BEYOND UPDATE COUNTS: A CONTROL MAP FOR EVALUATING TEST-TIME MODEL UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time updating lets a language model change persistent state after training—such as model weights, adapter parameters, or a memory module—before or while answering queries. Existing evaluations often compare systems using update count or frequency and performance measured immediately afterward. These summaries omit what is updated, where, when, and in what order. The same total can therefore hide an important difference: one system may answer after one update while another answers after two. We introduce the *Control Map*, an evaluation framework that records the committed updates preceding each scored answer and separately evaluates behavioral changes, immediate learning (*acquisition*), later recall (*retention*), preservation of unrelated behavior (*locality*), and performance in new settings. We prove when counts alone determine final state and show, under stated assumptions, how differences between pre-answer histories limit differences in average scores. Our experiments establish three lessons. First, choices hidden by update count matter. At one 760M-parameter checkpoint, 24 equal-count schedules span 19.7%–34.2% accuracy; updating the early and late layer groups more often than the middle groups performs best in this experiment. Second, apparent count effects can disappear after proper matching: in 1,600 pairs of runs assigned different nominal dose labels but executing the same updates in the same order, both runs end with the same model state and locality. Third, immediate learning does not guarantee a lasting or well-contained change. Across six selected CounterFact LoRA settings, new information is acquired in all six; in a separate ROME study, edits are retained at all 13 tested operating points but fail locality at all 13. The relative ordering of editors also changes across the tested combinations of models, benchmarks, and tests. Therefore, evaluations should record which updates precede each answer and assess these questions separately.

1 INTRODUCTION

Test-time model updates let a model change persistent state after training—such as model weights, adapter parameters, or a memory module—before or while answering queries. This includes systems based on fast weights, test-time training, adaptive memory, and model editing (Schlag et al., 2021; Sun et al., 2025; Behrouz et al., 2025b; Meng et al., 2022). A policy, or update plan, must choose what content to write, which parameters or layers to modify, which update mechanism to use (called the *writer* below), how strongly and often to update, and when those updates occur relative to each query. Representative studies vary or report layer location, update timing or frequency, update strength, and the number of edits processed together (Tandon et al., 2025; Lee et al., 2023; Wakayama, 2026; Meng et al., 2023; Yoon et al., 2024). Yet comparisons based only on policy labels, update frequencies, or total update counts compress these choices into a small number of summary values.

Two systems can perform the same total number of updates over a full run, yet an answer may be produced after one update in one system and after two in the other (Figure 1). This difference can be consequential: if the second update contains information needed for the query, one system is evaluated before it can use that information, whereas the other is evaluated afterward. Describing both as "two-update" systems therefore hides a real difference in the information available when they answer. Conversely, different update plans can produce the same ordered update history before an answer. In that case, comparing their policy labels cannot reveal whether the intended difference between the plans matters.

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.

Establishing that the histories differ is only the first step. A complete evaluation must ask four questions: (1) which ordered updates were completed before each answer, and did the compared plans differ; (2) did those differences change the model's answer or evaluation score; (3) was new information learned immediately (*acquisition*), recalled later (*retention*), and learned without changing unrelated behavior (*locality*); and (4) does the conclusion hold in a new model or benchmark setting? Evidence for one question cannot answer the others: different histories do not guarantee a behavioral change, immediate learning does not guarantee later recall or locality, and a result in one setting does not guarantee the same result in another.

We address these questions with the *Control Map*, an evaluation framework rather than a new update algorithm. Before examining outcome scores, the evaluator specifies a tracing rule that identifies which completed updates may affect each answer. For every update plan and example, the framework records these updates in order and groups plans that produce the same history. Its conclusions therefore depend on the stated tracing rule; the framework does not claim to discover every possible path to an answer automatically. It reports separately whether the histories differ, whether behavior changes, whether acquisition, retention, and locality each meet their requirements, and whether the conclusion is supported by direct evaluation in a new setting. This framework brings together three lines of prior work: methods that update model state during inference, evaluations of model editing and memory, and methods for identifying which intervention or computation actually affects an output.

1.1 Related work and positioning

Stateful test-time adaptation. Fast weights, test-time-training layers, and hierarchical memory systems study how models should update state during inference (Schlag et al., 2021; Yang et al., 2024; Sun et al., 2025; Behrouz et al., 2025b;a; 2026). Related work treats update frequency, layer location, routing, and update strength as design variables (Tandon et al., 2025; Hu et al., 2025; Sim, 2026; Wakayama, 2026). These studies primarily develop mechanisms for updating state. We study the complementary evaluation question: once an update method and plan have been chosen, which updates were actually completed before each evaluated answer?

Model editing and memory evaluation. ROME, MEMIT, AlphaEdit, and model-editing toolkits evaluate edit success, generalization to related prompts, locality, and propagation to related facts (Meng et al., 2022; 2023; Fang et al., 2025; Wang et al., 2024; Zhang et al., 2024; Cohen et al., 2024). Recent work further shows that reported editor rankings can change with the evaluation metric, whether edits are processed individually or in batches, and other protocol choices (Gupta et al., 2024; Pohl et al., 2025; Yoon et al., 2024; Liu et al., 2026). UniEdit extends evaluation to open-domain, multi-step test questions (Chen et al., 2025). This literature establishes the need to evaluate several dimensions of model editing. We add a complementary execution question: which ordered updates were completed before each evaluated answer? The Control Map links each result to that history and treats immediate learning, later recall, and preservation of unrelated behavior as separate requirements.

Assignments and computational provenance. Causal-inference work distinguishes an assigned intervention from what a case actually experiences (VanderWeele & Hernán, 2013; Aronow & Samii, 2017; Sävje, 2024; Adhikari et al., 2026; Huber et al., 2026), while program-analysis work traces which earlier operations can influence an output (Weiser, 1981; Korel & Laski, 1988; Zhang et al., 2020; Lee et al., 2021; Tian & Liu, 2026). These literatures explain why an assigned label may not describe what actually happened and how relevant operations can be traced. However, neither provides an evaluation that connects each model answer to the ordered updates completed before it and then separately tests behavioral change, retention, locality, and performance in a new setting. The Control Map provides this connection using a tracing rule fixed before scores are examined. Its guarantees depend on that rule capturing every way the update plan can affect the score, and claims about a new setting require direct evaluation there rather than assuming that the original result will carry over (Pearl & Bareinboim, 2014).

Contributions. Motivated by these complementary lines of work and the gaps identified above, we make three contributions:

1. **Evaluation framework.** We introduce the Control Map, which records the ordered updates completed before each answer and separately reports whether plans produce different histories, change behavior, satisfy acquisition, retention, and locality requirements, and continue to work in a new setting.
2. **Formal guarantees.** We show that greater overlap between recorded histories places a tighter bound on average-score differences under stated assumptions. We also characterize when update counts determine final state and extend the score bound to a known amount of mismatch in the scoring process.
3. **Empirical evidence.** We find that layer placement, query timing, and the number of updates affect outcomes in setting-dependent ways. At one checkpoint, 24 equal-count schedules span 19.7%–34.2% accuracy, while matching the same updates and order produces 1,600 exact terminal-state and locality ties. Across six selected CounterFact settings, acquisition passes in all six, but retention and locality pass in none: information learned immediately can later be forgotten or alter unrelated answers. Method rankings can also change across models, benchmarks, and tests.

Section 2 formally defines update plans, pre-answer histories, and the Control Map. Section 3 asks when different plans actually produce distinguishable histories. Section 4 studies the effects of layer placement, timing relative to the query, and the number of updates completed before a query. Section 5 separates immediate learning from later recall and preservation of unrelated behavior. Section 6 tests whether conclusions continue to hold across new models and benchmarks. Section 7 concludes. Additional experimental details, threshold definitions, sensitivity analyses, and provenance records appear in Appendices A–E; complete proofs appear in Appendix F.

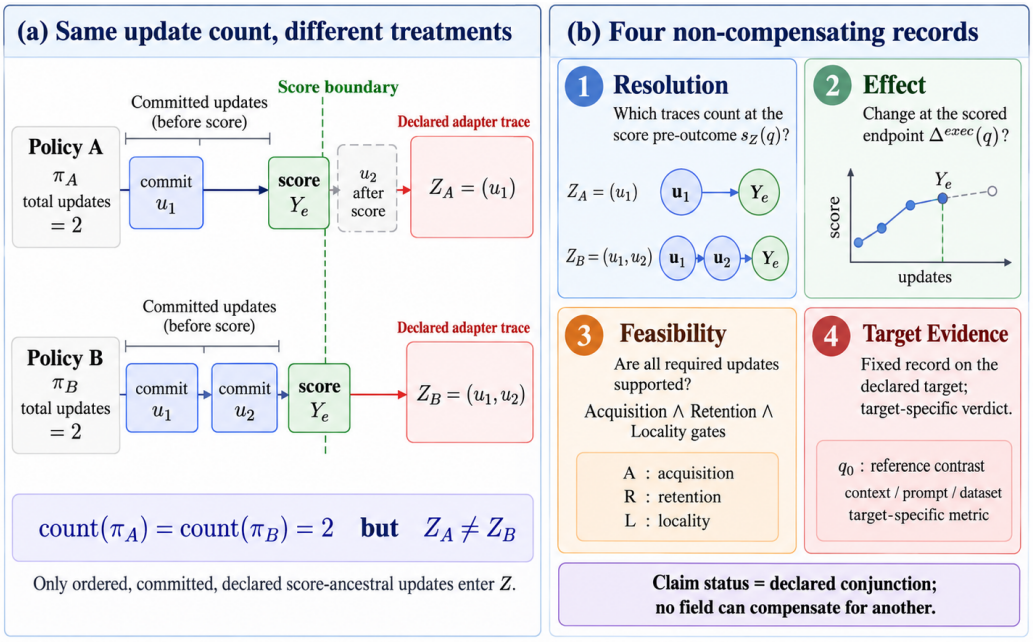


Figure 1: **The same update count can yield different histories at an answer.** Both plans perform u_1 and u_2, but only plan B performs u_2 before scoring, so their recorded histories are (u_1) and (u_1, u_2) under the stated tracing rule. The Control Map reports update history, measured effect, memory quality, and results in a new setting separately. The formal record also stores cost; values are schematic.

