REVISION COLLISIONS IN SEQUENTIAL MODEL EDITING: FAMILY-ALIGNED WRITE PATHS REDUCE FACTUAL INTERFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential corrections can remain individually successful while collectively corrupting unrelated facts: after 100 Llama-3.1-8B revisions, a single low-rank write path sustains 95.8% edit success yet adds 32.4 percentage points of net exact-match error on fixed, unedited factual probes. Write-path decomposition exposes the source of this interference by fixing MEMIT targets, edited layers, request order, summed read semantics, and a 2,752,512-parameter online budget while varying writable organization; among partitioned policies, online destination assignment is the experimental variable. REVSEP routes temporal, relational, and attribute revisions to family-specialized paths with shared overflow, while random, label-shuffled, route-blind, and shared controls separate semantic placement from generic dispersion. At matched partition geometry, REVSEP reduces edit-100 probe error from 30.6 to 27.1 points versus random assignment, a paired mean shift of -3.5 points (95% bootstrap interval $[-5.3, -1.7]$), and from 31.3 to 27.1 versus label-shuffled assignment (-4.2 $[-5.9, -2.5]$), while retaining 96.2% edit success. Relative to the shared path, the reduction is 5.3 points; the shared-to-random-to-aligned ordering also appears with zero subject–relation–template overlap, on Mistral-7B-v0.3, and in mixed-family histories, whereas route corruption and 91% occupancy progressively compress the gain. Write destination is therefore a controllable source of lifelong-edit retention: matching updates to specialized writable state protects factual behavior beyond equal-budget load dispersion.

1 INTRODUCTION

A language model can answer the newest correction correctly and still become less reliable with every accepted write. Dates, roles, and properties arrive as separate requests, but their parameter updates accumulate in the same evolving state. A later write can preserve its own target while increasing error on facts that were never requested for change. We call this accumulated contention a *revision collision*; it makes the destination of an update consequential even when the update itself succeeds.

Standard independent-edit evaluation resets the model after each request, so it cannot reveal this failure trajectory. Lifelong editors address sequential retention, but complete systems commonly change several factors together—target construction, routing, retrieval, writable capacity, and read-time activation. Their end-to-end gains do not identify whether interference changed because writes were dispersed, because content informed their placement, or because updates entered state specialized for their revision family. Isolating these explanations requires holding the incoming update and storage budget fixed while intervening on its destination.

Write-path decomposition provides that intervention. Every policy receives the same MEMIT residual targets at the same layers and checkpoints, in the same request order, with 2,752,512 online parameters. Untyped merges the budget into one path; Random Partition uses a balanced content-blind hash; Label-Shuffled deliberately breaks family–path correspondence; route-blind training matches auxiliary optimization using descriptor-derived pseudo-labels; and REVSEP assigns temporal, relational, and attribute revisions to family-specialized paths, with shared overflow for abstentions or saturated routes. Because the partitioned variants use the same path geometry, ranks, starting matrices, read rule, and

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.

write opportunities, aligned-versus-random and aligned-versus-shuffled contrasts test destination assignment rather than parameter count.

The primary study follows a prespecified 40/30/30 mixture of temporal, relational, and attribute single-family chains over ten paired seeds. At edit 100, Random reaches 30.6 points of net error increase on fixed unedited factual probes, Label-Shuffled reaches 31.3, and REVSEP reaches 27.1. The paired mean reductions are 3.5 points versus Random and 4.2 versus Label-Shuffled, with both 95% intervals excluding zero, while REVSEP retains 96.2% edit success. A joint-disjoint split, Mistral-7B-v0.3, and mixed-family histories preserve the same descriptive ordering; controlled route corruption and 500-edit chains reveal how assignment errors and occupancy erode the margin.

The primary contribution is a controlled factorization of writable destination policy inside one editing substrate. REVSEP supplies the constructive result: semantic alignment improves retention beyond equal-budget compartmentalization. The accompanying collision trajectories, route-count ablation, disjointness test, cross-backbone study, and occupancy diagnostics establish an operating principle for sequential editing—protecting factual behavior depends on where repeated updates accumulate, not only on whether each requested update is written.

2 RELATED WORK

2.1 DIRECT AND LIFELONG MODEL EDITING

ROME, MEMIT, MEND, KnowledgeEditor, Editable Neural Networks, and SERAC alter model behavior through direct updates, learned update rules, or external memory (Meng et al., 2022; 2023; Mitchell et al., 2022a; De Cao et al., 2021; Sinitsin et al., 2020; Mitchell et al., 2022b). PRUNE restrains update conditioning, AlphaEdit projects updates into preservation null spaces, and NMKE selects sparse neuron masks (Ma et al., 2025; Fang et al., 2025; Liu et al., 2025). GRACE and continual-knowledge training pursue durable behavior through discrete memory or continued optimization (Hartvigsen et al., 2023; Jang et al., 2022b). TemporalWiki evaluates acquisition and retention across changing corpus snapshots, while AToKE distinguishes temporal updates from corrections and tests access to historical facts (Jang et al., 2022a; Yin et al., 2024). Revision collisions instead track how repeated writes change a fixed population of unedited factual probes and use destination controls to identify the source of that interference.

