live weights while its influence survives inside every later update optimized in its presence. RExMap recovers the corresponding counterfactual history: its ledger unwinds the retained suffix, removes the revoked update, re-estimates each retained request on the corrected state, and evaluates the recovered model on the edited fact, 500 clean prompts, 200 attacked probes, and an off-target entity neighborhood. In 24,000 Llama-3.1-8B transactions, exact all-conditions restoration falls from 97.4% at depth 1 to 82.8% at depth 20 in FP32, 68.7% under INT8, and 60.9% with dual overlap and non-LIFO revocation. Re-estimating retained descendants recovers 12.0 points over direct subtraction at depth 20, identifying state-dependent descendants as a correctable source of rollback error. A transaction-state score predicts behavioral validity with 0.914 AUROC and a 2.7% false-valid rate; routing uncertain cases through checkpoint replay raises restoration by 22.0 points over uniform subtraction to 97.8%, with a 60.1% direct fast path and 0.31 s median rollback. The depth and serving-representation ordering recurs across a reversible ROME-style writer, Mistral-7B-v0.3, and MQuAKE. RExMap therefore turns reversibility from a write-time algebraic property into a behaviorally testable recovery decision for evolving edited models.

1 INTRODUCTION

Consider a factual update e_k that is correct when committed, followed by updates whose factors are optimized on a model already containing e_k. Revoking e_k later should preserve those descendants as if they had instead been generated from the history without e_k. Direct subtraction preserves their original factors and therefore returns a different computation. Quantization, adapter merging, update-subspace overlap, and non-LIFO revocation can widen that difference further. The object that must be recovered is thus a behavioral counterfactual over an edit history, not a parameter delta in isolation.

This distinction changes observed reliability by tens of points. Across 24,000 matched rollback transactions, exact FP32 restoration declines from 97.4% at depth 1 to 82.8% at depth 20 and 67.3% at depth 50. At depth 20, INT8 yields 68.7%, while dual update/entity overlap with non-LIFO revocation yields 60.9%. Edited-fact and attacked answers fail before clean protected outputs, exposing failures that a touched-tensor norm does not identify.

RExMap reconstructs and measures the required history. The transaction ledger records each request and inverse factor together with commit order, update-subspace and entity dependencies, serving representation, and checkpoint identity. To revoke e_k, ledger replay reverses the retained suffix, removes e_k, and reruns the frozen writer for every retained request on the corrected intermediate state. The matched reference independently starts from the pre-e_k checkpoint and executes the same retained history without e_k. Their behavioral agreement defines a rollback-validity surface over composition depth, overlap, undo order, and serving mode. At depth 20, descendant re-estimation raises all-conditions restoration by 12.0 points over direct subtraction (95% CI [9.1,14.7]).

RExMap also converts this measurement into an online recovery decision. A calibrated score combines observable history, dependency, serving, checkpoint, and live-representation features. On 12,000 chain-disjoint events, it reaches 0.914 AUROC and a 2.7% false-valid rate, 5.2 points below

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.

tensor-only routing. The resulting subtraction-or-checkpoint policy improves restoration by 22.0 points over uniform subtraction and reaches 97.8% restoration with a 0.31 s median. All six tested model–writer–workload regimes preserve the depth ordering while locating their checkpoint crossing at different depths.

The contribution hierarchy follows this recovery path:

1. a matched retained-history endpoint and execution-complete replay operator that distinguish algebraic deletion from behavioral recovery while preserving every non-revoked descendant;
2. a 24,000-transaction controlled map that resolves how composition, dependency overlap, undo order, and serving representation move rollback reliability; and
3. a calibrated state score and hybrid recovery rule that turn the measured surface into a low-false-valid, cost-aware subtraction-versus-checkpoint decision.

2 Behavioral Rollback as a Retained-History Counterfactual

ROME, MEMIT, MEND, SERAC, Editable Neural Networks, in-context editing, AlphaEdit, and constrained updates optimize the forward write and measure efficacy, generalization, and post-write locality (Sinitsin et al., 2020; De Cao et al., 2021; Mitchell et al., 2022a;b; Meng et al., 2022; 2023; Zheng et al., 2023; Fang et al., 2025; Deng et al., 2025; Li et al., 2025b). EasyEdit and broader studies standardize these endpoints (Yao et al., 2023; Wang et al., 2024b; Zhang et al., 2024). RExMap extends the evaluation trajectory beyond the write: it revokes one transaction after subsequent state-dependent writes and serving transformations, then compares the result with the history in which that transaction was never committed.

Lifelong and modular editors reduce sequential interference by isolating, retaining, or routing updates (Jang et al., 2022; Hartvigsen et al., 2023; Li et al., 2025a; Wang et al., 2025; Luo et al., 2026). Semantic-routing revocation removes a routing key and therefore avoids reconstructing additive descendants. Tracing-and-reversal methods instead recover or neutralize isolated rank-one edits from modified weights, while reversal tokens counter in-context edits (Youssef et al., 2026; 2025). RExMap addresses the complementary additive-history problem: the revoked edit is absent, every retained descendant is regenerated, non-LIFO order is explicit, both candidate histories receive the same serving transformation, and the measured state selects an executable recovery action. This endpoint aligns the closest reversal mechanisms by asking whether the post-undo model realizes the same retained transaction history.

Behavioral editing suites reveal propagation, reasoning, and locality effects after writing (Cohen et al., 2024; Gu et al., 2024; Zhong et al., 2023; Cheng et al., 2023; Wang et al., 2024c;a; Wei et al., 2025; Li et al., 2025c). RExMap carries the behavior-first principle through revocation and connects it to low-rank adaptation, task arithmetic, quantization, and model merging (Hu et al., 2022; Ilharco et al., 2023; Frantar et al., 2023; Liu et al., 2025). Its technical object is a serving-conditioned map from the current transaction history to post-undo behavior and the recovery operation that restores that behavior.