2 THE CONTROL MAP AND PROBLEM FORMULATION

2.1 STATEFUL POLICIES AND OPERATING POINTS

A setting is $c = (M, \mathcal{T}, e, \mathcal{B})$, where M is the model, $\mathcal{T}$ the task and interface, e the output being evaluated, and $\mathcal{B}$ an optional cost budget; a row r is one evaluation example. An operating point pairs a setting with a fully specified update policy and configuration.

Definition 1 (Policy and declared pre-answer update history)**.** Fix a setting $c = (M, \mathcal{T}, e, \mathcal{B})$ and a filtration $(\mathcal{F}_t)_t$ containing only information available before update opportunity t. A test-time update policy π consists of the action kernel

$$a_t = (g_t, x_t, \ell_t, w_t, \eta_t) \sim \pi_{\mathrm{act}}(\cdot \mid \mathcal{F}_t, c; \kappa, \phi), \tag{1}$$

together with its cadence κ, query-relative phase ϕ, and commit rule. Here g_t admits an update, x_t selects available content, ℓ_t selects parameter support, w_t selects the writer, and η_t its strength. With commit indicator δ_t, model state θ_t, and writer state ω_t, the complete mutable state obeys

$$(\theta_{t+1}, \omega_{t+1}) = \begin{cases} F_{w_t,\ell_t,\eta_t}(\theta_t, \omega_t; x_t), & \delta_t g_t = 1, \\ (\theta_t, \omega_t), & \delta_t g_t = 0. \end{cases} \tag{2}$$

For row r and scored endpoint e, let $\Lambda_{r,\pi,e}$ be the ordered log of committed transitions preceding e. A dependency specification $\mathcal{D}_{c,e}$, maps this log to the canonical ordered sequence

$$z_{r,\pi,e} := \mathcal{D}_{c,e}(\Lambda_{r,\pi,e})$$

of committed transitions it declares able to influence the score. We call $z_{r,\pi,e}$ the *declared pre-answer update history*. Policies π and π' are history-equivalent on (r, e), written $\pi \sim_{r,e} \pi'$, iff $z_{r,\pi,e} = z_{r,\pi',e}$.

2.2 PRE-ANSWER HISTORIES AND COMPARISON RESOLUTION

Definition 2 (Contrast-specific history resolution and score sufficiency)**.** Fix the scored endpoint e and the dependency specification of Definition 1. Let Π be a finite policy catalogue and $\mathcal{R}$ a finite row set with policy-independent weights $w_r \geq 0$ and $\sum_r w_r = 1$. Assume the history and outcome spaces are finite or standard Borel. For each $\pi \in \Pi$, let $K_{\pi,e}(\mathrm{d}z \mid r)$ be the conditional law of its declared history. Let R denote a row sampled according to the weights and Z its declared history, and define

$$P^{\pi}_{R,Z}(\{r\} \times A) = w_r K_{\pi,e}(A \mid r).$$

A contrast q is a weighted comparison of policies whose coefficients sum to zero. For nonzero $q \in \mathbb{R}^{|\Pi|}$ with $\mathbf{1}^\top q = 0$, write $q^+_\pi = \max\{q_\pi, 0\}$ and $q^-_\pi = \max\{-q_\pi, 0\}$. Since $\|q^+\|_1 = \|q^-\|_1 = \|q\|_1/2$, define the two contrast-induced probability laws by

$$P^{q^\pm}_{R,Z} = \sum_{\pi \in \Pi} \frac{2q^\pm_\pi}{\|q\|_1} P^{\pi}_{R,Z}.$$

For probability measures P, Q, use the convention

$$\mathrm{TV}(P, Q) := \sup_A |P(A) - Q(A)| = \frac{1}{2} \int \left| \frac{\mathrm{d}P}{\mathrm{d}\lambda} - \frac{\mathrm{d}Q}{\mathrm{d}\lambda} \right| \mathrm{d}\lambda,$$

where $\lambda = P + Q$. The *history resolution* of q is

$$s_Z(q) := \mathrm{TV}(P^{q^+}_{R,Z}, P^{q^-}_{R,Z}).$$

For deterministic histories, if row r has K_r history-equivalence classes, let $D_r \in \{0,1\}^{|\Pi| \times K_r}$ have $(D_r)_{\pi k} = 1$ exactly when policy π belongs to history class k on row r. Then

$$s_Z(q) = \frac{\sum_r w_r \|D_r^\top q\|_1}{\|q\|_1}. \tag{3}$$

Writing Y for the scored outcome and $P^{\pi}_{R,Z,Y}$ for the joint law induced by policy π, the declared history is *score-sufficient for* q if there is one Markov kernel $L(\mathrm{d}y \mid r, z)$ such that $P^{\pi}_{R,Z,Y} = P^{\pi}_{R,Z} L$ for every $\pi \in \mathrm{supp}(q)$.

For fully specified deterministic update types, their count vector determines the composite update map—and hence the final state from any initial state—for every finite ordering exactly when the corresponding update maps commute pairwise (Appendix Lemma 1). This result concerns the final state: an intermediate answer can still depend on a prefix, and a scalar total count is coarser than a count for each update type.

Theorem 1 (History-resolution certificate)**.** *Assume a common initial state and coupled exogenous randomness (shared random inputs or seeds), the policy-independent row law in Definition 2, and that the declared history is score-sufficient for q in the sense of that definition. Let $P_Y^{q^\pm} = P_{R,Z}^{q^\pm} L$ be the induced outcome laws and define $\Delta_e^{\mathrm{exec}}(q) = \int y \, \mathrm{d}P_Y^{q^+}(y) - \int y \, \mathrm{d}P_Y^{q^-}(y)$. For a measurable summary h, let $h_\# P$ denote the distribution obtained by applying h to a draw from P. If $Y_e \in [a, b]$ almost surely, with $a < b$, then*

$$|\Delta_e^{\mathrm{exec}}(q)| \le (b-a) s_Z(q), \qquad \mathrm{TV}(h_\# P_{R,Z}^{q^+}, h_\# P_{R,Z}^{q^-}) \le s_Z(q) \tag{4}$$

for every measurable summary $h(R, Z)$.

Full proof. See Appendix F.3.

History resolution lies in $[0, 1]$. Zero means that the two weighted policy groups produce identical recorded-history mixtures and, under the theorem's assumptions, must have the same average score. One means that their recorded histories do not overlap, in which case the bound spans the full score range and does not constrain the effect magnitude. A complete derivation of Equation 3 appears in Appendix F.1. The probability facts are classical; the tracing procedure makes the bound available before outcomes are examined.

The appendix derives the minimum history resolution needed to support a specified score difference and extends the bound to a known amount of mismatch in the scoring process. These results require a defensible mismatch bound; they do not verify that the tracing rule captures every relevant path (Appendices F.4 and F.5).

2.3 Control Map records

Definition 3 (Control Map)**.** For each comparison, the Control Map keeps four kinds of evidence separate: the recorded histories and how well they distinguish the update plans, the estimated behavioral change and its uncertainty, acquisition, retention, and locality, and measured cost. Missing or invalid evidence is reported as UNDEFINED. To evaluate a conclusion in a new setting, we apply the same record there under an explicit mapping of the tested update plans. Appendix A.2 gives the complete record definitions.

Distinguishable histories rule out complete aliasing, but they neither determine effect direction nor replace the design assumptions needed for effect estimation.

3 Which Updates Reach Each Answer?

Nominally different schedules are not always distinguishable at an answer: short inputs permit fewer updates and therefore collapse more schedules to the same pre-answer history. Longer inputs expose progressively more of the schedule grid.

3.1 Formulation

For the Memory-as-Context (MAC) schedule experiment and an input containing $L \ge 1$ tokens, the number of complete writes available before scoring is

$$H(L) = \left\lfloor \frac{L + 256\lfloor (L-1)/512 \rfloor}{16} \right\rfloor, \quad a_p(H) = (\mathbf{1}\{t \bmod p = 0\})_{t=1}^{H}. \tag{5}$$

For $\Pi = \mathrm{Perm}(1, 2, 4, 8)$ and $s = (p_1, \ldots, p_4) \in \Pi$, let $z_H(s) = (a_{p_1}(H), \ldots, a_{p_4}(H))$. Exact enumeration yields

$$|\{z_H(s) : s \in \Pi\}| = \begin{cases} 1, & H = 0, \\ 4, & H = 1, \\ 12, & H \in \{2, 3\}, \\ 24, & H \geq 4. \end{cases} \tag{6}$$

For a pair $q = \mathbf{e}_s - \mathbf{e}_{s'}$, where $\mathbf{e}_s$ is the policy-coordinate basis vector for schedule s, $s_Z(q) = 1 - \sum_r w_r \mathbf{1}\{z_{r,s,e} = z_{r,s',e}\}$. The class count and pairwise-resolution identity are proved in Appendix F.6.

Appendix B reports which layer–period comparisons remain distinguishable when schedules collapse to the same history, together with the resulting rank and conditioning diagnostics. The corresponding estimability statement and proof appear in Appendix F.7.

3.2 Experimental protocol and results

The history audit uses a fixed 24-layer, approximately 760M-parameter Memory-as-Context evaluator (MAC-760M) and enumerates all 24 assignments of periods $\{1, 2, 4, 8\}$ to four six-layer depth bands on 27,072 rows from eight evaluation sets fixed before outcomes were examined (hereafter, *frozen*). Policies share a nominal update count but differ in pre-answer opportunity. Appendix Figure 3 reports both the exact class staircase and the benchmark-specific row partition.

Of the 5,153 LAMBADA (Paperno et al., 2016) rows, 5,141 distinguish all schedules. For a uniformly sampled row and one of the 276 schedule pairs, the probability of identical histories is 0.1285. A retrospective SocialIQA (Sap et al., 2019) check makes Theorem 1 concrete: each period-4 versus period-8 swap changes the declared pre-answer history on only one of 1,954 rows. Thus $s_Z = 1/1954$, any binary-accuracy contrast is bounded by 0.0512 percentage points, and all six observed contrasts are zero.