2.2 ROUTED AND MODULAR WRITABLE CAPACITY

Adapters, LoRA, modular composition, and prompt- or prefix-based tuning create compact trainable paths (Houlsby et al., 2019; Hu et al., 2022; Pfeiffer et al., 2021; 2020; Mahabadi et al., 2021; Lester et al., 2021; Li & Liang, 2021; Ben Zaken et al., 2022; Wang et al., 2021). Conditional computation routes inputs through specialized components (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022), while replay and constrained optimization protect prior tasks (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019; Rebuffi et al., 2017). Closest to our setting, WISE shards edited knowledge into a side memory and uses a read-time router; ELDER assigns edits to a mixture of LoRA modules; MEMOIR combines residual memory, sparse masks, and retrieval; RILKE clusters edits into routed representation-intervention modules; and SoLA activates frozen per-edit LoRA modules through semantic matching (Wang et al., 2024; Li et al., 2025; Wang et al., 2025; Liu et al., 2026; Luo et al., 2026). These systems optimize complete lifelong-editing contracts in which allocation, retrieval, update construction, or read-time selection can change together. Write-path decomposition targets a complementary causal question inside one substrate: with incoming MEMIT targets and summed read semantics fixed, how much retention comes from generic compartmentalization, content-aware assignment, and correspondence with specialized state? Random, shuffled, and route-blind controls make those alternatives directly observable.

2.3 EVALUATION UNDER ACCUMULATED EDITS

Editing evaluations measure efficacy, locality, ripple effects, multi-hop consistency, temporal knowledge, and general-ability retention (Cohen et al., 2024; Zhong et al., 2023; Dhingra et al., 2022; Gu et al., 2024). Evidence that factual associations concentrate in transformer submodules motivates layer-selective intervention (Petroni et al., 2019; Roberts et al., 2020; Geva et al., 2021). Revision-

collision chains add accumulated writable-state interference to this evaluation space. Their matched controls distinguish successful update computation from compatibility with the state receiving those updates, while joint-disjoint data and routing perturbations test whether the assignment ordering survives unfamiliar descriptors and imperfect decisions.

3 WRITE-PATH DECOMPOSITION

3.1 WRITABLE SUBSTRATE AND READ PATH

REVSEP places low-rank writable modules at three edit-sensitive layers: 17, 21, and 25 for Llama-3.1-8B (Grattafiori et al., 2024), and 42, 54, and 66 for Llama-3.1-70B. Each layer contains temporal, relational, and attribute down/up matrix pairs of rank 32, plus a rank-16 shared overflow path. Temporal routes receive year or date changes, relational routes receive entity–role substitutions, and attribute routes receive property–value changes. A frozen linear probe over the subject's last-layer representation concatenated with its relation descriptor selects a route; its confidence threshold is selected before the online chains, and low-confidence assignments or capacity failures enter overflow.

Let $h^\ell \in \mathbb{R}^d$ be the residual stream entering edited layer ℓ and F^ℓ the frozen backbone map. With $A_z^\ell \in \mathbb{R}^{r_z \times d}$ and $B_z^\ell \in \mathbb{R}^{d \times r_z}$, the read-time computation is

$$h^{\ell+1} = F^\ell(h^\ell) + \sum_{z \in \mathcal{Z}} B_z^\ell A_z^\ell h^\ell.$$

All stored paths are summed for edited prompts, paraphrases, unrelated factual probes, MMLU, and MQuAKE. Routing therefore changes only which factor pair receives a write; it does not selectively hide paths at evaluation. Untyped supplies the same residual through its single merged pair, while all partitioned controls use the identical summed read rule. This separation rules out selective read-time activation as an explanation for their retention differences.

Counting matrices that can receive online MEMIT writes, the substrate contains

$$3 \times 3 \times 2 \times d \times 32 + 3 \times 2 \times d \times 16 = 672d$$

parameters. With $d = 4096$, this is exactly 2,752,512 online writable parameters, approximately 0.034% of the Llama-3.1-8B backbone. Every policy uses this total budget at the same layers. Untyped merges the route-specific and overflow ranks into one rank-112 shared factor pair, whereas a partitioned write activates rank 32 or rank 16. The shared comparison therefore tests the complete writable organization, including its active-rank geometry. Random, shuffled, route-blind, and aligned policies provide the sharper assignment comparisons because they match route count, available ranks, overflow capacity, initialization, warm-start schedule, and update opportunities.

3.2 DESTINATION-POLICY LADDER

The intervention ladder below records the control logic. Every partitioned policy starts a chain from the same post-warm-start route matrices, retains three route-specific paths plus shared overflow, and uses a destination function frozen before the first online request.

Policy	Geometry	Assignment signal	Identified comparison
Untyped	Shared	Common destination	Shared-path reference
Random	Partitioned	Balanced fixed hash	Generic compartmentalization
Label-Shuffled	Partitioned	Permuted family labels	Broken label correspondence
Route-blind	Partitioned	Descriptor pseudo-routes	Matched auxiliary optimization
REVSEP	Partitioned	Revision-family probe	Semantic alignment

Random Partition assigns requests through a seed-fixed balanced hash without inspecting request content. Label-Shuffled applies a seed-fixed derangement of the three family labels, so every family is sent to a different specialized path while label frequency and geometry are preserved. The route-blind

Edit request → frozen route probe → capacity check → routed MEMIT write

Temporal slot	Relational slot	Attribute slot	Overflow
Year / time updates	Entity–relation changes	Attribute edits	Ambiguous / saturated

Target layers: Llama-3.1-8B uses 17/21/25; Llama-3.1-70B uses 42/54/66. The backbone remains frozen, and only the selected route or overflow receives each online update.