3 RExMap: Counterfactual Rollback over Transaction Histories

RExMap represents editing as a sequence of state-dependent transactions. Edit request q_j produces a reversible low-rank update

$$\Delta_j(\theta_{j-1}) = U_j \Sigma_j V_j^\top, \qquad \theta_j = \theta_{j-1} + \Delta_j(\theta_{j-1}), \quad r \le 16. \tag{1}$$

The ledger stores q_j, FP64 factors, commit position, touched layers, entity neighborhood, serving signature, and checkpoint hash. Let P_m pack effective FP32 weights into serving mode m (FP32, BF16, INT8/GPTQ, or merged adapter), and let U_m materialize a deployed state back into effective FP32 tensors. For merged adapters, P_m performs the merge exactly once and U_m returns the resulting effective weights. The live deployed state is $w_t = P_m(\theta_t)$.

The ledger DAG $\mathcal{L} = (E, \prec)$ records chronological and explicit dependencies. For edits sharing a touched layer, update-subspace overlap is the maximum cosine between their orthonormalized row or

column bases; a calibration-frozen threshold assigns parameter overlap. Entity overlap is exact after alias normalization. Thus "none," parameter-only, entity-only, and dual overlap remain distinct even though the primary writer draws touched layers from the common range 17–25.

For target transaction e_k, let $R_k = (e_{k+1}, \ldots, e_t)$ be the retained suffix in commit order. All non-checkpoint operators begin from the same live representation $U_m(w_t)$ and are repacked once after recovery:

$$\begin{aligned}
\mathsf{Direct}_{k,m}(w_t) &= P_m(U_m(w_t) - \Delta_k)\,, \\
\mathsf{Fixed}_{k,m}(w_t) &= P_m\left(U_m(w_t) - \overleftarrow{\sum_{j\in R_k}} \Delta_j - \Delta_k + \sum_{j\in R_k} \Delta_j\right), \\
\mathsf{Ledger}_{k,m}(w_t) &= P_m\left(\mathrm{Replay}\left(U_m(w_t) - \overleftarrow{\sum_{j\in R_k}} \Delta_j - \Delta_k;\ R_k\right)\right).
\end{aligned} \tag{2}$$

Here each unindexed Δ_j denotes its stored factor. Direct subtraction leaves every later state-dependent factor fixed. Fixed replay reverses the suffix, removes e_k, and reapplies the same stored suffix factors; its additive FP32 computation equals direct subtraction up to operation order and isolates numerical bookkeeping. Ledger replay replaces those fixed suffix factors with $\Delta'_j = \mathrm{Writer}(q_j, \theta'_{j-1})$, estimated sequentially on the corrected state. Checkpoint recovery starts from the exact pre-e_k checkpoint and performs the same ordered re-estimation. LIFO rollback has an empty suffix; non-LIFO rollback preserves every retained transaction. The distinction remains active in the no-explicit-overlap slice because a low measured subspace cosine does not remove the writer's dependence on the full intermediate state.

The matched no-edit reference $\theta_{\mathrm{ref}}(e_k, c)$ starts from the pre-e_k checkpoint, omits e_k, and replays R_k with the frozen writer. Candidate and reference are both packed with P_m before inference. Their matched-history touched-tensor error is

$$E_{\mathrm{hist}} = 100\frac{\|U_m(w_{\mathrm{rb}}) - U_m(P_m(\theta_{\mathrm{ref}}))\|_F}{\|U_m(P_m(\theta_{\mathrm{ref}}))\|_F}, \tag{3}$$

with both norms restricted to chain-touched tensors. A separate round-trip diagnostic, E_{chain}, compares a deterministic rollback candidate with its immediate pre-chain FP32 checkpoint; it appears only in Figure 1 and Appendix Table 6, never as the behavioral endpoint or a routing feature. Four indicators compare recovered behavior with the matched history: B_{fact} for the edited fact and retained descendants, B_{clean} over 500 clean prompts, B_{attack} over 200 attacked probes, and B_{neigh} over the off-target entity neighborhood. Transaction success is

$$S_{\mathrm{rb}} = \mathbb{1}[B_{\mathrm{fact}} \wedge B_{\mathrm{clean}} \wedge B_{\mathrm{attack}} \wedge B_{\mathrm{neigh}}]. \tag{4}$$

Deterministic decoding defines the answer comparisons; B_{fact} also requires the retained descendants to remain forward-valid, and B_{neigh} requires a post-undo change rate below the threshold selected on disjoint chains. Checkpoint recovery can therefore score below 100% even after exact state reload because replayed descendant edits must still satisfy these behavioral criteria.

Definition 1 (Rollback-validity surface)**.** *Let the live rollback context be $c = (d, o, r, m, h)$, containing chain depth d, update/entity overlap class o, rollback order r, serving mode m, and checkpoint/serving signature state h. RExMap estimates*

$$\mathcal{V}_{\mathcal{C}}(c) = \Pr(S_{\mathrm{rb}} = 1 \mid c, \mathcal{C}) \tag{5}$$

over independent transaction draws. The calibrated expiry boundary separates tested contexts that authorize direct subtraction from contexts routed to checkpoint recovery.

Continuous state score. Routing uses $p_k = \sigma(\beta^\top \phi_k)$, where ϕ_k contains log depth, retained-suffix length, update/entity overlap class, undo order, serving and checkpoint mismatch indicators, dependency counts, and a live-representation sketch. For fixed random projection Π and ledger sketch $s_t = \Pi\theta_t$, the last feature compares $\Pi(U_m(w_t) - \Delta_k)$ with the additive canonical prediction $s_t - \Pi\Delta_k$. It is available from the live state and ledger without loading a checkpoint or constructing θ_{ref}; E_{hist} remains evaluation-only. Chain-disjoint predictor data are split into 6,000 fitting, 6,000 threshold-calibration, and 12,000 evaluation events. Logistic loss fits β; the calibration split selects the largest fast-path region whose false-valid rate is at most 3%. Model, writer, and serving-stack transfers refit the score and threshold on their calibration chains. The continuous p_k produces AUROC/AUPRC, while the frozen threshold produces `valid`/`stale` actions.