Live neural validation. On a randomly initialized two-layer Llama model updated with LoRA (Hu et al., 2022), the tracing procedure recovers the value $s_Z = 1/2$ specified in advance from 16 rows and rejects all 16 injected trace corruptions; validation passes again after the original traces are restored. Each row performs updates A then B, but B precedes scoring in 6/8 rows for one policy and 2/8 for the other. This controlled test was fixed locally before model construction. It validates the recorded nonzero-gradient paths in this setup, complete internal coverage of published editors, or automatic discovery of every relevant path.

Finally, an apparent dose–locality curve, where dose is the number of updates completed before the focal query, disappears after matching model, writer, strength, seed, family, the focal item's position, and complete event order: all 1,600 pairs have identical terminal-state hashes and locality scores. After matching, data no longer support interpreting dose as the cause of the terminal-locality pattern.

4 Behavioral Effects of Control Choices

For each policy comparison with sufficient support from the recorded-history design, we report the mean behavioral difference, its uncertainty, the inference unit, and when the analysis was specified. Unless stated otherwise, 95% ranges come from paired family- or block-level reweighting of one fixed evaluation set; they describe sensitivity to those units. In the editing studies, a family groups related rows for one target fact. Appendix B.6 gives the formal record.

4.1 Experimental protocol and results

Where and when updates occur can matter even when their total number is fixed, but the resulting pattern is specific to the tested setting. Figure 2 shows the complete equal-count placement surface and its additive decomposition.

At this checkpoint, the fitted additive effects of frequent updates are positive in the earliest and latest layer groups and negative in the middle groups. We treat this as a local placement–frequency

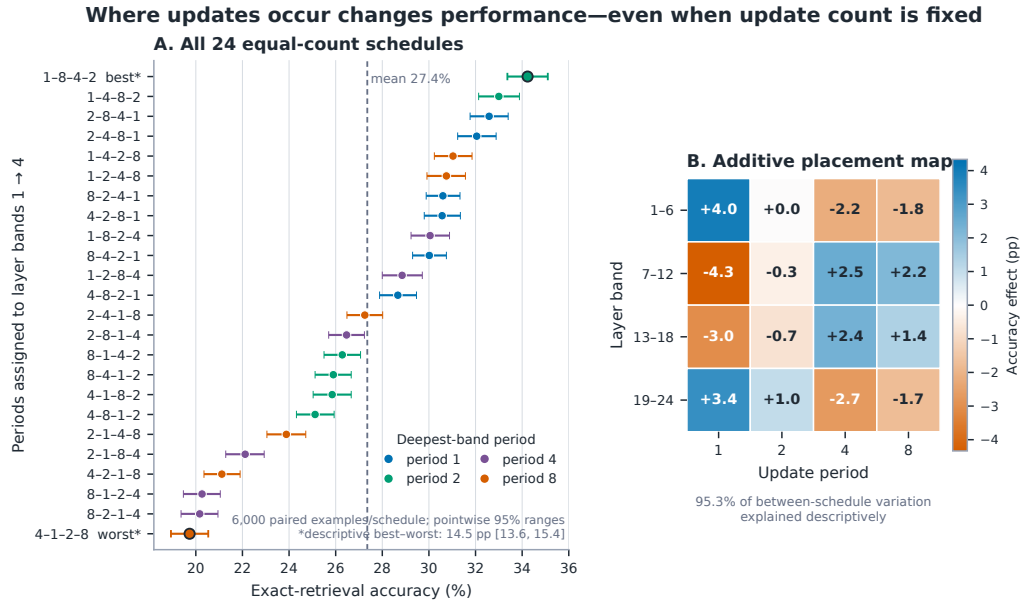


Figure 2: **Where updates occur changes performance even when update count is fixed.** All 24 assignments of periods $\{1, 2, 4, 8\}$ to four ordered layer bands are evaluated on the same 6,000-example single-needle exact retrieval (Single-NIAH) evaluation set. Schedule codes list periods from the earliest to the latest layer band; whiskers are pointwise paired 95% bootstrap ranges over the fixed evaluation set. The additive assignment map explains 95.3% of between-schedule variation descriptively (leave-one-schedule-out $R^2 = 0.863$, root-mean-square error $= 1.59$ percentage points). Best–worst labels and the decomposition are retrospective, so this is a local response surface rather than a universal schedule ranking.

interaction rather than a universal layer rule. Table 1 summarizes this placement result together with the timing and dose comparisons.

Table 1: **Controlled comparisons of placement, timing, and dose.** Each row changes one choice while preserving the listed controls; conclusions remain specific to the tested setting.

Choice	Comparison and controls	Observed result	Supported takeaway
Placement	24 band–period assignments; one MAC-760M checkpoint; 6,000 paired Single-NIAH rows; equal prompt-only writes	Accuracy 19.733%–34.233%; the band–period model descriptively explains 95.322% of between-schedule variation	Placement matters in this task and checkpoint;
Timing	Same selected events and terminal actions; writes before versus after the focal query	Immediate-minus-deferred acquisition: $+0.1196$, $+0.0598$, $+0.4320$, and $+0.1688$ nats across four GPT-2 sizes (Radford et al., 2019); Pythia-410M (Biderman et al., 2023): $+0.1592$ nats, range $[-0.4484, +0.7667]$	The estimated effect is positive in all four GPT-2 assays; the Pythia range crosses zero
Dose	Eight objectives, eight actions, 32 stochastic-gradient-descent steps, and terminal adapter–optimizer construction fixed	Across 18 deliberately varied LoRA settings, the best tested doses $1/2/4/8$ occur in $9/7/2/0$ settings; the best-dose prediction specified before outcomes is confirmed in 7/18 settings	The best tested dose varies across settings

5 FEASIBILITY OF ACQUIRED MEMORY

5.1 DECISION RULES

We evaluate acquisition, retention, and locality separately using thresholds specified for each study. We orient each decision metric so that larger values are better; for locality, this means using the negative of measured damage. A valid uncertainty range lies above its threshold for PASS, below it for FAIL, and across it for INCONCLUSIVE; an absent or invalid measurement is UNDEFINED. The retention check also requires that at least 20% of edit pairs meet the immediate-learning condition (installation) and that at least half of those installed edits are retained. The overall memory-quality decision passes only when every required check passes, so strong acquisition cannot compensate for failed retention or locality. Cost is reported separately and becomes a requirement only when a budget was specified. The complete decision rules appear in Appendix C.1.

5.2 EXPERIMENTAL PROTOCOL AND RESULTS

Immediate learning can overstate memory quality: a method may acquire the requested information yet fail to retain it or preserve unrelated behavior. The LoRA study measures query-time acquisition, terminal retention, and terminal locality separately. Across six selected CounterFact (Meng et al., 2022) settings, acquisition passes 6/6 while retention and locality pass 0/6; the prespecified four-setting Stage-0 screen passes in 0/4 and triggers its stopping rule (Appendix C.2). A study of published editors exposes the same logical gap: ROME passes acquisition and retention at all 13 operating points, but locality and therefore full feasibility pass at 0/13 points. A separate shared-cohort comparison of dose-one writer profiles appears in Appendix Figure 4; acquisition–locality tradeoffs and dose profiles appear in Appendix Figure 5.

A separate 160-family positive-control evaluation set shows that the criteria are attainable. MEMIT-high has acquisition 8.30 nats and locality damage 2.08 percentage points and passes the full CounterFact decision rule. ROME-low has acquisition 7.52 nats and damage 4.64 percentage points, but its damage range $[3.08, 6.42]$ crosses the threshold. The MEMIT-minus-ROME damage contrast is -2.56 percentage points $[-4.27, -1.06]$. Because the dose-four query is terminal, its terminal-recall measurement is taken at zero lag; installation and retention among installed edits remain separate checks. In a separate shared-cohort eight-edit sequence, AlphaEdit meets all declared memory-quality criteria at each of the four tested doses for these setting-specific operating points.

The five-percentage-point locality boundary is a tolerance specified before outcomes. Exact threshold analysis (Appendix A.6) places the CounterFact pass breakpoints at 3.20 percentage points for MEMIT-high and 6.42 for ROME-low, versus 52.73 and 50.78 on WikiRecent. Relaxing locality cannot rescue LoRA because retention passes in 0/6 settings. New deployments should specify setting-specific thresholds before examining outcomes, using an explicit tolerated loss; every required component must pass.

6 GENERALIZATION ACROSS MODELS AND BENCHMARKS

6.1 COMPARING SETTINGS

A *bundle* is a fixed model and editor configuration. To ask whether a result carries to another bundle or benchmark, we specify in advance how the tested update plans correspond across the two settings and preserve which plan is the positive side of the comparison. Within each setting, a valid uncertainty range entirely above zero gives a positive sign, one entirely below zero a negative sign, and one crossing zero an unresolved sign. These endpoint-specific measurements are reported separately rather than combined into a single utility. Data from one setting do not determine an unevaluated setting without an additional invariance assumption, so our strongest claims use direct evaluation in the new setting. Appendix D.1 gives the formal definitions and the off-support construction.

6.2 RESULTS ACROSS TWO MODEL BUNDLES

The relative performance of the tested editors varies across the evaluated model–benchmark–question bundles; a source-setting ranking is therefore not something that should be assumed to carry over.

GPT-2XL source. With checkpoint, editor strengths, execution, dose, and estimator fixed, the MEMIT-minus-ROME acquisition contrast is positive on all three endpoints ($+0.7850, +0.5772, +1.2930$ nats). Its locality-success contrast changes from $+2.5625$ percentage points on CounterFact to -7.5000 on zsRE (Levy et al., 2017; Zhang et al., 2024) and -15.9180 on WikiData-CounterFact (Zhang et al., 2024); the evaluation sets and questions differ.

Appendix Figure 6 displays the complete endpoint trajectories with separate pointwise ranges for each fixed evaluation set.