Figure 1: Route-partitioned containment in REVSEP. Routing selects a route-specific path or overflow, while the MEMIT update rule itself is unchanged.

control trains the same probe for the same number of updates on balanced pseudo-routes from a seed-fixed random projection of subject–relation descriptors. REVSEP aligns the frozen assignment policy with temporal, relational, and attribute revision families. All partitioned policies use the same capacity test and rank-16 overflow destination; shuffled and aligned policies additionally share the same abstention decisions before the label map is applied. Thus the aligned–shuffled contrast changes family–destination correspondence without changing accepted requests, confidence decisions, or active write rank.

3.3 COMMON TARGET, DESTINATION-LOCAL WRITE, AND RENEWAL

For an edit $(s, r, o \to o')$, the probe predicts temporal, relational, or attribute type; the capacity check selects its route or shared overflow; MEMIT updates only that destination at each target layer; and a per-route FIFO log records the accepted write. The implementation reports occupancy as accepted routed load divided by the route's preallocated nominal load budget, clipped to $[0, 1]$ and averaged across edited layers. We use this quantity as capacity telemetry, not as an independent causal variable. A full reset restores typed matrices to their post-warm-start state while retaining overflow; a partial reset then replays the newer half of each restored route's log.

Let K_t^ℓ and R_t^ℓ denote MEMIT's key and residual-target matrices for edit t. They are computed once from the common frozen post-warm-start checkpoint, before policy-specific projection, and reused by every policy. The full-matrix target at layer ℓ is

$$D_t^\ell = R_t^\ell K_t^{\ell\top} \left(C^\ell + K_t^\ell K_t^{\ell\top} \right)^{-1}.$$

For destination z_t with down/up factors $A_{z_t}^\ell, B_{z_t}^\ell$ and admissible rank r_{z_t}, the online write is

$$B_{z_t}^{\ell,t+1} A_{z_t}^{\ell,t+1} = \Pi_{r_{z_t}} \left(B_{z_t}^{\ell,t} A_{z_t}^{\ell,t} + D_t^\ell \right),$$

where $\Pi_r(X) = U_{:r} \Sigma_{:r,:r} V_{:r}^\top$ is the truncated-SVD projection; the factors are recovered as $B = U_{:r} \Sigma_{:r,:r}^{1/2}$ and $A = \Sigma_{:r,:r}^{1/2} V_{:r}^\top$, and every unselected pair is unchanged. The covariance term C^ℓ is the same frozen MEMIT covariance cache for every policy. All policies also share K_t^ℓ, R_t^ℓ, solve tolerances, and optimization schedule. Comparisons among partitioned policies therefore preserve the incoming target and active projection rank while varying its destination.

3.4 WARM START

Warm start uses 12,000 temporal revisions derived from Wikidata (Vrandečić & Krötzsch, 2014), 2,000 CounterFact edits (Meng et al., 2022), and a separate 5K labeled router pool. The backbone remains frozen; slot parameters train with Adam at 10^{-4} for 20K alternating language-modeling and edit episodes, and the router is frozen before sequential evaluation. The resulting specialized slot checkpoint is copied unchanged into Random, Label-Shuffled, route-blind, and REVSEP before each matched chain. The retention objective is

$$\mathcal{L} = \sum_{t=1}^{T} \left[\mathcal{L}_{\text{edit}}^t + \lambda \sum_{j<t} \mathcal{L}_{\text{retain}}^{t,j} \right], \tag{1}$$

with one fixed λ and identical normalization, schedules, and update counts across matched variants.

4 EVALUATION

4.1 WORKLOADS, METRICS, AND INFERENCE

The evaluation centers on the matched assignment contrast after 100 edits, then varies descriptor overlap, backbone, traffic mixture, route error, and chain length. CounterFact contributes non-overlapping 2,000-example warm-start and sequential-evaluation subsets; its relation-schema map is fixed before sampling and separates entity–role substitutions from property–value corrections. The Temporal Revisions workload contains 5,000 held-out Wikidata updates from 2020–2024 whose target is a year or date replacement. MQuAKE measures multi-hop consistency (Zhong et al., 2023), and MMLU measures retained general ability (Hendrycks et al., 2021).

Each primary chain contains one temporal, relational, or attribute family, with family identity and request order shared across policies for each seed. The paired cohort uses 4/3/3 temporal/relational/attribute chains and weights seeds equally. Temporal chains come from Temporal Revisions; CounterFact triples are separated by relation schema into entity–role substitutions for relational chains and property–value corrections for attribute chains before sampling. Mixed traffic places temporal, relational, and attribute requests in one history at proportions 50/25/25. Out-of-family traffic exercises abstention and overflow. The primary grid is $k \in \{0, 10, \ldots, 100\}$, and the long-chain grid is $\{0, 50, 100, 150, 200, 250, 300, 400, 500\}$.