Optional proposal admission. Before a transaction enters the ledger, the pre-commit score $\delta(e) = \delta_{\text{margin}}(e) + \delta_{\text{search}}(e)$ combines clean-margin fragility with observed protected-prompt flips. The operating point admits $\delta(e) \leq 5.0\%$. This screen changes which edits enter service; it is separate from the post-commit score p_k, which chooses how an admitted edit is revoked. Appendix D gives the complete equations and threshold sweep.

4 Instrumentation and Online Recovery

RExMap accepts a reversible writer, protected distribution, probe generator, chain schedule, serving modes, and recovery policies. The primary instantiation uses constrained MEMIT; transfer applies the same transaction observables and state logic to a reversible ROME-style writer.

Independent behavior probes. For each edit, the 200-probe admission audit contains 50 paraphrases, 30 multi-turn conversations, 40 entity-linked queries, and 80 decomposition attacks. Templates, relation instances, and generator seeds are shared across admission methods and fixed before threshold selection. The post-commit pool replaces all four families with unseen paraphrase styles, long-context distractors, multi-hop decompositions, and cross-turn entity substitutions. Budgets of 50, 100, 200, 400, and 800 measure how ranking, calibration, protected flips, and post-undo failures evolve with probe strength.

Commit and rollback path. An admitted request, its factors, canonical sketch, serving signature, checkpoint identity, and dependency edges are recorded atomically. At rollback, the frozen state score first evaluates the live direct-subtraction candidate. A `valid` decision returns that candidate; a `stale` decision loads the pre-e_k checkpoint and regenerates retained descendants in commit order. Appendix Algorithm 1 specifies both branches and the ledger-replay evaluation operator.

5 Experimental Setup

The experiments follow four linked questions: how composition and serving representation reshape behavioral rollback; how much descendant re-estimation recovers over fixed-factor subtraction; whether observable transaction state predicts a safe fast path; and whether state-routed recovery preserves the ordering across editing stacks while improving the reliability–cost frontier.

5.1 Benchmarks, Models, and Writers

CounterFact (Meng et al., 2022) supplies 2,000 proposals and 1,000 protected facts for the controlled study. MQuAKE (Zhong et al., 2023) contributes 500 multi-hop edits for transfer. The primary model is Llama-3.1-8B (Grattafiori et al., 2024), edited by constrained MEMIT-style rank-16 updates whose touched layers lie in 17–25. Transfer crosses this writer with a reversible ROME-style rank-one update and crosses both writers with the documented Mistral-7B-v0.3 checkpoint (Mistral AI, 2024). Protected prompts are selected before editing from fixed relation templates and entity neighborhoods.

5.2 Rollback Protocol and Uncertainty

The primary mode-by-depth grid contains 200 independently drawn chains per seed at depths 1, 5, 10, 20, 35, and 50, for five seeds and 1,000 target rollback transactions per point. Four serving modes—FP32, BF16, INT8/GPTQ, and merged adapter—produce 24,000 transactions. A paired depth-20 FP32 slice varies no, parameter-only, entity-only, and dual overlap together with LIFO and non-LIFO undo, again using 1,000 transactions per condition. Each transaction evaluates the edited fact, 500 clean prompts, 200 attacked probes, and 50 off-target prompts.

Every edit chain is generated in a canonical FP32 transaction state and packed once into its serving mode. Direct, fixed-factor, and ledger operators materialize the same deployed state with U_m, perform their updates in FP32, and repack once with P_m; checkpoint recovery starts from the exact canonical checkpoint and then uses the same P_m. The matched no-edit history is independently reconstructed and packed with the identical serving signature. Reported mode effects are therefore serving-conditioned rollback outcomes, not estimates of transformation error in isolation. Table 8's

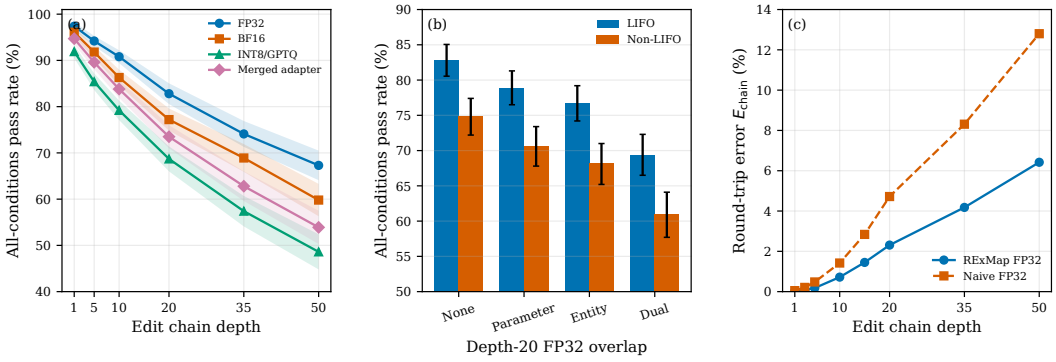

Figure 1: Behavioral rollback-validity surface on CounterFact with Llama-3.1-8B and the constrained rank-16 writer. Left: mean all-conditions success across six depths and four serving modes. Center: depth-20 FP32 success for LIFO and non-LIFO rollback under no, parameter-only, entity-only, and dual overlap. Right: the separate round-trip diagnostic E_{chain} for ledger and direct FP32 rollback against the immediate pre-chain checkpoint. Behavioral points contain 1,000 transactions over five seeds.

0.1% value is single-edit commit-time error; the approximately 0.02% depth-1 value and Appendix Table 6's depth-5 value of 0.18% are E_{chain} round-trip diagnostics against the immediate pre-chain checkpoint. E_{hist} instead compares a target rollback with its retained-history counterfactual.