Llama-3.2-3B follow-up. Using Llama-3.2-3B (Meta AI, 2024b), the analysis was fixed before outcomes. A later reconstruction of the recorded editor-call sequences gives terminal $s_Z = 1$, although it neither inspects editor internals nor verifies the cross-model policy mapping. Ordered as CounterFact, zsRE, and WikiData-CounterFact, the predicted $(+, -, -)$ locality pattern does not recur: the result is $(+, +, ?)$. Acquisition favors ROME on all three endpoints $(-0.3017, -0.2116, -0.2247)$ nats; both editors fail every locality requirement.

Prospective WikiRecent. The only cross-setting workflow for which the policy mapping, tracing procedure, outcome measures, and decision rules were all fixed before setup validation or outcomes were inspected uses 64 WikiRecent (Zhang et al., 2024) families, four cyclic orders, and two GPT-2XL policies. All 512 editor-call sequences match their specifications. MEMIT's gain after four edits relative to the reset baseline exceeds ROME's by $+0.2142$ nats per target token on canonical prompts and $+0.7178$ on a rephrase, satisfying the prespecified requirement that MEMIT's mean gain exceed ROME's. Yet both editors fail the two-probe locality requirement (-43.36 and -41.41 percentage points), so the overall evidence requirement fails.

Alternative outcome measures. With rows, model states, probes, and bootstrap draws fixed, binary success, full-vocabulary Kullback–Leibler divergence, and greedy-token change all favor MEMIT at the source setting when each measure is oriented so that positive means better performance. Complete results and scope appear in Appendix A.6.

6.3 What the combined evidence supports

The four questions require separate evidence. Update counts must be interpreted through the recorded histories; acquisition, retention, and locality must be tested separately; and claims about a new setting require histories and outcomes from that setting. Distinct histories do not determine the direction of an effect, and a relative ranking does not show that either method meets the absolute memory-quality requirements.

Appendix Table 25 summarizes four common shortcuts and the corresponding evaluation alternatives.

Limitations. The guarantees depend on the stated tracing rule. Our checks verify that the recorded event sequences follow their specifications, if a relevant path is omitted, Theorem 1 does not apply. The cross-setting grid also changes the model, benchmark, and test questions together, so it demonstrates heterogeneity without identifying its cause. Detailed conformance counts, robustness bounds, timing labels, and reproducibility fields appear in Appendices A.6 and E.

7 Conclusion

Update count is not enough to describe test-time adaptation. Evaluations should record the updates that precede each answer and then separately test behavioral effects, immediate learning, later recall, changes to unrelated behavior, and performance in new settings. Our results show that these dimensions can disagree: choices hidden by update count can matter, an apparent count effect can disappear after matching the actual update sequence, and immediate success does not guarantee retention or preservation of unrelated behavior. A method that succeeds immediately or ranks first in one setting should therefore not be treated as broadly effective without evidence for retention, locality, and the intended target setting.

AI USE STATEMENT

Generative AI tools assisted with literature discovery; drafting and editing manuscript text; drafting and debugging research code; cleaning and reformatting data; constructing synthetic evaluation panels; preparing analyses, tables, and figures; and auditing and assisting code writing, math proof, statistical and textual claims. The authors determined the research questions, experimental designs, estimands, decision rules, and final interpretations. All AI-assisted outputs were reviewed, and claim-bearing computations were tested or independently reproduced against versioned, hash-bound artifacts. AI-generated suggestions were not treated as evidence. The authors take full responsibility for the paper.

ETHICS STATEMENT

This work evaluates failure modes in model memory and editing; it involves no human participants and no deployment. Such mechanisms can store false or sensitive associations and damage unrelated behavior, so we report locality and retention failures rather than treating target recall as sufficient. We restrict claims to tested systems and condition artifact release on checks for anonymity, privacy, redistribution rights, and third-party licenses.

REPRODUCIBILITY STATEMENT

The paper specifies estimands, units, schedules, decision rules, tracing rules, and claim boundaries. Our reproducibility bundle includes an implementation that constructs finite deterministic histories, displays rendered from hash-bound aggregate projections, and sanitized WikiRecent sufficient statistics. Offline tests recompute structural outputs and the complete WikiRecent record. For one CPU run, the source and protocol were recorded in version control before execution; a separate implementation checks its history mixtures and corruption controls. A second CPU-only audit recomputes the finite omitted-path example, decision-threshold breakpoints, and structural LoRA history tests. Legacy family-level measurements are omitted, so their intervals are not independently recomputed. This is neither a raw-data release nor a full model reproduction and does not expose third-party editor internals. After de-anonymization, we plan to release redistributable author-created code, protocols, aggregates, and validators; third-party data and checkpoints retain their original licenses.

REFERENCES

Shishir Adhikari, Sourav Medya, and Elena Zheleva. Learning exposure mapping functions for inferring heterogeneous peer effects. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bXYp5AIMju.

Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017. doi: 10.1214/16-AOAS1005.

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. Nested learning: The illusion of deep learning architectures. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46968–47002, 2025a. doi: 10.52202/085713-1565.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113506–113543, 2025b. doi: 10.52202/085713-3786.

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. It's all connected: A journey through test-time memorization, attentional bias, retention, and online optimization. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2504.13173.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine*

Learning, volume 202 of *Proceedings of Machine Learning Research*, 2023. URL https://arxiv.org/abs/2304.01373.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

Qizhou Chen, Dakan Wang, Taolin Zhang, Zaoming Yan, Chengsong You, Chengyu Wang, and Xiaofeng He. UniEdit: A unified knowledge editing benchmark for large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.12345.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, pp. 2924–2936, 2019. doi: 10.18653/v1/N19-1300.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018.

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/.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HvSytvg3Jh.

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

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

Jinwu Hu, Zitian Zhang, Guohao Chen, Xutao Wen, Chao Shuai, Wei Luo, Bin Xiao, Yuanqing Li, and Mingkui Tan. Test-time learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24823–24849, 2025. URL https://proceedings.mlr.press/v267/hu25z.html.

Martin Huber, Jannis Kueck, and Mara Mattes. Learning and testing exposure mappings of interference using graph convolutional autoencoders. In *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pp. 656–675, 2026. URL https://proceedings.mlr.press/v323/huber26a.html.

Bogdan Korel and Janusz Laski. Dynamic program slicing. *Information Processing Letters*, 29(3): 155–163, 1988. doi: 10.1016/0020-0190(88)90054-3.

Seongmin Lee, Dave Binkley, Robert Feldt, Nicolas Gold, and Shin Yoo. Causal program dependence analysis. *arXiv preprint arXiv:2104.09107*, 2021. URL https://arxiv.org/abs/2104.09107.

Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=r8juz2t749J.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pp. 333–342, 2017. doi: 10.18653/v1/K17-1034. URL https://aclanthology.org/K17-1034/.

Wei Liu, Haomei Xu, Hongkai Liu, Zhiying Deng, Ruixuan Li, Heng Huang, Yee Whye Teh, and Wee Sun Lee. Are we evaluating the edit locality of LLM model editing properly? *arXiv preprint arXiv:2601.17343*, 2026. URL https://arxiv.org/abs/2601.17343.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 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 *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Meta AI. Llama 3.2 1B model card, 2024a. URL https://huggingface.co/meta-llama/Llama-3.2-1B.

Meta AI. Llama 3.2 3B model card, 2024b. URL https://huggingface.co/meta-llama/Llama-3.2-3B.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1525–1534, 2016. doi: 10.18653/v1/P16-1144.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014. doi: 10.1214/14-STS486.

Sebastian Pohl, Max Ploner, and Alan Akbik. Towards a principled evaluation of knowledge editors. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pp. 47–60, 2025. doi: 10.18653/v1/2025.l2m2-1.4. URL https://aclanthology.org/2025.l2m2-1.4/.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of EMNLP-IJCNLP*, pp. 4463–4473, 2019. doi: 10.18653/v1/D19-1454.

Fredrik Sävje. Causal inference with misspecified exposure mappings: Separating definitions and assumptions. *Biometrika*, 111(1):1–15, 2024. doi: 10.1093/biomet/asad019. URL https://doi.org/10.1093/biomet/asad019.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366, 2021. URL https://proceedings.mlr.press/v139/schlag21a.html.

Gihyeon Sim. When to ponder: Adaptive compute allocation for code generation via test-time training. *arXiv preprint arXiv:2601.00894*, 2026. URL https://arxiv.org/abs/2601.00894.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57503–57522, 2025. URL https://arxiv.org/abs/2407.04620.

Arnuv Tandon, Karan Dalal, Xinhao Li, Daniel Koceja, Marcel Rød, Sam Buchanan, Xiaolong Wang, Jure Leskovec, Sanmi Koyejo, Tatsunori Hashimoto, Carlos Guestrin, Jed McCaleb, Yejin Choi, and Yu Sun. End-to-end test-time training for long context. *arXiv preprint arXiv:2512.23675*, 2025. URL https://arxiv.org/abs/2512.23675.

Zilu Tian and Liying Liu. Provenance tracking in AI compilers through the lens of coalgebra. *arXiv preprint arXiv:2606.10937*, 2026. URL https://arxiv.org/abs/2606.10937.

Tyler J. VanderWeele and Miguel A. Hernán. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013. doi: 10.1515/jci-2012-0002.

Tomoya Wakayama. A decision-theoretic view of test-time training: When, how far, and which directions to adapt. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2606.15569.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 82–93, 2024. doi: 10.18653/v1/2024.acl-demos.9. URL https://aclanthology.org/2024.acl-demos.9/.

Mark Weiser. Program slicing. In *Proceedings of the 5th International Conference on Software Engineering*, pp. 439–449, 1981.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, pp. 115491–115522, 2024. doi: 10.52202/079017-3668.

Junsang Yoon, Akshat Gupta, and Gopala Anumanchipalli. Is bigger edit batch size always better?—an empirical study on model editing with Llama-3. *arXiv preprint arXiv:2405.00664*, 2024. URL https://arxiv.org/abs/2405.00664.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472.