At checkpoint k, edit success is exact match averaged over all requested objects seen through k, and paraphrase accuracy uses their benchmark paraphrases. For each seed, $\mathcal{P}_i$ contains 100 benchmark factual paraphrases that are excluded from that chain's edit targets and selected once at the common post-warm-start checkpoint. The same probes are scored at every checkpoint under every policy. We measure *net factual interference* as the percentage-point change in exact-match error:

$$I_{i,m,k} = \frac{100}{|\mathcal{P}_i|} \sum_{q \in \mathcal{P}_i} \left(\mathbf{1}[f_{i,m,k}(q) \neq o_q] - \mathbf{1}[f_{i,0}(q) \neq o_q] \right).$$

This baseline-referenced quantity is a net error change: wrong-to-correct and correct-to-wrong transitions enter with opposite signs. Accordingly, we interpret I as factual-probe interference rather than as the fraction of all behavior damaged. By construction $I_{i,m,0} = 0$. With $\bar{I}_{m,k} = 10^{-1} \sum_i I_{i,m,k}$, the reported primary interference slope is the ordinary least-squares coefficient

$$\beta_m = \frac{\sum_k (k - \bar{k})(\bar{I}_{m,k} - \bar{I}_m)}{\sum_k (k - \bar{k})^2},$$

in percentage points per edit. This definition makes Table 1 exactly recoverable from Appendix A.

The Llama-3.1-8B control cohort uses 10 paired seeds. Each seed shares edit order, checkpoint schedule, and evaluation slice across methods. The 11,000 measurements per policy factor as 10 seeds $\times$ 11 checkpoints $\times$ 100 fixed probes; seeds, not these repeated measurements, are the inferential units. Table 1, Table 2, and Figure 2 report the common cohort; ROME and MEMIT remain five-seed editor references. Trajectory bands jointly resample matched seed tuples at each checkpoint, contrast intervals bootstrap paired mean endpoint differences, and method intervals for edit and paraphrase accuracy bootstrap seed aggregates. Experiments use 8$\times$A100-80GB GPUs, CUDA 12.4, and PyTorch 2.3.1.

4.2 CONTROLLED ESTIMANDS

Let $y_{i,m} = I_{i,m,100}$. For policies a and b, the paired endpoint estimand is the arithmetic mean of seedwise differences,

$$\widehat{\Delta}_{a-b} = \frac{1}{10} \sum_{i=1}^{10} (y_{i,a} - y_{i,b}).$$

For shared path u, random partition r, shuffled partition h, and aligned policy s, we report $\widehat{\Delta}_{s-u}$ for the complete organization change, $\widehat{\Delta}_{r-u}$ for content-blind partitioning, and the same-geometry assignment contrasts $\widehat{\Delta}_{s-r}$ and $\widehat{\Delta}_{s-h}$. The aligned–shuffled contrast additionally matches abstention and realized active rank request by request. Negative values mean less induced error. Paired bootstrap intervals jointly resample the ten seed tuples, preserving request order and evaluation slices within every resample.

Table 1: Matched destination policies after 100 sequential Llama-3.1-8B edits. I_{100} is net error increase on fixed unedited factual probes, in percentage points. Slopes are OLS fits to Appendix A; brackets are 95% intervals.

Policy	Edit ↑	Para. ↑	$\beta_I \downarrow$	$I_{100} \downarrow$	MMLU ↑	MQuAKE ↑
Untyped Control	95.8 ± 0.4	93.7 ± 0.4	0.333	32.4 [30.9,33.9]	61.1	40.5
Random Partition	96.0 ± 0.3	94.1 ± 0.4	0.313	30.6 [29.0,32.2]	61.8	42.4
Label-Shuffled	95.9 ± 0.4	94.0 ± 0.4	0.320	31.3 [29.7,32.9]	61.6	41.9
REVSEP	96.2 ± 0.3	94.6 ± 0.4	0.276	27.1 [25.6,28.6]	62.4	44.3

Table 2: Paired mean edit-100 shifts on Llama-3.1-8B. Point estimates equal the arithmetic differences of Table 1 means; intervals jointly bootstrap paired seed tuples. Negative values indicate less net probe error.

Contrast	ΔI_{100}	What changes
Random − Untyped	-1.8 $[-2.9, -0.3]$	Complete shared-to-partitioned organization
REVSEP − Untyped	-5.3 $[-6.4, -3.2]$	Complete shared-to-aligned organization
REVSEP − Random	-3.5 $[-5.3, -1.7]$	Family alignment at matched partition geometry
REVSEP − Label-Shuffled	-4.2 $[-5.9, -2.5]$	Restored family–path correspondence

4.3 STRICT GENERALIZATION AND SECOND ANCHOR

The joint-disjoint stress test draws its 100-edit chains from the 5,000-example evaluation pool while enforcing zero shared subjects, relations, or prompt templates with warm-start and router training. It evaluates Untyped, Random, and REVSEP with 10 paired seeds. A second matched anchor uses Mistral-7B-v0.3 (Mistral AI, 2024) with target layers 13/17/21 and 8 paired seeds.

5 RESULTS

5.1 ALIGNMENT IMPROVES ON LOAD DISPERSION

The partitioned controls isolate the central result. Random hashing disperses each homogeneous chain, yet REVSEP lowers I_{100} from 30.6 to 27.1 points; deranging family correspondence reaches 31.3. The same ordering appears in the OLS accumulation rates, 0.313, 0.320, and 0.276 points per edit. Untyped reaches 32.4 points, while REVSEP retains 96.2% edit success. Separately estimated editor references attain 97.1/96.8% edit success and slopes of 0.42/0.38 for ROME/MEMIT; the destination-policy cohort is the controlled comparison because it shares targets, checkpoints, and probe slices.