Mode–depth tables report five-seed means and raw transaction counts; paired seed-level bootstrap intervals accompany the two primary contrasts. Deterministic tensor sweeps are labeled explicitly. Admission rates use all 2,000 proposals; commit-conditional metrics use the 1,924 accepted proposals at the main operating point. State quality uses chain-disjoint fitting, calibration, and evaluation sets.

5.3 Comparators

Rollback comparisons use RExMap ledger replay, direct target-delta subtraction, the algebraically equivalent frozen-factor replay null, and checkpoint restoration on identical chains. State baselines use the live tensor sketch, depth, parameter overlap, entity overlap, both overlaps, or serving mode alone, plus always-checkpoint recovery. Admission comparisons hold the rank-16 writer, 2,000 proposals, protected set, and probe budget fixed while varying always commit, acceptance-matched random, search-only, margin-only, combined, and full-policy rules.

Constrained low-rank editing is the matched forward comparator because it shares the writer, rank, layers, and edit regime; MEMIT and ROME provide established efficacy anchors. Every recovery contrast reuses the same target transactions, so the direct–fixed–ledger–checkpoint sequence changes one recovery operation at a time.

6 Results

6.1 Composition and Serving State Move the Rollback Boundary

Figure 1 resolves the validity surface first over depth and serving mode, then over overlap and rollback order. Each mode–depth point averages 1,000 paired transactions over five independent seeds.

The surface contracts along both composition and representation axes. FP32 declines from 97.4% to 82.8% by depth 20 (paired 95% CI [80.5,85.0]) and 67.3% by depth 50. At depth 20, INT8 and merged-adapter serving trail FP32 by 14.1 and 9.3 points. An edit that supports direct rollback in a shallow FP32 history can therefore require checkpoint recovery after either further composition or a serving-representation change.

Table 1: Transaction-level all-conditions pass rate (%) by serving mode and chain depth. Every cell is a five-seed mean over 1,000 paired rollback transactions; raw pass/fail counts appear in Appendix Table 13.

Mode	1	5	10	20	35	50
FP32	97.4%	94.2%	90.8%	82.8%	74.1%	67.3%
BF16	96.1%	91.8%	86.3%	77.2%	68.9%	59.8%
INT8/GPTQ	91.9%	85.4%	79.2%	68.7%	57.4%	48.6%
Merged adapter	94.7%	89.6%	83.8%	73.5%	62.8%	53.9%

Table 2: Post-undo behavioral restoration (%) at representative conditions. Component rates and their transaction-level conjunction use the same transactions. Ledger and direct rows use identical chains; checkpoint includes retained-update replay.

Policy	Condition	Fact	Clean	Attack	Neigh.	All pass
Ledger	FP32, depth 20	94.6%	97.3%	94.0%	95.5%	82.8%
Ledger	INT8, depth 20	89.7%	93.1%	88.6%	91.3%	68.7%
Ledger	Dual, non-LIFO, depth 20	87.8%	91.9%	85.4%	88.1%	60.9%
Direct	FP32, depth 20	88.9%	93.5%	86.8%	90.6%	70.8%
Checkpoint	FP32, depth 20	99.4%	99.8%	99.3%	99.7%	98.3%

6.2 Re-estimating Descendants Recovers the Missing History

At depth 20, ledger replay improves all-conditions success over direct subtraction by 12.0 points (paired 95% CI [9.1,14.7]). Frozen-factor replay is algebraically identical to direct subtraction, so the gain comes from rerunning the writer on corrected intermediate states rather than from recording and reordering factors. Attacked and edited-fact answers expose this difference before clean outputs. Parameter-only and entity-only overlap yield 78.9% and 76.7%; their combination yields 69.4%, and non-LIFO dual overlap reaches 60.9%.

6.3 Transaction State Predicts a Reliable Fast Path

The continuous score and its 3%-calibrated threshold label 7,214 of 12,000 chain-disjoint events valid and 4,786 stale. Table 3 reports both discrimination and the decision errors induced by this operating point.

The joint score reaches 0.914 AUROC and 0.953 AUPRC. Although each single-coordinate comparator authorizes a similar or larger fast share, the joint score's 2.7% false-valid rate is substantially below 7.9% for the sketch-only score, 10.8% for depth alone, and 12.4% for mode alone. The full score authorizes direct subtraction for 60.1% of events; Appendix Table 12 provides all feature ablations, false-stale rates, and fast shares.

6.4 State-Routed Recovery Improves the Reliability–Cost Frontier

Recovery policies are evaluated on the same independent cohort drawn from the default depth, overlap, order, and mode mixture; the denominator is every requested rollback. The predictor test in Table 3 is a separate chain-disjoint population. Latency uses batch size 1, FP32 factors, 20 warm-ups, and 200 repetitions on 8×A100-80GB with local NVMe checkpoints.

Routing improves restoration by 22.0 points over uniform subtraction and finishes within 1.4 points of always checkpoint. Its 60.1% fast share produces a 0.31 s median and 152.4 successful recoveries per minute. The hybrid retains the same 30.5 GB shared checkpoint capacity as always checkpoint and adds 21.6 MB of factors, sketches, and metadata per edit.

Table 3: Rollback-state prediction against transaction-level behavior. Counts cover 12,000 evaluation transactions; rates in the lower block use the row-specific decision denominator. False-valid is a failed rollback labeled valid, while false-stale is a successful rollback routed to checkpoint recovery.

Predicted state	Behavior pass	Behavior fail
Valid	7,019	195
Stale	1,289	3,497
Metric	**Value**	
Precision / recall	97.3% / 84.5%	
Specificity / NPV	94.7% / 73.1%	
AUROC / AUPRC	0.914 / 0.953	

Table 4: Recovery reliability and like-for-like cost on the common default-mixture cohort. Timing uses the separate 200-repetition protocol. Shared checkpoint capacity is per snapshot; per-edit storage is incremental.