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. *arXiv preprint arXiv:2401.01286*, 2024. URL https://arxiv.org/abs/2401.01286.

Ziqi Zhang, Yuanchun Li, Yao Guo, Xiangqun Chen, and Yunxin Liu. Dynamic slicing for deep neural networks. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 838–850, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-7043-1. doi: 10.1145/3368089.3409676. URL https://doi.org/10.1145/3368089.3409676.

A AUDIT CONTRACT AND MATHEMATICAL BOUNDARY

This appendix is organized by the four evidentiary records used in the main paper. Section A gives complete definitions, the classical resolution calculus, and the adapter-robustness and gate-threshold audits; Section B gives execution and effect details; Section C reports absolute-gate evidence; Section D reports endpoint transport; and Section E states the reproducibility boundary. Only material supporting a main-text claim is included.

The operational schematic appears as Figure 1 in the main paper. Here LoRA denotes low-rank adaptation and SGD denotes stochastic gradient descent.

Table 2: **Claim-to-evidence index.** Each row identifies the study, analysis timing, inference or enumeration unit, and fastest route to the supporting record. Assays answer separate questions and are not pooled into a joint response surface. "Frozen" means internally hash-bound before the named outcome was accessed, not publicly preregistered.

Claim / audit question	Study	Timing	Unit and scale	Location
Equal counts need not identify equal executed treatments	MAC-760M trace atlas; 24 depth–cadence schedules on eight panels	Outcome-blind exact enumeration	27,072 benchmark rows; 24 policies per row	Section B; Table 6
Layer placement can change accuracy at fixed update count	MAC-760M placement surface on Single-NIAH	One fixed checkpoint; additive summary descriptive	6,000 paired rows; all 24 schedules	Figure 2; Section B.4
Order can masquerade as dose–locality structure	GPT-2XL LoRA complete-order correction	Post-hoc deterministic correction	1,600 order-matched pairs	Section B
Query-relative phase can change acquisition	GPT-2 family and Pythia-410M phase studies	Per-model frozen gates; cross-model synthesis descriptive	Four LoRA initializations per model	Section B
The best tested dose varies across settings	LoRA/SGD four-dose grids	Purposive, partly adaptive setting collection; within-setting calls frozen	18 settings; four receiver seeds per setting	Section B.5
Acquisition need not imply feasible memory	LoRA and ROME grids; fixed MEMIT, ROME, and AlphaEdit controls	Cohort-specific frozen or declared timing	Family/block resampling within each assay	Table 7; Sections C.2–C.10
Observed editor ordering varies across evaluated bundles	GPT-2XL and Llama-3.2-3B across CounterFact, zsRE, and WikiData-CF	Endpoints separately frozen; Llama follow-up fixed before Llama outcomes	160/160/128 families; four orders per bundle	Sections D.3 and D.4
Strict prospective end-to-end verdict	GPT-2XL MEMIT-high/ROME-low on WikiRecent	Full workflow frozen before preflight or outcomes	64 families; four Latin orders; 512 policy rows	Section D.2
Whether the tracing procedure survives live neural mutation and tampering	Random two-layer Llama with PEFT-LoRA and two SGD writes	Source and protocol committed locally before model construction; not externally preregistered	16 policy rows; two history classes; 16 corruptions plus restoration	Section A.6
How conclusions vary with hidden paths, adapter scope, and gate tolerance	Finite hidden-cache witness; adapter-conformance audits; exact breakpoint analysis	Deterministic post-hoc robustness audits	Exact finite witness, recorded traces, and frozen gate records	Section A.6

A.1 SECONDARY RESOLUTION GUARANTEES

Lemma 1 (When counts are sufficient)**.** *Fix a finite atom alphabet $\mathcal{A}$ and a state space $\mathcal{X}$. For each $a \in \mathcal{A}$, let $T_a : \mathcal{X} \to \mathcal{X}$ be a deterministic self-map. For a word $\Gamma = a_1 \cdots a_T$, set $F(\Gamma) = T_{a_T} \circ \cdots \circ T_{a_1}$ and let $\nu(\Gamma)$ be its atom-count vector, with $F(\varnothing) = \mathrm{id}_{\mathcal{X}}$. Over all finite words, there exists $Q : \mathbb{N}_0^{|\mathcal{A}|} \to \{f \mid f : \mathcal{X} \to \mathcal{X}\}$ such that*

$$F(\Gamma) = Q(\nu(\Gamma)) \quad \textit{for every } \Gamma \tag{7}$$

if and only if the maps T_a commute pairwise.

The proof appears in Appendix F.2.

Remark 1 (Final-state scope). An intermediate query can depend on a prefix even when final maps commute, and scalar total count is coarser than ν. On a restricted word catalogue, counts may also suffice accidentally without global commutativity.

Corollary 1 (Resolution floor for a target effect)**.** *Under Theorem 1 with $a < b$, any claim that the realized assignment effect satisfies $|\Delta_e^{\mathrm{exec}}(q)| \geq \delta > 0$ requires $s_Z(q) \geq \delta/(b-a)$. Thus $s_Z(q) < \delta/(b-a)$ establishes before outcome access that the target effect is unattainable under the theorem's assumptions.*

The proof appears in Appendix F.4.

Proposition 1 (Robustness to conditional-kernel mismatch)**.** *Let $X = (R, Z)$, let $K_+(\mathrm{d}y \mid x)$ and $K_-(\mathrm{d}y \mid x)$ be the actual outcome kernels in the two contrast arms, and let $L(\mathrm{d}y \mid x)$ be any common reference kernel. Define $P_Y^{q^\pm} = P_X^{q^\pm} K_\pm$ and $\epsilon_\pm = \int \mathrm{TV}(K_\pm(\cdot \mid x), L(\cdot \mid x))\, \mathrm{d}P_X^{q^\pm}(x)$. Let $\Delta_e^{\mathrm{exec}}(q)$ be the difference between the means under these two output laws. If $Y_e \in [a, b]$ almost surely under both laws, with $a < b$, and the two pointwise total-variation functions in the definition of $\epsilon_\pm$ are measurable, then*

$$\mathrm{TV}(P_Y^{q^+}, P_Y^{q^-}) \le \min\{1, s_Z(q) + \epsilon_+ + \epsilon_-\}, \qquad |\Delta_e^{\mathrm{exec}}(q)| \le (b-a)\min\{1, s_Z(q) + \epsilon_+ + \epsilon_-\}. \tag{8}$$

The proof appears in Appendix F.5. This extension is informative only when an external bound on ϵ_+ and ϵ_- is defensible; it does not verify that the tracing rule is complete. The finite hidden-cache experiment in Appendix A.6 illustrates the boundary exactly.

A.2 The complete typed evidence record

We use *context* and *setting* synonymously in this appendix; c denotes the setting defined in the main text. The common object across our studies is an evidence record, not a shared response surface. Effect is a relative, contrast-level property; feasibility is an absolute, policy-level property; and transport is a relation between two context-specific records. The framework combines them only through a typed conjunction. It never pools estimates across assays whose outcome instruments are not commensurable. Write e_c for the endpoint component of context c.

Let q_c be a nonzero, zero-sum contrast over the evaluated policy catalogue Π_c^{eval}. A named estimand m fixes the intended treatment-version semantics, outcome transform, aggregation, orientation, and endpoint; u_c^m is its inference unit. Its local effect record is

$$\mathsf{Eff}_c^m(q_c) = \big(m, \widehat{\Delta}_c^m(q_c), I_c^m(q_c), H_c^m(q_c), u_c^m, \mathsf{Asn}_c^m, \mathsf{Int}_c^m\big), \tag{9}$$

where I is the declared uncertainty interval, $H \in \mathcal{S}$ is the protocol's effect decision, $\mathsf{Asn}_c^m \in \{\text{PROSPECTIVE}, \text{DESCRIPTIVE}, \text{DIAGNOSTIC}\}$ records when the analysis was fixed, and $\mathsf{Int}_c^m \in \{\text{VALID}, \text{INVALID}\}$ records provenance and integrity separately. Superscripts Z, K, T, and ι, and the gate subscript j, use these same timing and integrity vocabularies. Here $\mathcal{S} = \{\text{PASS}, \text{FAIL}, \text{INCONCLUSIVE}, \text{UNDEFINED}\}$. Its conjunction operator $\bigwedge_{\mathcal{S}}$ uses the precedence FAIL $>$ UNDEFINED $>$ INCONCLUSIVE $>$ PASS. We write $\perp$ for a field that was not measured or not declared, and $\mathcal{M}_c$ for the set of named estimands in context c.

Let $\mathcal{D}_{c,e_c}$ be the declared score-dependency adapter and Cert_{c,e_c} its source/input binding. The resolution record

$$\begin{aligned}\mathsf{Res}_{c,e_c}(q_c) = \big(&\mathcal{D}_{c,e_c}, \{w_r, z_{r,s,e_c}\}_{r,s}, \mathsf{Alias}_{c,e_c}, \mathsf{Exposure}_{c,e_c},\\ &s_{Z,c,e_c}(q_c), S_{c,e_c}^Z(q_c), \mathsf{Asn}_{c,e_c}^Z, \mathsf{Int}_{c,e_c}^Z, \mathsf{Cert}_{c,e_c}\big)\end{aligned} \tag{10}$$

retains the bound traces (or their digests), row weights, aliases, exposure, resolution, protocol-specific support status $S^Z \in \mathcal{S}$, timing, integrity, and certificate. A typed gate record is $G_{c,j}(\pi) = (j, \psi_{c,j}, \widehat{\nu}_{c,j}(\pi), I_{c,j}(\pi), S_{c,j}(\pi), \mathsf{Asn}_{c,j}(\pi), \mathsf{Int}_{c,j}(\pi))$, containing the gate identity, declared rule, estimate, interval, and status $S_{c,j}(\pi) \in \mathcal{S}$ together with its timing and integrity. By construction, $S_{c,j} = \text{UNDEFINED}$ when a required measurement or its integrity record is missing or invalid; a valid but nonprospective gate is FAIL whenever prospective timing is part of the declared rule. Retain the full vector of typed gate records $\mathbf{G}_c(q_c) = (G_{c,j}(\pi))_{\pi \in \mathrm{supp}(q_c), j}$ rather than averaging it. A separate cost record $K_c(\pi) = (\boldsymbol{\kappa}_c(\pi), \mathbf{u}_c^K, g_c, \mathcal{B}_c, \mathsf{Asn}_c^K, \mathsf{Int}_c^K)$ reports named cost coordinates and units, an optional prespecified scalarization g_c, a finite protocol budget or $\mathcal{B}_c = \perp$, and cost-record integrity. When $\mathcal{B}_c < \infty$, $g_c(\boldsymbol{\kappa})$ must have the budget's units; when no scalar budget was declared, $g_c = \mathcal{B}_c = \perp$. A vector such as outer-call counts is not silently scalarized or compared across writers. Let $\mathbf{K}_c(q_c) = (K_c(\pi))_{\pi \in \mathrm{supp}(q_c)}$. The local evidence record is

$$\Gamma_c^m(q_c) = \big(\mathsf{Res}_{c,e_c}(q_c), \mathsf{Eff}_c^m(q_c), \mathbf{G}_c(q_c), \mathbf{K}_c(q_c)\big). \tag{11}$$