Every paired endpoint interval in Table 2 excludes zero. The aligned-versus-random and aligned-versus-shuffled rows hold path geometry and nominal ranks fixed; aligned versus shuffled also matches abstention and realized active rank. The route-blind auxiliary control reaches $I_{100} = 31.4 \pm 1.3$, close to shuffled assignment despite matched auxiliary optimization. A separate equal-budget route-count ablation gives slopes 0.360, 0.350, 0.344, and 0.346 for one, two, three, and four content-blind paths, versus 0.310 for three semantic paths (Table 8). Assignment semantics, rather than route multiplicity or probe-training compute, account for the additional separation.

Figure 2 shows the assignment effect accumulating rather than appearing as an isolated endpoint fluctuation. The trajectories separate most clearly after 40 edits, reaching 32.4 versus 27.1 points for Untyped and REVSEP. Post-chain MMLU and MQuAKE in Table 1 follow the same ordering, with REVSEP at 62.4 and 44.3.

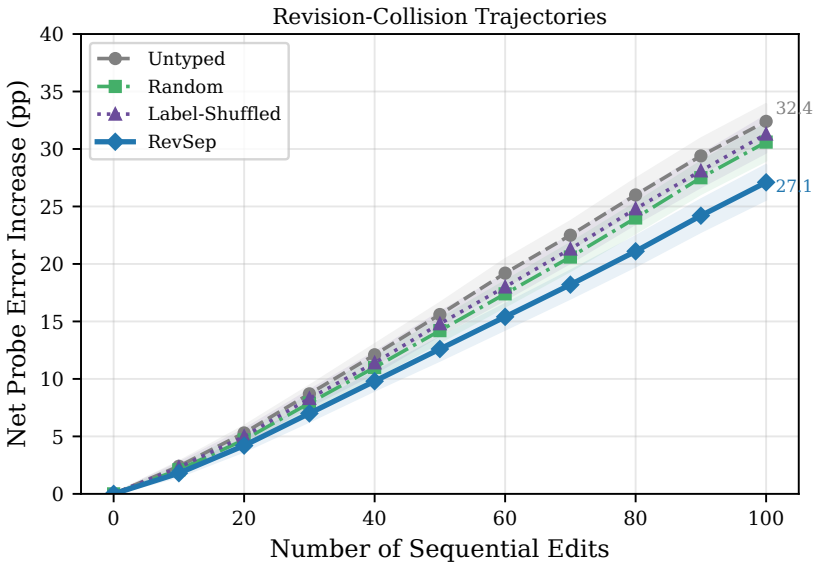


Figure 2: Net factual interference accumulates over the 100-edit chain. Markers are checkpoint means, lines connect them without smoothing, and bands are 95% joint paired-seed bootstrap intervals. The late-chain ordering agrees with Table 2.

Table 3: Joint-disjoint Llama-3.1-8B collision stress test. Subjects, relations, and templates have zero overlap with warm-start and router data; I_{100} is in percentage points.

Method	Edit ↑	β_I ↓	I_{100} ↓
Untyped	95.3 ± 0.4	0.384 ± 0.018	34.8 ± 1.5
Random	95.4 ± 0.4	0.367 ± 0.019	33.2 ± 1.6
RevSep	95.5 ± 0.3	0.338 ± 0.017	30.1 ± 1.4

5.2 The effect survives strict disjointness and a second backbone

The stricter split raises net probe error for every method but preserves the shared-to-random-to-aligned ordering. RevSep is 3.1 points below Random descriptively and improves on Untyped by -4.7 $[-6.8, -2.6]$ points with an exact endpoint sign-flip $p = 0.004$, while edit success remains closely matched. Family-aligned retention therefore survives removal of entity, relation, and template reuse between training and evaluation.

On Mistral, the RevSep–Untyped slope contrast is -0.028 $[-0.046, -0.011]$ with an exact sign-flip $p = 0.016$ and 7/8 seedwise wins. Endpoint interference follows the shared-to-random-to-aligned sequence 35.0, 33.9, and 31.9 points, a descriptive 2.0-point aligned-versus-random separation on a second backbone.

5.3 Mixed-family histories exercise all routes

The mixed history places all three families in one chain and preserves the shared-to-random-to-aligned endpoint ordering: 35.0, 33.6, and 31.8 points. The aligned policy is 1.8 points below Random descriptively, and its paired shift versus Untyped is -3.2 $[-5.1, -1.3]$ with an endpoint sign-flip $p = 0.018$. On a jointly held-out 2,000-request evaluation, the frozen probe reaches 90.6% macro-F1 with 3.15% abstention. Predicted temporal/relational/attribute/abstain frequencies are 35.2/32.3/29.4/3.15%, route occupancies after 100 in-family edits are 38/32/34%, and expected calibration error is 4.7%. Table 5(c) shows a monotone 27.3-to-30.5-point response as controlled route-label flips rise from 0% to 25%, tying assignment quality to retention without skipped writes.

Table 4: Matched Mistral-7B-v0.3 destination controls after 100 in-family edits. I_{100} is net factual-probe error increase in percentage points; values report means with 95% intervals where shown.