Policy	Success	p50	p95	Shared ckpt.	Per edit	Successes/min
Direct subtraction	75.8%	0.18 s	0.27 s	–	21.6 MB	260.4
Always checkpoint	99.2%	7.11 s	7.86 s	30.5 GB	–	8.2
RExMap hybrid	97.8%	0.31 s	7.43 s	30.5 GB	21.6 MB	152.4

6.5 Round-Trip Tensor Error Does Not Determine Behavior

The separate E_{chain} round-trip diagnostic follows the coarse behavioral ordering: FP32 stays below 0.5% through depth 5 and reaches 2.31% at depth 20, where BF16, INT8, and merged adapters reach 3.12%, 3.81%, and 3.48%. Yet Table 2 shows why it cannot define rollback validity: similar tensor distances can restore different edited-fact, attacked, and neighborhood answers. Appendix Table 6 reports the complete deterministic sweep; the online score uses the distinct live-representation sketch defined in Section 3.

6.6 Pre-Commit Screening Complements Post-Commit Recovery

The matched admission comparison holds the writer, proposals, protected set, and 200-probe budget fixed; proposal-level target success counts refusals as unsuccessful, while protected flips count them as zero. At 96.2% acceptance, the combined score lowers proposal-level protected flips by 1.2 points relative to matched random refusal (95% CI [0.8,1.6]) and by 0.4 points relative to search-only refusal, while target success remains within 0.2 points of search only. Appendix Table 9 reports the complete matched frontier. This pre-commit result controls entry into the transaction history; the rollback surface and routing results address recovery after entry.

6.7 Held-Out Probe Budgets Preserve Risk Ordering

The independently generated pool excludes admission templates, generator seeds, and relation instances. As the audit grows from 50 to 800 probes, detected protected flips rise from 3.7 to 7.8 and rollback failure from 6.9 to 10.2; AUROC reaches 0.88 and ECE reaches 0.053. Appendix Table 10 shows the complete budget sweep. This experiment establishes held-out ranking and calibration across probe budgets, complementing the matched operating-point comparison rather than re-estimating its frozen 5% decision.

6.8 The Expiry Ordering Recurs across Editing Stacks

All six model–writer–workload regimes in Appendix Table 7 exhibit finite expiry with stack-specific checkpoint crossings. At depth 20, restoration is 87.1 for the Llama ROME-style writer, 83.4 for its Mistral counterpart, and 76.8 on MQuAKE. The descriptive five-seed means preserve the depth

ordering while placing majority checkpoint routing at depth 10 or 20, demonstrating why the boundary is calibrated per editing stack.

6.9 Ablations Expose the Efficacy–Reversibility Tradeoff

On a matched operating slice, increasing writer rank from 4 to 16 raises target success from 89.7% to 94.8% while restoration moves from 94.8% to 89.3%. The state-feature ablations show a parallel systems tradeoff: single-coordinate rules authorize similar or larger fast shares but incur 7.9%–12.4% false-valid rates, versus 2.7% for the joint score. For proposal admission, tightening $\delta_{\max}$ from 7.0% to 3.0% moves acceptance from 98.8% to 91.0% and protected flips from 4.8% to 3.4%. These ablations locate the operating point jointly in writer capacity, state evidence, and admission strictness.

7 Implications for Editable-Model Evaluation and Recovery

RExMap separates four layers that should be reported independently: executable subtraction, round-trip tensor drift, retained-history behavior, and recovery choice. Their divergence is operationally consequential: the same stored factor can support direct rollback in one history and require checkpoint replay after more composition or a serving-state change.

The evaluated regime spans two models, two additive low-rank writers, CounterFact and MQuAKE, depths through 50, four serving modes, four overlap classes, and two undo orders. The hybrid's 60.1% fast share explains its 0.31 s median, while its 26.9% false-stale rate and 7.43 s p95 show that conservative checkpoint routing still determines the tail. Its 2.7% decision error concentrates most strongly in INT8, dual-overlap, and depth-50 slices. These measured tradeoffs support per-stack calibration for additive reversible writers and make the retained-history endpoint directly applicable when evaluating other revocation mechanisms.

8 Conclusion

RExMap establishes that an inverse factor is not a permanent rollback guarantee: composed descendants and serving transformations change the history that must be recovered. Re-estimating descendants gains 12.0 restoration points over direct subtraction at depth 20, and state-aware checkpoint routing raises end-to-end restoration by 22.0 points while retaining a 60.1% direct fast path. This turns rollback from a write-time declaration into a behaviorally verified, history-conditioned recovery capability for editable models.

Ethics Statement

The experiments use factual editing benchmarks and model-generated attack probes without human subjects. For safety-sensitive factual patches, the evaluated operating policy combines bounded pre-commit screening, explicit human authorization, behavior-validated rollback state, and checkpoint recovery.

Reproducibility Statement

Sections 3–5 specify transaction inputs, observables, audit composition, model and writer settings, chain construction, serving modes, comparators, uncertainty, and hardware. The appendices provide the execution order, threshold sweeps, raw-count summaries, and secondary diagnostics used to reproduce every reported comparison.

AI Use Statement

Generative AI tools assisted with drafting, copyediting, and presentation of manuscript text, including revisions to the title, abstract, and introduction. The available manuscript-preparation record does not document additional AI use in developing the research questions, methods, experiments, analyses, or interpretations. The authors reviewed all AI-assisted material and take responsibility for the submitted manuscript.

REFERENCES

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? In *EMNLP*, pp. 13877–13888, 2023. doi: 10.18653/v1/2023.emnlp-main.856. URL https://aclanthology.org/2023.emnlp-main.856/.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. In *Transactions of the Association for Computational Linguistics*, volume 12, pp. 283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://aclanthology.org/2024.tacl-1.16/.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *EMNLP*, pp. 6491–6506, 2021. doi: 10.18653/v1/2021.emnlp-main.522. URL https://aclanthology.org/2021.emnlp-main.522/.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. Everything is editable: Extend knowledge editing to unstructured data in large language models. In *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/02763667a5761ff92bb15d8751bcd223-Paper-Conference.pdf.