It exists without a target context and can therefore represent the paper's resolution, effect, and feasibility studies on their own terms. For a resolution-only study with no outcome measurement,

set $\widehat{\Delta} = I = \perp$ and $H =$ UNDEFINED; $\perp$ records absence, not a numerical zero. When several named estimands are reported, the empirical record is the bundle $(\Gamma_c^m)_{m\in\mathcal{M}_c}$; it is not collapsed to one scalar. Formally, the base *Control Map* is the partial function

$$\mathfrak{C} : (c, q_c, m_c) \;\rightharpoonup\; \Gamma_c^{m_c}(q_c), \tag{12}$$

whose domain requires a well-formed declared contrast, dependency adapter, and valid trace provenance. Zero resolution remains in the domain. Under the assumptions of Equation 4, its positive and negative trace mixtures coincide and the realized assignment contrast is zero; individual policies need not be pairwise aliased. If the intended estimand requires distinct implemented mixtures, its effect field is UNDEFINED rather than a null estimate. Policy-level feasibility fields may still be evaluated. Positive contrast support is required only before estimating the intended effect.

Transport is an optional extension of two local records and requires an explicit policy correspondence. Before target outcomes are opened, define

$$\iota_{c\to c^*} : \mathrm{supp}(q_c) \longrightarrow \Pi_{c^*}^{\mathrm{eval}}, \qquad q_{c^*} = \iota_{c\to c^*}\#q_c, \tag{13}$$

where the pushforward sums source coefficients mapped to each target policy. The correspondence is admissible only when q_{c^*} remains nonzero and the predeclared positive-versus-negative orientation of the contrast is preserved; otherwise transport is UNDEFINED. Its certificate records which model, writer, support, strength, execution template, and endpoint fields are held fixed or changed. The typed correspondence record $\mathsf{Iota}_{c\to c^*} = (\iota, \mathsf{Cert}^\iota, \mathsf{Asn}^\iota, \mathsf{Int}^\iota, S^\iota)$ also records whether the mapping was fixed before target outcomes and whether its provenance is valid. If it is absent, the target contrast is not silently inferred from matching names.

Define the target local record analogously and let $\mathbf{R}_{c\to c^*}(q_c, q_{c^*}) = (s_{Z,c,e_c}(q_c), s_{Z,c^*,e_{c^*}}(q_{c^*}))$. With $\mathbf{m} = (m_c, m_{c^*})$, let the transport record be $\mathsf{Tr}^{\mathbf{m}}_{c\to c^*} = (\rho, T, \mathsf{Asn}^T, \mathsf{Int}^T)$, where ρ is the declared relational rule and $T \in \mathcal{S}$ its result. The transport-extended record is

$$\begin{aligned}\Gamma^{\mathbf{m},\iota}_{c\to c^*}(q_c) = \big(&\mathsf{Iota}_{c\to c^*}, \Gamma_c^{m_c}(q_c), \Gamma_{c^*}^{m_{c^*}}(q_{c^*}), \mathbf{R}_{c\to c^*}(q_c, q_{c^*}),\\ &\mathsf{Tr}^{\mathbf{m}}_{c\to c^*}(q_c, q_{c^*})\big).\end{aligned} \tag{14}$$

The optional transport lift is

$$\mathfrak{C}^{\to} : (c, c^*, q_c, \mathsf{Iota}, \mathbf{m}) \;\rightharpoonup\; \Gamma^{\mathbf{m},\iota}_{c\to c^*}(q_c). \tag{15}$$

Its domain requires well-formed source and target contexts and available, valid provenance. The value $\mathsf{Iota} = \perp$ is admitted only to represent an attempted transport claim with $S^\iota = T =$ UNDEFINED; no mapping is inferred from matching names. Missing measurements remain typed UNDEFINED fields rather than being imputed, pooled, or interpreted as zero. Thus local records can pass, fail, be inconclusive, or be unavailable without requiring a target, while the transport lift records whether its additional obligations are met. The transport field $T \in \mathcal{S}$ applies the recorded relational rule to the two local records, such as preservation of direction under the certified policy correspondence. If the correspondence, target trace, required target outcome, or required instrument is absent, T is UNDEFINED. The tuple is deliberately typed: $\mathbf{R}$ is design geometry, each Eff is an estimate plus decision and provenance, $\mathbf{G}$ contains absolute per-policy gates, $\mathbf{K}$ contains cost coordinates and any declared budget, and T is relational. Addition or averaging across these fields has no defined meaning.

Let $\mathcal{R}^{\mathrm{pass}} \subseteq (0,1]^2$ be the protocol-declared set of source–target resolution pairs that license its transport analysis. For the strongest prospective transport claim, let $S^\iota =$ PASS when the correspondence is present, admissible, prospectively fixed, and valid; FAIL when it is valid but not prospectively fixed; and UNDEFINED when it is absent, inadmissible, or invalid. Let $S^R =$ PASS when both resolution records are valid and prospective and their positive pair lies in $\mathcal{R}^{\mathrm{pass}}$; UNDEFINED when either record is unavailable, invalid, or has zero resolution; and FAIL when the positive records are valid but retrospective or their pair lies outside $\mathcal{R}^{\mathrm{pass}}$. Finally, let $S_c^{P,m} =$ PASS when $\mathsf{Int}_c^m =$ VALID and $\mathsf{Asn}_c^m =$ PROSPECTIVE, FAIL for a valid but nonprospective record, and UNDEFINED for invalid or missing provenance. Define $S^{T,P}$ by preserving T when its rule is valid and prospectively fixed, setting it to FAIL when valid but nonprospective, and to UNDEFINED when the rule or integrity record is absent or invalid. Then define

$$\begin{aligned}&\mathsf{TransportSupported}^{\mathbf{m}}_{c\to c^*}(q_c)\\ &\quad= \bigwedge_{\mathcal{S}}\Big(S^\iota, S^R, S_c^{P,m_c}, S_{c^*}^{P,m_{c^*}},\\ &\qquad\qquad H_c^{m_c}(q_c), H_{c^*}^{m_{c^*}}(q_{c^*}), S^{T,P}_{c\to c^*}(q_c, q_{c^*})\Big).\end{aligned} \tag{16}$$

This is a deliberately strong predicate, not a declaration that descriptive or diagnostic records are scientifically useless.

For a policy $\pi^\dagger \in \Pi_{c^*}^{\mathrm{eval}}$ considered after target evaluation, let $\mathcal{J}_{c^*}^{\mathrm{req}}$ index the protocol's required behavioral gates. Let $S_{c^*}^K(\pi) = \textsc{pass}$ when $\mathcal{B}_{c^*} = \bot$ (cost is then neutral, not certified) or when a valid, prospectively fixed finite-budget record satisfies $g_{c^*}(\boldsymbol{\kappa}_{c^*}(\pi)) \le \mathcal{B}_{c^*}$; let it be FAIL for a valid prospective budget exceedance or a valid but retrospectively selected budget, and UNDEFINED when a required cost record is absent or invalid. Define absolute target feasibility by

$$\mathsf{TargetFeasible}_{c^*}(\pi^\dagger) = \bigwedge_{\mathcal{S}} \big(S_{c^*}^K(\pi^\dagger), \{S_{c^*,j}(\pi^\dagger) : j \in \mathcal{J}_{c^*}^{\mathrm{req}}\}\big)\,. \tag{17}$$

Thus cost is always a reported coordinate but is a pass/fail requirement only for a protocol that declares a finite budget and a matching scalar cost coordinate or prespecified scalarization. A resolved failure suffices to reject the conjunction; otherwise a missing required component propagates UNDEFINED; otherwise any unresolved component propagates INCONCLUSIVE. The empirical "full gate" in this paper sets $\mathcal{B}_c = \bot$ and conjoins acquisition, retention, and locality; it must not be read as satisfying an unstated deployment budget. For a pairwise target contrast, let $\mathrm{cand}(q_{c^*}) \in \mathrm{supp}(q_{c^*})$ be the arm prospectively designated as positive/preferred; a more general contrast requires an equally explicit designation rule. Let $S^{\mathrm{cand}} = \textsc{pass}$ when $\pi^\dagger$ is that designated candidate, UNDEFINED when no designation exists, and FAIL when another policy is substituted. The complete evidence condition is

$$\begin{aligned}&\mathsf{EvidenceComplete}_{c\to c^*}^{\mathbf{m}}(q_c, \pi^\dagger)\\ &\quad= \bigwedge_{\mathcal{S}} \big(S^{\mathrm{cand}}, \mathsf{TransportSupported}_{c\to c^*}^{\mathbf{m}}(q_c),\\ &\qquad\qquad \mathsf{TargetFeasible}_{c^*}(\pi^\dagger)\big)\,.\end{aligned} \tag{18}$$

This adjudicates one evaluated target point after target evaluation; it is not an online deployment decision. A comparator's infeasibility does not automatically make the candidate infeasible: the record retains every arm's absolute gates, while Equation 17 evaluates $\pi^\dagger$ itself. Only the all-PASS conjunction is collapsed. Mixed FAIL, INCONCLUSIVE, and UNDEFINED components remain visible because they answer different questions.