Policy	Edit ↑	β_I ↓	I_{100} ↓	MMLU ↑	MQuAKE ↑
MEMIT	95.9 ± 0.5	0.41 ± 0.023	37.2 ± 1.8	59.2	39.8
Untyped	95.2 ± 0.5	0.389 ± 0.021	35.0 ± 1.7	59.6	40.4
Random	95.3 ± 0.4	0.375 ± 0.022	33.9 ± 1.8	59.9	41.0
REVSEP	95.4 ± 0.4	0.36 ± 0.020	31.9 ± 1.6	60.4	42.1

Table 5: Operating-envelope measurements. Panel (a) reports edit-100 net factual interference when 50% of requests are temporal and the remainder is split between relational and attribute revisions; panel (b) reports overflow over fixed request batches; panel (c) perturbs route labels in the primary cohort.

(a) Mixed-traffic retention Method	I_{100} ↓	Paired Δ vs. Untyped
MEMIT	36.1 ± 1.8	—
Untyped	35.0 ± 1.7	—
Random	33.6 ± 1.6	—
REVSEP	31.8 ± 1.5	$-3.2\ [-5.1, -1.3]$
(b) Overflow by workload		
In-family	48/1000 (4.8%)	
Mixed	157/1000 (15.7%)	
Out-of-family	591/1000 (59.1%)	
(c) Controlled route corruption		
Route-label flips	I_{100} ↓	95% interval half-width
0%	27.3	± 1.1
10%	28.7	± 1.3
25%	30.5	± 1.5

6 MECHANISM AND OPERATING ENVELOPE

The same-geometry comparisons identify specialization compatibility as the key mechanism. Random hashing reduces load concentration but ignores path semantics; label shuffling preserves routing machinery while breaking correspondence; route-blind training controls auxiliary optimization; and controlled label flips provide a dose response. REVSEP improves over all four conditions. The joint-disjoint test removes familiar descriptors, the Mistral study changes the backbone, and the mixed history puts all families into one online sequence.

Capacity measurements add a distinct diagnostic. During early accumulation, local trajectory intervals overlap; after 40 edits, assignment policy visibly separates retention. Figure 3 shows that REVSEP maintains lower interference through 300 in-family edits, while the margin compresses by 500 edits as mean occupancy approaches 91%. Traffic mismatch produces a parallel pattern as overflow rises from in-family through mixed to out-of-family requests.

Reset and replay show how writable paths can recover plasticity. After 100 edits, full reset raises new-edit success from 94.8% to 96.0% and returns prior-target accuracy to its 72.1% pre-chain baseline. Partial replay reaches 95.6% new-edit success and 84.7% absolute prior-target accuracy, a 12.6-point gain over that baseline. Over 1,000 edits, online latency is +4.1 $\pm$ 0.7% for Untyped and +7.3 $\pm$ 0.8% for REVSEP relative to MEMIT; peak-memory overhead is +3.0 $\pm$ 0.5% and +3.4 $\pm$ 0.5%. Occupancy and overflow are therefore useful monitoring signals, while the reset experiment directly demonstrates the retention–plasticity tradeoff.

The identified effect is the destination-policy contrast within the fixed MEMIT substrate. All partitioned policies match path geometry, starting matrices, and summed read semantics; aligned and shuffled assignment also match abstentions, accepted requests, and active write rank. The

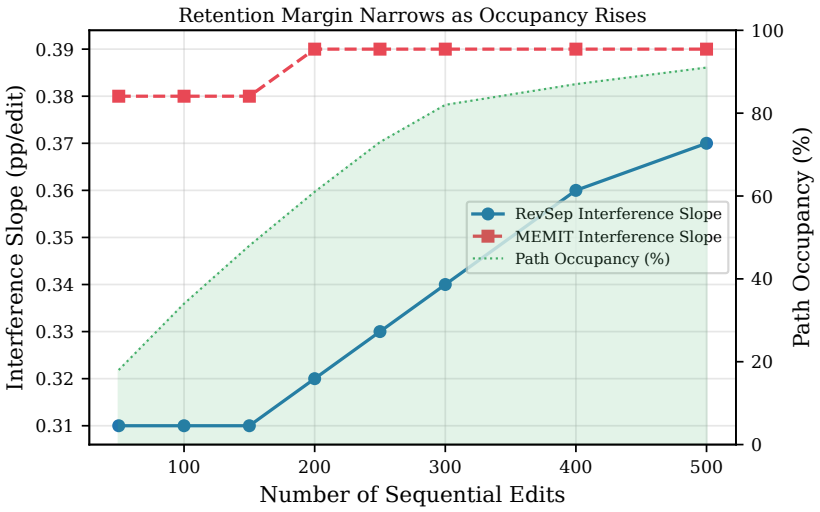


Figure 3: Long-chain operating envelope. REVSEP remains below the MEMIT interference slope through 500 edits, while its margin narrows as mean path occupancy rises from 34% at 100 edits to 91% at 500. Occupancy is telemetry measured on the same chain, not an independently randomized cause.

shared-path contrast additionally changes active projection rank and therefore measures the complete organization. The primary inference averages the prespecified 4/3/3 single-family seed mixture, while mixed traffic supplies the complementary heterogeneous history and cross-backbone/random gaps are descriptive where paired assignment intervals are unavailable. This scope keeps the central aligned-versus-random/shuffled result precise, and the route-noise and occupancy studies show how it changes under measured operating pressure.