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 *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/29c8c615b3187ee995029284702d3f43-Paper-Conference.pdf.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *ICLR*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783v3.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhenhua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16801–16819, 2024. doi: 10.18653/v1/2024.emnlp-main.934. URL https://aclanthology.org/2024.emnlp-main.934/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *NeurIPS*, volume 36, pp. 47934–47959, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/95b6e2ff961580e03c0a662a63a71812-Abstract-Conference.html.

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 *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=vfsRB5MImo9.

Jiaang Li, Quan Wang, Zhongnan Wang, Yongdong Zhang, and Zhendong Mao. ELDER: Enhancing lifelong model editing with mixture-of-LoRA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24440–24448, 2025a. doi: 10.1609/aaai.v39i23.34622. URL https://ojs.aaai.org/index.php/AAAI/article/view/34622.

Yanhong Li, Min Yang, Xiping Hu, and Chengming Li. Forget for get: A lightweight two-phase gradient method for knowledge editing in large language models. In *EMNLP Findings*, pp. 7604–7623, 2025b. doi: 10.18653/v1/2025.findings-emnlp.402. URL https://aclanthology.org/2025.findings-emnlp.402/.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. Unveiling the pitfalls of knowledge editing for large language models. In *ICLR*, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/56f187f2b3e87da2fb103af6df611970-Abstract-Conference.html.

Shuqi Liu, Yuxuan Yao, Bowei He, Zehua Liu, Xiongwei Han, Mingxuan Yuan, Han Wu, and Linqi Song. 1bit-merging: Dynamic quantized merging for large language models, 2025. URL https://arxiv.org/abs/2502.10743v2.

Haihua Luo, Xuming Ran, Tommi Kärkkäinen, Zhonghua Chen, Jiangrong Shen, Qi Xu, and Fengyu Cong. Reversible lifelong model editing via semantic routing-based LoRA, 2026. URL https://arxiv.org/abs/2603.11239.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, volume 35, pp. 17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *ICLR*, 2023. URL https://openreview.net/forum?id=MAljzdU9iC.

Mistral AI. Model card for mistral-7b-v0.3. Hugging Face model card, 2024. URL https://huggingface.co/mistralai/Mistral-7B-v0.3.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *ICLR*, 2022a. URL https://openreview.net/forum?id=0DcZxeWfOPt.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *ICML*, volume 162, pp. 15817–15831, 2022b. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry V Pyrkin, Sergei Popov, and Artem Babenko. Editable neural networks. In *ICLR*, 2020. URL https://openreview.net/forum?id=HJedXaEtvS.

Ke Wang, Yiming Qin, Nikolaos Dimitriadis, Alessandro Favero, and Pascal Frossard. MEMOIR: Lifelong model editing with minimal overwrite and informed retention for LLMs. In *NeurIPS*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/809a560bf60984be05180c071c352eaa-Abstract-Conference.html.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. Detoxifying large language models via knowledge editing. In *ACL*, pp. 3093–3118, 2024a. doi: 10.18653/v1/2024.acl-long.171. URL https://aclanthology.org/2024.acl-long.171/.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *ACL*, pp. 82–93, 2024b. doi: 10.18653/v1/2024.acl-demos.9. URL https://aclanthology.org/2024.acl-demos.9/.

Xiaohan Wang, Shengyu Mao, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, Huajun Chen, and Ningyu Zhang. Editing conceptual knowledge for large language models. In *EMNLP Findings*, pp. 706–724, 2024c. doi: 10.18653/v1/2024.findings-emnlp.40. URL https://aclanthology.org/2024.findings-emnlp.40/.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. MLaKE: Multilingual knowledge editing benchmark for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4457–4473, 2025. URL https://aclanthology.org/2025.coling-main.301/.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In *EMNLP*, pp. 10222–10240, 2023. doi: 10.18653/v1/2023.emnlp-main.632. URL https://aclanthology.org/2023.emnlp-main.632/.

Paul Youssef, Zhixue Zhao, Jörg Schlötterer, and Christin Seifert. How to make LLMs forget: On reversing in-context knowledge edits. In *NAACL*, pp. 12656–12669, 2025. doi: 10.18653/v1/2025.naacl-long.630. URL https://aclanthology.org/2025.naacl-long.630/.

Paul Youssef, Zhixue Zhao, Christin Seifert, and Jörg Schlötterer. Tracing and reversing edits in LLMs. In *ICLR*, 2026. URL https://openreview.net/forum?id=AiT8F6pbfi.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. A comprehensive study of knowledge editing for large language models, 2024. URL https://arxiv.org/abs/2401.01286v5.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *EMNLP*, pp. 4862–4876, 2023. doi: 10.18653/v1/2023.emnlp-main.296. URL https://aclanthology.org/2023.emnlp-main.296/.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *EMNLP*, pp. 15686–15702, 2023. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

A Experimental and Implementation Details

The 380 A100-GPU-hour evaluation ran on eight 80-GB A100s with CUDA 12.4 and PyTorch 2.3.1. Constrained MEMIT uses rank $r = 16$ with touched layers in 17–25 of Llama-3.1-8B. The paraphrase generator is Llama-3.1-8B with diverse decoding; entity-linked, multi-turn, and decomposition probes use fixed templates and relation graphs. Each transaction stores the edit request, FP64 factors, touched layers, update-subspace and entity metadata, a canonical live-state sketch, and $(\mathrm{dtype}, \mathrm{quantizer}, \mathrm{adapter_mode}, \mathrm{checkpoint_hash})$. INT8 uses the GPTQ implementation of OPTQ (Frantar et al., 2023). Every serving sweep uses identical chains and the P_m/U_m order defined in Section 3.

Table 5: Transaction objects, meanings, and behaviorally tested invalidation rules.