Operational recipe. The formalism induces a five-step evaluation procedure. (1) Freeze the candidate catalogue, endpoint, contrasts, dependency adapter, gates, and source-to-target correspondence. (2) Compile traces before opening outcomes and mark zero-resolution contrasts unavailable. (3) Estimate effects only for supported contrasts, retaining uncertainty and analysis timing. (4) Evaluate acquisition, retention, and locality as separate absolute gates for each candidate, record cost separately, and apply it as a gate only if a budget was declared. (5) Test the frozen relational transport rule in the target setting and otherwise return the unresolved or undefined component rather than a winner. This is the executable product of the Control Map; the experiments below instantiate each step on different controlled slices.

Why the types do not collapse. The separation is not cosmetic. Resolution can only make a contrast represented for estimation; it does not establish causal identification, and the bound in Equation 4 permits a zero effect at any positive resolution. An acquisition effect need not be feasible: the tested LoRA writer acquires its targets while missing the joint retention/locality gate. Source feasibility need not yield target stability: fixed MEMIT passes the named CounterFact criterion while its locality advantage over ROME reverses on both tested KnowEdit endpoints. These are three different failed implications— resolution $\not\Rightarrow$ effect, effect $\not\Rightarrow$ feasibility, and source feasibility $\not\Rightarrow$ cross-endpoint verdict stability—so a single accuracy, write-count, or leaderboard scalar cannot replace the typed record.

A.3 TRACE-RESOLUTION CONSEQUENCES

Definition 2 gives the deterministic and stochastic history laws, fixes the total-variation convention, and states the exact score-sufficiency premise. The deterministic formula is derived in Appendix F.1, and the complete Markov-contraction and bounded-score argument is in Appendix F.3.

At $s_Z = 0$, the positive and negative implemented trace mixtures coincide, but individual policies need not be pairwise aliased. Their assignment contrast is zero. If the named estimand requires

distinct implemented mixtures, the typed record marks that intended scientific contrast UNDEFINED; it does not reinterpret the zero as an estimate of distinct treatment versions. The bound does not apply if an action changes the row distribution or an omitted state path reaches the score; those are failures of the declared sufficiency premise. No finite magnitude cap follows for unbounded log probability or negative log likelihood without clipping, a bounded transform, or additional tail conditions; the empirical cap used in the paper is for binary correctness.

For any measurable summary $h(R, Z)$, define

$$s_h(q) = \mathrm{TV}\Big(h_{\#}P^{q^+}_{R,Z}, h_{\#}P^{q^-}_{R,Z}\Big). \tag{19}$$

Theorem 1 gives $0 \le s_h(q) \le s_Z(q)$. Equality holds when a Hahn sign partition for the two contrast measures is measurable through h; in particular, it holds when the full pair (R, Z) is recoverable almost surely from h. Recovering Z alone need not suffice because mixing rows can cancel contrast mass. Strict inequality is information discarded by the summary. This is a possibility result, not a claim that the observed task metric differs.

A.4 Why counts can fail, and when they are sufficient

Fix a context c and let every admissible update atom a induce a deterministic map T^c_a on the complete mutable state. For an ordered word $\Gamma = a_1 \cdots a_T$, define $R_c(\Gamma) = T^c_{a_T} \circ \cdots \circ T^c_{a_1}$ and let $\nu(\Gamma)$ count each atom.

The complete proof of Lemma 1 appears in Appendix F.2. On a restricted domain, factorization holds exactly when every admitted pair with the same count vector induces the same map; pairwise commutativity remains sufficient but need not be necessary. Total write count is coarser than ν, and an intermediate read can depend on a prefix even when final maps commute. Thus equal counts identify endpoint treatment only under stronger conditions than ordinary budget matching. This is explanatory algebra, not a claimed new commutativity theorem.

Compiler soundness under declared assumptions. Given a common initial state and randomness, a complete adapter-relative trace, and represented state sufficient for the score, equal traces induce the same score law. Unequal traces permit but do not prove an effect. Given policies Π, rows $\mathcal{R}$, endpoint e, dependency map $\mathcal{D}$, and contrast q, the compiler returns

$$\mathsf{Compile}(\Pi, \mathcal{R}, e, \mathcal{D}, q) \mapsto \big(\{z_{r,s,e}\}, \{D_r\}, \mathsf{alias}, \mathsf{exposure}, s_Z(q), \mathsf{certificate}\big). \tag{20}$$

It replays each policy only through the endpoint, removes transitions outside the dependency map, canonicalizes and hashes the remaining ordered trace, forms rowwise equivalence classes, computes resolution, and emits a source- and input-bound certificate before outcomes are opened. The result determines which comparisons may proceed; it never supplies their sign or magnitude.

A.5 Why separate coordinate studies do not identify a joint controller

Let $\mathcal{C}$ and E denote the context and endpoint sets, and let $\mathcal{H} \subsetneq \mathcal{C} \times \Pi \times \mathsf{E}$ be the tested context–policy–endpoint tuples and let $x^* \notin \mathcal{H}$. Over the unrestricted class of bounded response surfaces, any two distinct values in the outcome range can be assigned at x^* by two surfaces that agree at every point in $\mathcal{H}$. Hence separately measured coordinate effects do not identify a composed controller outside $\mathcal{H}$ without a declared separability, smoothness, or invariance restriction.

For a concrete two-coordinate case, suppose the design observes Y_{00}, Y_{10}, Y_{01} but not Y_{11}, with $Y \in [a, b]$. Then

$$Y_{11} - Y_{10} - Y_{01} + Y_{00} \in [a - Y_{10} - Y_{01} + Y_{00},\ b - Y_{10} - Y_{01} + Y_{00}]. \tag{21}$$

The missing cell may take any value in $[a, b]$; separate coordinate effects do not determine their composition.

If a prospectively fixed, commensurable utility U_c compares two policies π_a, π_b feasible in contexts c_1, c_2 and

$$[U_{c_1}(\pi_a) - U_{c_1}(\pi_b)][U_{c_2}(\pi_a) - U_{c_2}(\pi_b)] < 0, \tag{22}$$

then neither constant choice is optimal in both contexts within $\{\pi_a, \pi_b\}$; a maximizing rule on this restricted set must distinguish contexts. Abstention is a separate option only after its utility is defined. We do not apply this statement to endpoint-native metrics whose scales are not commensurable.

For a declared endpoint set $E \subseteq \mathsf{E}$, let m_e be endpoint e's metric and let I_e be the declared decision range for its estimated within-endpoint contrast. Let $\rho_e = +$ when $\inf I_e > 0$, $\rho_e = -$ when $\sup I_e < 0$, and $\rho_e =?$ otherwise. With $\mathbf{m} = (m_e)_{e \in E}$, define the contrast and resolved-sign status signature by

$$\widehat{\Delta}_e^{m_e}(\pi, \pi') = \widehat{\mu}_e^{m_e}(\pi) - \widehat{\mu}_e^{m_e}(\pi'), \qquad \boldsymbol{\sigma}_E^{\mathbf{m}}(\pi, \pi') = (\rho_e)_{e \in E} \in \{+, -, ?\}^{|E|}. \tag{23}$$

When instruments differ, this is an endpoint-status signature rather than evidence about one shared latent locality quantity or an identified causal effect of dataset identity.

A.6 Adapter robustness, coverage evidence, and gate sensitivity

This section separates three questions that a trace audit can otherwise blur: whether recorded executions conform to a declared trace, whether that trace contains every score ancestor, and how conclusions change when either the conditional-sufficiency premise or a gate threshold is varied. The calculations below are deterministic sensitivity analyses. They add no model outcomes and do not estimate the hidden-path error rate of any editor.

Conditional-kernel mismatch bound. The complete proof of Proposition 1, including the kernel-averaging and bounded-expectation steps, appears in Appendix F.5. Here $\epsilon_+ + \epsilon_-$ is one aggregate deviation budget; if instead each arm is known only to be within ϵ of the common reference kernel L, the bound uses $s_Z(q) + 2\epsilon$ inside the cap at one.

Controlled hidden-cache state-machine ablation. Let $N \geq 1$ and $m, k \geq 0$ be integers with $m + k \leq N$, let R be uniform over the N rows, and choose disjoint row sets C, H of sizes m, k. The positive arm records visible bit $V = \mathbf{1}\{R \in C\}$ and writes hidden cache bit $U = \mathbf{1}\{R \in H\}$; the negative arm writes neither. The declared trace records V but omits U. Holding these executions and declared traces fixed, toggle only the scorer: $Y_+^{(\lambda)} = \mathbf{1}\{R \in C\} + \lambda \mathbf{1}\{R \in H\}$, $Y_-^{(\lambda)} = 0$, for $\lambda \in \{0, 1\}$. When $\lambda = 0$, the cache is non-ancestral and the declared exact bound holds. When $\lambda = 1$, the cache becomes an omitted score ancestor. For this mode, define the common reference kernel on the declared input $x = (r, v)$ by

$$L(\mathrm{d}y \mid r, v) = \delta_{v + \mathbf{1}\{r \in H\}}(\mathrm{d}y).$$

Because $C \cap H = \varnothing$, this is a binary-output kernel on every input in the support of either arm. It matches the positive-arm kernel everywhere; on the negative-arm support it differs from the constant-zero kernel exactly on rows in H. Consequently,

$$s_Z(q) = \frac{m}{N}, \qquad \epsilon_+ = 0, \qquad \epsilon_- = \frac{k}{N}, \qquad \mathrm{TV}(P_Y^{q^+}, P_Y^{q^-}) = \frac{m + k}{N}. \tag{24}$$

Verification of Equation 24. For each $r \in C$, the positive arm has declared input $(r, 1)$ and the negative arm has $(r, 0)$; for each $r \notin C$, both arms have $(r, 0)$. Because the row coordinate is part of the input and every row has mass $1/N$, the two declared-input laws differ on exactly the m rows in C. Hence $s_Z(q) = m/N$.