7 CONCLUSION

Sequential editing fails along two distinct axes: a requested correction can succeed while accumulated writes increase error elsewhere. Write-path decomposition makes the second axis controllable. After 100 Llama-3.1-8B revisions, family-aligned paths reduce net error on fixed unedited factual probes by 3.5 points versus random dispersion and 4.2 versus label-corrupted assignment, with paired intervals excluding zero and 96.2% edit success. The ordering reappears under joint subject–relation–template disjointness, on Mistral-7B-v0.3, and in mixed-family traffic; route corruption and rising occupancy progressively narrow the margin. Treating write destination as part of the editing algorithm therefore converts revision collisions from an unobserved side effect into a measurable and governable retention mechanism.

REPRODUCIBILITY STATEMENT

Sections 3–6 specify the intervention, controls, estimands, workloads, metrics, hardware, pairing, and uncertainty protocol; Appendix D consolidates training, routing, checkpoint, and reset details.

AI USE STATEMENT

Generative AI tools assisted with drafting and revising manuscript prose, presentation, and submission-preparation checks. The authors are responsible for the research design, experimental results, analysis, and all scientific claims.

REFERENCES

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL*, 2022.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 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.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, et al. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10: 257–273, 2022.

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 *The Thirteenth International Conference on Learning Representations*, 2025.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *EMNLP*, 2021.

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

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua 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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.934. URL https://aclanthology.org/2024.emnlp-main.934/.

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, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, et al. Parameter-efficient transfer learning for NLP. In *ICML*, pp. 2790–2799, 2019.

Edward J Hu et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6237–6250. Association for Computational Linguistics, 2022a. doi: 10.18653/v1/2022.emnlp-main.418. URL https://aclanthology.org/2022.emnlp-main.418/.

Joel Jang, Seonghyeon Ye, Sohee Yang, et al. Towards continual knowledge learning of language models. In *ICLR*, 2022b.

James Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521–3526, 2017.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

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, 2025. doi: 10.1609/aaai.v39i23.34622. URL https://ojs.aaai.org/index.php/AAAI/article/view/34622.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

Jinzhe Liu, Junshu Sun, Shufan Shen, Chenxue Yang, and Shuhui Wang. Edit less, achieve more: Dynamic sparse neuron masking for lifelong knowledge editing in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xuyuan Liu, Shengyu Chen, Xinshuai Dong, Yanchi Liu, Xujiang Zhao, Haoyu Wang, Yujun Yan, Haifeng Chen, and Zhengzhang Chen. Representation interventions enable lifelong knowledge memory control in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5414–5436. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.246.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, 2017.

Haihua Luo, Xuming Ran, Tommi Kärkkäinen, Zhonghua Chen, Jiangrong Shen, Qi Xu, and Fengyu Cong. Reversible lifelong model editing via semantic routing-based LoRA. *arXiv preprint arXiv:2603.11239*, 2026.

Jun-Yu Ma, Hong Wang, Hao-Xiang Xu, Zhen-Hua Ling, and Jia-Chen Gu. Perturbation-restrained sequential model editing. In *The Thirteenth International Conference on Learning Representations*, 2025.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *NeurIPS*, 2021.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, volume 35, 2022.

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

Mistral AI. Model card for Mistral-7B-v0.3, May 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.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *ICML*, 2022b.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *EMNLP*, pp. 2463–2473, 2019.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *EMNLP*, 2020.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.39. URL https://aclanthology.org/2021.eacl-main.39/.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *CVPR*, 2017.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *EMNLP*, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

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

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489.

Ke Wang, Yiming Qin, Nikolaos Dimitriadis, Alessandro Favero, and Pascal Frossard. MEMOIR: Lifelong model editing with minimal overwrite and informed retention for LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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*, volume 37, 2024. URL https://arxiv.org/abs/2405.14768.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of ACL-IJCNLP*, 2021.

Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. History matters: Temporal knowledge editing in large language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (17):19413–19421, 2024. doi: 10.1609/aaai.v38i17.29912. URL https://ojs.aaai.org/index.php/AAAI/article/view/29912.

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*, 2023.

A Estimator Checkpoints

Table 6: Checkpoint net factual interference I_k with 95% joint paired-seed bootstrap intervals. The final row is the OLS slope of each displayed mean trajectory in percentage points per edit, providing a direct reconciliation with Table 1.

k	Untyped	Random	Label-Shuf.	RevSep
0	0.0	0.0	0.0	0.0
10	2.4 [2.0,2.8]	2.1 [1.7,2.5]	2.3 [1.9,2.7]	1.8 [1.5,2.2]
20	5.3 [4.7,5.9]	4.7 [4.1,5.3]	5.0 [4.4,5.6]	4.2 [3.7,4.8]
30	8.7 [8.0,9.4]	7.9 [7.2,8.6]	8.3 [7.6,9.0]	7.0 [6.3,7.7]
40	12.1 [11.2,13.0]	11.0 [10.1,11.9]	11.4 [10.5,12.3]	9.8 [9.0,10.6]
50	15.6 [14.6,16.6]	14.2 [13.2,15.2]	14.8 [13.8,15.8]	12.6 [11.6,13.6]
60	19.2 [18.0,20.4]	17.4 [16.3,18.5]	18.0 [16.9,19.1]	15.4 [14.3,16.5]
70	22.5 [21.3,23.7]	20.6 [19.4,21.8]	21.3 [20.1,22.5]	18.2 [17.0,19.4]
80	26.0 [24.6,27.4]	24.0 [22.7,25.3]	24.8 [23.5,26.1]	21.1 [19.8,22.4]
90	29.4 [27.9,30.9]	27.5 [26.1,28.9]	28.1 [26.7,29.5]	24.2 [22.8,25.6]
100	32.4 [30.9,33.9]	30.6 [29.0,32.2]	31.3 [29.7,32.9]	27.1 [25.6,28.6]
β_I	0.333	0.313	0.320	0.276

B Router and Capacity Measurements

The held-out confusion matrix resolves assignment errors by revision family, while the separate route-count ablation tests whether adding content-blind paths monotonically improves retention.