Object	Meaning in this paper	Remains valid while	Invalidated by
Protected-risk estimate	Admission screening summary under the fixed protected set and 200-probe audit	Audit inputs and writer family match	Recomputed for a new protected set, budget, or writer
Commit decision	Proposal-level admit/refuse action at $\delta_{\max}$	Model state and policy match the audited proposal	Re-evaluated before a changed proposal is committed
Rollback state valid	Direct subtraction lies inside the S_{rb}-calibrated region	Dependencies, overlap, depth, signatures, and the live sketch support the fast path	The frozen score crosses its threshold
Rollback state stale	Checkpoint recovery and ordered replay are selected	The boundary-crossing evidence remains active	Recalibration places a later context inside the tested region

Table 6: Round-trip error E_{chain} (%) across deployment modes and chain depths. The deterministic sweep compares the rolled-back candidate with its immediate pre-chain FP32 checkpoint. This diagnostic is distinct from both the matched-history error E_{hist} and the post-undo behavior in Table 1.

Mode	5 edits	10 edits	20 edits	35 edits	50 edits
FP32	0.18	0.72	2.31	4.18	6.42
BF16	0.31	1.14	3.12	5.41	8.17
INT8 (GPTQ)	0.91	1.84	3.81	6.73	10.24
Merged adapter	0.42	1.28	3.48	5.92	9.01

Table 7: Transfer across the evaluated additive low-rank stacks. Forward and behavior are percentages; behavior is transaction-level rollback success at depth 20. Fallback is the first tested depth at which checkpoint routing becomes the majority. Entries are descriptive five-seed means.

Model	Writer	Data	Forward	Behavior	AUROC	Fallback
Llama-3.1-8B	Rank-16	CF	94.8	82.8	0.914	20
Llama-3.1-8B	ROME-style	CF	96.3	87.1	0.887	20
Mistral-7B-v0.3	Rank-16	CF	92.6	85.8	0.892	20
Mistral-7B-v0.3	ROME-style	CF	94.1	83.4	0.864	10
Llama-3.1-8B	Rank-16	MQuAKE	78.6	81.2	0.879	10
Mistral-7B-v0.3	ROME-style	MQuAKE	75.3	76.8	0.842	10

B Online Transaction-State Update

At commit, screening precedes the write and the request, factors, sketch, signatures, and dependency edges are recorded atomically. Algorithm 1 specifies target selection, fast routing, checkpoint recovery, and the offline ledger comparator.

Algorithm 1 RExMap rollback operators and state-routed recovery

Require: Deployed state w_t, target e_k, ledger $\mathcal{L}$, live mode m, fitted (β, τ)
Select retained suffix $R_k = (e_{k+1}, \ldots, e_t)$ from commit order
$\theta_{\text{live}} \leftarrow U_m(w_t)$; $\theta_{\text{direct}} \leftarrow \theta_{\text{live}} - \Delta_k$
Compute live-sketch discrepancy; set $p_k \leftarrow \sigma(\beta^\top \phi_k)$
if $p_k \geq \tau$ and checkpoint/serving signatures satisfy the calibrated rule **then**
  $w_{\text{rb}} \leftarrow P_m(\theta_{\text{direct}})$; action $\leftarrow$ `direct`
else
  $\theta_{\text{rb}} \leftarrow$ load exact pre-e_k FP32 checkpoint
  for e_j in R_k in commit order **do**
    Re-estimate $\Delta'_j \leftarrow \text{Writer}(q_j, \theta_{\text{rb}})$; $\theta_{\text{rb}} \leftarrow \theta_{\text{rb}} + \Delta'_j$
  end for
  $w_{\text{rb}} \leftarrow P_m(\theta_{\text{rb}})$; action $\leftarrow$ `checkpoint`
end if
Record action and, for evaluation, $(B_{\text{fact}}, B_{\text{clean}}, B_{\text{attack}}, B_{\text{neigh}}, E_{\text{hist}})$
return w_{rb}, action

For the ledger-replay comparator, the checkpoint load in Algorithm 1 is replaced by reverse-order subtraction of the stored suffix and target from θ_{live} before the same writer re-estimation loop. Direct subtraction returns the packed line-2 candidate without routing; fixed replay replaces each Δ'_j by its stored Δ_j. These operators isolate descendant re-estimation, state routing, and checkpoint access.

C Forward Editing Reference Results

Table 8: Forward editing and single-edit commit-time rollback on CounterFact. Refusal rows are commit-conditional for target, paraphrase, and protected flips; zero-refusal rows are proposal-level. Values are means ± 95% confidence intervals over five seeds. The final row applies direct undo to the unconstrained MEMIT update.

Method	Target ↑	Paraphrase ↑	Prot. flip ↓	Refusal	Rb. error ↓
Constrained LR (matched)	95.1 ± 0.4	93.8 ± 0.4	5.4 ± 0.4	0%	0.1%
RExMap (matched)	94.8 ± 0.4	93.1 ± 0.4	4.2 ± 0.3	3.8%	0.1%
MEMIT (unconstrained)	96.8 ± 0.3	95.2 ± 0.4	6.1 ± 0.4	0%	–
ROME	97.1 ± 0.3	94.8 ± 0.4	7.2 ± 0.5	0%	–
MEMIT + direct undo	96.8 ± 0.3	95.2 ± 0.4	6.1 ± 0.4	0%	0.3%

The constrained low-rank row is the matched forward reference. RExMap uses that identical writer after proposal admission; the final row instead tests immediate direct undo for unconstrained MEMIT. The 0.1% rollback value is a single-edit commit-time measurement, distinct from the depth-1 and depth-5 chain protocols in Section 5.2.

D Admission Score Details

For a protected prompt x, let $\hat{y}_{\theta_0}(x) = \arg\max_y p_{\theta_0}(y \mid x)$ and

$$m(x) = z_{\theta_0}(x, \hat{y}_{\theta_0}(x)) - \max_{y \neq \hat{y}_{\theta_0}(x)} z_{\theta_0}(x, y), \qquad M(e) = \sum_{l \in L(e)} \|\Delta_{e,l}\|_2. \tag{6}$$

The clean component normalizes edit magnitude by available logit margin:

$$\delta_{\text{margin}}(e) = \frac{100}{500} \sum_{x \in \mathcal{P}_{\text{clean}}(e)} \min\left\{1, \frac{M(e)}{m(x) + \varepsilon_{\text{mach}}}\right\}. \tag{7}$$

For the 200 attacked probes $\widetilde{\mathcal{P}}(e)$,

$$\delta_{\text{search}}(e) = \frac{100}{200} \sum_{\tilde{x} \in \widetilde{\mathcal{P}}(e)} \mathbf{1}[\hat{y}_{\theta_e}(\tilde{x}) \neq \hat{y}_{\theta_0}(x(\tilde{x}))]. \tag{8}$$

Both terms are percentage-valued screening coordinates; their sum is used only for matched proposal admission, while rollback routing is trained directly against S_{rb}.