Table 7: Held-out router confusion counts. Columns are predicted temporal, relational, attribute, and abstain.

True route	Temporal	Relational	Attribute	Abstain
Temporal	651	28	18	23
Relational	31	588	24	17
Attribute	22	29	546	23

Table 8: Separate route-count ablation at fixed total online writable and overflow capacity. Values are seedwise slopes from the ablation cohort.

Writable geometry	Interference slope
One shared path	0.36
Two random paths	0.350 ± 0.017
Three random paths	0.344 ± 0.015
Four random paths	0.346 ± 0.018
Three Label-Shuffled paths	0.351 ± 0.016
Three semantic RevSep paths	0.31

The non-monotonic random-path series shows that simply increasing route count is insufficient. Three semantic paths outperform every equal-budget random or shuffled organization, consistent with alignment between revision family and writable destination.

C Reset Diagnostics

Reset diagnostics distinguish recovery of new-edit plasticity from absolute accuracy on previously requested targets.

Table 9: Reset behavior after 100 Llama-3.1-8B edits. "Prior retained" is absolute exact match on the chain targets; the fresh row measures their pre-edit accuracy at the common post-warm-start checkpoint.

Condition	New edit ↑	Prior retained	MMLU
Fresh model	96.2%	72.1%	63.2
After 100 edits	94.8%	91.4%	62.4
Full slot reset	96.0%	72.1%	62.9
Partial replay reset	95.6%	84.7%	62.6

Full slot reset returns prior-target accuracy to the fresh baseline; retained overflow writes account for the small residual differences in new-edit success and MMLU. Partial replay restores additional chain-specific corrections while recovering new-edit plasticity.

D Implementation and Experiment Summary

All experiments ran on 8×A100-80GB GPUs with CUDA 12.4 and PyTorch 2.3.1. Writable modules are LoRA-variant residuals; all paths are summed at inference, and the frozen router selects only the destination updated online. The route probe is linear on the concatenation of the edit subject's last-layer hidden state and its relation descriptor. Its 5K training pool uses the same temporal, relational, and attribute schema as evaluation; the held-out evaluation contains 720, 660, and 620

examples of the three classes. The abstention threshold is selected on router validation and then frozen. Low-confidence predictions enter overflow, and a capacity failure occurs when accepting the next write would raise normalized route occupancy above one. Causal tracing uses 500 prompts per layer. Warm start alternates language-modeling and edit episodes with Adam at 10^{-4} for 20K steps. The resulting slot matrices are copied identically across partitioned policies; Shuffled uses one of the two three-label derangements for each seed. Online evaluation freezes the backbone and every destination function.

The primary pool-disjoint protocol keeps 12,000 temporal warm-start edits, 5,000 temporal evaluation edits, disjoint 2,000-example CounterFact warm-start and evaluation subsets, and 5K router tuples example-disjoint. The joint-disjoint protocol additionally excludes shared subjects, relations, and prompt templates. Each matched seed fixes request order, the 100 unrelated-probe slice, and checkpoint cadence across methods. A seed-fixed balanced hash over canonical request identifiers defines Random assignments; a seed-fixed family-label derangement defines Shuffled assignments. Primary chains evaluate every 10 edits, producing 10 seeds $\times$ 11 checkpoints $\times$ 100 probes. Long chains evaluate at $0, 50, 100, 150, 200, 250, 300, 400, 500$. Full reset restores typed matrices and clears their FIFO logs while retaining overflow; partial reset then replays the newer half.

An additional Llama-3.1-70B scale check uses target layers 42/54/66. MEMIT and REVSEP attain edit success of 97.4 and 97.1, interference slopes of 0.29 and 0.24, MMLU of 78.1 and 79.3, and MQuAKE of 52.4 and 55.8, respectively. This larger-model result follows the direction established by the fully matched Llama-3.1-8B and Mistral studies.

Table 10: Experiment summary and statistical units.

Experiment	Statistical units	Purpose
Primary controlled chain	Common paired destination-policy cohort	Method estimates and paired effects in Tables 1–2
Checkpoint trajectory	Same seedwise primary cohort	Paired-bootstrap progression in Figure 2
Joint-disjoint chain	Matched Untyped, Random, and REVSEP runs	Generalization beyond shared subjects, relations, and templates
Mistral chain	Matched four-method runs	Second-backbone estimate in Table 4
Mixed traffic	Matched four-method runs	Heterogeneous-traffic and overflow behavior
Route-blind control	Same primary-chain geometry and update count	Auxiliary-optimization alternative
Route corruption	Same paired primary cohort at three flip rates	Assignment-error dose response
Long chain	Checkpoints through 500 sequential edits	Occupancy-linked margin compression
Reset and replay	Fresh, post-chain, full-reset, and partial-replay states	Plasticity recovery and prior-target retention
Online cost	Repeated 1,000-edit measurements	Latency and peak-memory overhead