Table 9: Same-writer admission frontier on CounterFact, mean over five seeds. Counts are accepted/refused proposals. Target$_p$ and flip$_p$ are proposal-level; target$_c$, flip$_c$, and paraphrase are commit-conditional. Recovery is subtraction-only for the first five rows and state-routed subtraction/checkpoint recovery for the full policy. The final two rows share the admitted set and isolate recovery routing.

Rule	Count	Target$_p$	Target$_c$	Flip$_p$	Flip$_c$	Para.	Recovery
Always commit	2000/0	95.1	95.1	5.4	5.4	93.8	75.8%
Random matched	1924/76	91.3%	94.9%	5.2%	5.4%	93.6%	75.6%
Search only	1920/80	91.0%	94.8%	4.4%	4.6%	93.2%	80.1%
Margin only	1926/74	91.4%	94.9%	4.6%	4.8%	93.4%	78.7%
Combined score	1924/76	91.2%	94.8	4.0%	4.2	93.1	82.9%
Full RExMap policy	1924/76	91.2%	94.8	4.0%	4.2	93.1	97.8%

Table 10: Probe-budget sensitivity over 600 disjoint CounterFact proposals and five seeds. Templates, generator seeds, and relation instances are excluded from threshold selection. Flip and rollback failure are percentages; rollback uses ledger replay.

Budget	Flip	Spearman	AUROC	AUPRC	ECE	Brier	Rb. fail
50	3.7	0.74	0.79	0.64	0.071	0.112	6.9
100	4.3	0.79	0.83	0.70	0.064	0.106	7.4
200	5.1	0.81	0.86	0.75	0.058	0.097	8.1
400	6.4	0.80	0.85	0.74	0.061	0.101	8.9
800	7.8	0.83	0.88	0.78	0.053	0.091	10.2

E ADMISSION THRESHOLD SWEEP

Table 11: Combined-score threshold sweep on the same 2,000 CounterFact proposals. Target and protected flips are proposal-level; counts are accepted/refused.

$\delta_{\max}$	Acceptance	Count	Target	Prot. flip
3.0%	91.0%	1820/180	86.4%	3.4%
4.0%	93.5%	1870/130	88.7%	3.7%
5.0%	96.2%	1924/76	91.2%	4.0%
6.0%	98.0%	1960/40	93.1%	4.5%
7.0%	98.8%	1976/24	93.9%	4.8%
Always commit	100%	2000/0	95.1%	5.4%

The sweep traces a continuous utility–harm frontier. The 5.0% point retains 91.2% proposal-level target success and reduces proposal-level protected flips by 1.4 points relative to always commit. The strictest point improves harm further while trading away accepted target updates.

F Behavioral Raw Counts and State Slices

The predictor analysis is reported at three resolutions: matched decision rules expose which signals matter, raw mode–depth counts audit the validity surface, and deployment slices locate the remaining false-valid actions.

Table 12: State-rule and ablation errors (%). False-valid and false-stale use predicted-valid and predicted-stale denominators, respectively; fast share is the fraction authorized for subtraction.

Rule	False valid	False stale	Fast share
Full state logic	2.7%	26.9%	60.1%
Sketch threshold	7.9%	22.4%	61.8%
Depth threshold	10.8%	19.7%	64.4%
Parameter overlap only	8.6%	20.9%	64.1%
Entity overlap only	9.4%	20.2%	64.8%
Both overlaps	5.5%	23.7%	62.5%
Mode invalidation only	12.4%	18.5%	66.7%
Always checkpoint	0.0%	69.2%	0.0%

Table 13: Raw transaction pass/fail counts behind Table 1. Each cell contains 1,000 paired rollback transactions across five seeds.

Mode	1	5	10	20	35	50
FP32	974/26	942/58	908/92	828/172	741/259	673/327
BF16	961/39	918/82	863/137	772/228	689/311	598/402
INT8/GPTQ	919/81	854/146	792/208	687/313	574/426	486/514
Merged adapter	947/53	896/104	838/162	735/265	628/372	539/461

Table 14: False-valid rate (%) for the full state rule by deployment slice. Each mode and overlap slice contains 3,000 transactions; each depth slice contains 2,000.

Slice family	False-valid rate					
Mode	FP32: 1.4	BF16: 2.2	INT8: 4.1	Merge: 3.1		
Overlap	None: 1.2	Param.: 2.4	Entity: 2.7	Dual: 4.5		
Depth	1: 0.8	5: 1.1	10: 1.9	20: 2.8	35: 4.0	50: 5.6

The hardest slices are INT8, dual overlap, and depth 50, matching the behavioral expiry map. The state rule retains visible residual error in these slices and routes their majority toward checkpoint recovery.

G Refusal Recovery

Table 15 evaluates three revised transactions for the 76 proposals refused at the primary operating point, complementing the held-out calibration analysis (Guo et al., 2017). Narrowing scope recovers the largest share, whereas protected-set expansion gives the highest new target success.

Table 15: Refusal recovery on the 76 CounterFact proposals rejected at the 5.0% operating point. Rows are counterfactual retries on the same subset.

Recovery strategy	Recovery rate	Harm avoided (pts)	New target success
Narrowed scope	72%	3.1	91.4%
Reduced rank (r=8)	58%	2.4	88.7%
Protected-set expansion	41%	1.8	93.2%