When $\lambda = 0$, the common kernel $L_0(\mathrm{d}y \mid r, v) = \delta_v(\mathrm{d}y)$ agrees with both arms on their respective supports. Theorem 1 therefore applies exactly, and the output total variation is $m/N = s_Z(q)$.

Under the positive arm, $v = \mathbf{1}\{r \in C\}$, so the reference kernel returns $v + \mathbf{1}\{r \in H\} = Y_+^{(1)}$ on every row. Thus $\epsilon_+ = 0$. Under the negative arm, $v = 0$ and the actual output is zero, whereas the reference kernel returns one exactly for $r \in H$. The pointwise total variation is therefore $\mathbf{1}\{r \in H\}$, whose uniform-row average is $\epsilon_- = k/N$.

Finally, $P_Y^{q^+}$ is Bernoulli with success probability $(m + k)/N$, while $P_Y^{q^-} = \delta_0$. Their total variation is the positive law's mass at one, namely $(m + k)/N$. This equals $s_Z(q) + \epsilon_+ + \epsilon_-$, proving that the bound in Proposition 1 is attained. If the declared history is expanded to (V, U), the two arms differ exactly on $C \cup H$, so $s_Z^* = (m + k)/N$; the common kernel δ_{v+u} then restores the premise and bound of Theorem 1. □

Thus the incomplete exact bound misses by exactly k/N, Proposition 1 is tight, and expanding the trace to include U restores Theorem 1 with $s_Z^* = (m+k)/N$. A hash-bound CPU check evaluates both scorer modes on all 16 cells with $N = 1000$, $m \in \{0, 1, 10, 100\}$, and $k \in \{0, 1, 10, 50\}$. It verifies the non-ancestral bound in 16/16 cells, the omitted-ancestor failure in all 12 cells with $k > 0$, and restored coverage in 16/16 cells. Paired execution and declared-trace digests are identical across the scorer toggle. This is a finite state-machine construction, not a transformer run, KV-cache prevalence claim, or estimate of EasyEdit behavior.

Version-controlled live-neural conformance check. A separate CPU protocol was committed locally before its random model was constructed. Its design was informed by two earlier pilots and is not a fresh hypothesis test. It instantiates a two-layer Llama with PEFT-LoRA on the query and value projections. Every one of 16 frozen rows performs exactly two SGD writes in A-then-B order and reaches one common terminal model and optimizer state; only whether B precedes the score changes. The negative mixture contains two full pre-score traces and six A-only pre-score traces, while the positive mixture contains six and two, respectively, so the predeclared and separately recomputed resolution is exactly $s_Z = 1/2$.

For every live commit, the runtime records trainable-parameter transitions, parameter/buffer and optimizer digests, all module train/eval modes, Python, NumPy, and Torch random-number-generator states, and the nonzero scalar-score gradient paths. A separately loaded compiler validates all 16 chronological chains, rejects 16 precommitted corruptions (ten design-listed fault families plus six compiler-closure faults), and accepts the restored trace; the post-run checker regenerated the outputs byte-identically. The pre-run Git commit contains the manifest, design, dependency versions, runtime, compiler, and tests but neither output. The source and protocol were committed locally before execution; this is not external preregistration, a published-checkpoint result, an editor-quality measurement, or proof that unrecorded zero-gradient or hidden paths do not exist.

For the SocialIQA swaps, $s_Z = 1/1954$. Table 3 shows how a declared aggregate budget $\bar{\epsilon} = \epsilon_+ + \epsilon_-$ changes the binary-outcome cap. These are conditional calculations: the study does not supply an empirical value of $\bar{\epsilon}$.

Table 3: **SocialIQA certificate sensitivity.** The right column is $100 \min\{1, 1/1954 + \bar{\epsilon}\}$ percentage points.

Declared aggregate deviation $\bar{\epsilon}$	Binary-effect cap (points)
0	0.0512
0.001	0.1512
0.010	1.0512
0.050	5.0512

Conformance is not ancestry coverage. Table 4 reports the strongest claim supported at each recorded boundary. A conformance check asks whether an observed or replayed execution matches the declared trace. A coverage check would require an independent oracle for the complete set of score ancestors. No such oracle was recorded for EasyEdit internals, so false-negative and false-positive ancestry rates are unavailable rather than zero. A false negative can violate the common-kernel premise. A false positive does not omit a score path, but can split outcome-equivalent policies and make the quotient unnecessarily fine.

Table 4: **Adapter assurance at the recorded boundary.** Counts establish conformance only at the level named in the second column.

Study	Recorded boundary	Conformance evidence	Ancestry-coverage claim
MAC trace atlas	Declared deterministic scheduler graph	Complete 27,072-row enumeration; 184 outcome-blind checks	Exact relative to the declared scheduler adapter; no whole-program claim
Constructed neural LoRA test	Live trainable transitions plus model-state, optimizer, RNG, mode, and nonzero score-gradient records	16/16 chains validate; $s_Z = 1/2$; 16 checks reject and restoration passes	Exact conformance for recorded paths in one random tiny Llama; no ancestry-discovery theorem
GPT-2XL LoRA phase study	Declared PEFT-LoRA/SGD event and optimizer-step trace	Separate implementation matches 3,072/3,072 fields; 15/15 tests reject five omission/alias/substitution fault classes	Structural reconstruction only; no live-state observation or completeness theorem
WikiRecent	Score-ancestral editor outer calls	512/512 policy rows match precomputed certificates	EasyEdit-internal false-positive/negative rates unavailable
GPT-2XL endpoints	Source- and receipt-bound outer-call replay	448/448 compiled-cell digests; 1,792 family–seed units and 3,584 policy traces match	No observed internal-step ledger, parameter hashes, or random-number-generator snapshots
Llama-3.2-3B endpoints	Declared dose-four/terminal atomic outer-call traces	Post-outcome projection: 3,584/3,584 rows conform; all 1,792 paired traces separate ($s_Z = 1$)	EasyEdit internals and cross-bundle correspondence unavailable; diagnostic component only

Exact locality-threshold breakpoints. Let $d \geq 0$ be an allowed damage in percentage points, so the success threshold is $-d$. For point $\widehat{L}$ and range $[\ell, u]$ with $\ell \leq \widehat{L} \leq u \leq 0$, the main-text joint rule is FAIL when $d < -u$, INCONCLUSIVE when $-u \leq d \leq -\ell$, and PASS when $d > -\ell$. Under this containment assumption, the joint rule reduces to the interval rule, so $\widehat{L}$ introduces no additional breakpoint. Table 5 recomputes this rule from four hash-bound frozen records; it preserves their frozen five-point pass/non-pass decisions and the typed labels used in the current evidence assembly.

Table 5: **Locality verdict as a function of damage tolerance d.** Endpoint records are not pooled. Bounds are percentage points.

Context	Policy	FAIL	INCONCLUSIVE	PASS	At $d = 5$
CounterFact-new160	MEMIT-high	$d < 1.125000$	$1.125000 \leq d \leq 3.203125$	$d > 3.203125$	PASS
CounterFact-new160	ROME-low	$d < 3.078125$	$3.078125 \leq d \leq 6.421875$	$d > 6.421875$	INCONCLUSIVE
WikiRecent shared locality	MEMIT-high	$d < 33.984375$	$33.984375 \leq d \leq 52.734375$	$d > 52.734375$	FAIL
WikiRecent shared locality	ROME-low	$d < 32.2265625$	$32.2265625 \leq d \leq 50.781250$	$d > 50.781250$	FAIL

The six selected LoRA settings have point damage from 30.625 to 53.203 points. More importantly, retention passes in 0/6, so relaxing locality alone cannot turn any of their full conjunctions into a pass. The submitted analysis uses the main-text joint rule, matching the typed implementation: equality at either interval boundary is INCONCLUSIVE. Applying an independent fail-precedence conjunction would instead relabel several inherited threshold-crossing ROME records from INCONCLUSIVE to FAIL. An earlier WikiRecent replay helper also used a point miss or interval miss and treated upper-bound equality as failure, although its frozen protocol specified only the pass predicate. Every observed WikiRecent interval lies wholly below the five-point threshold, so harmonization changes none of its four verdicts or any headline result. The package validator binds the thresholds and applies one convention to future reporting.

The prospective WikiRecent record also exposes the closest non-locality threshold: ROME's rephrase installation point is 0.20703125, only 0.00703125 above the frozen 0.20 cutoff (its descriptive reweighting range crosses that cutoff, but the registered installation gate is point-based). Raising the installation cutoff above 0.20703125 would change that component to FAIL; it would not change the existing full verdict, which already fails locality. All four WikiRecent acquisition and continuous-retention ranges clear zero, and every retained-among-installs point clears 0.50.

Controlled locality-readout sensitivity. All three summaries use the same 1,280 editor-by-order CounterFact rows, reset and terminal checkpoints, ten neighborhood prompts per family, family weights, and 10,000 paired-bootstrap index draws; only the readout changes. MEMIT-minus-ROME is $+2.562500$ percentage points $[+1.062500, +4.265625]$ for neighborhood success, where higher is better; -0.094463 $[-0.151609, -0.044748]$ for full-vocabulary KL from reset; and -0.111094 $[-0.135156, -0.088750]$ for greedy-next-token change, where lower is better for the latter two. The primary success endpoint and secondary diagnostics were encoded in the frozen analysis; interpreting their agreement as factor-isolation evidence is post hoc. This fixes neither probe selection nor cross-bundle factors, so it rules out dependence on the binary score at this source point but does not attribute the observed reversals. Relative ordering is robust across the three readouts; absolute feasibility under KL, greedy change, or UniEdit remains UNDEFINED without separately frozen gates.

Threshold sensitivity and instrument sensitivity are different audits. The table varies d while holding the locality probe and estimator fixed. Replacing binary relation-specificity with full-vocabulary KL, greedy token change, or a compositional benchmark instead defines a new metric m, gate, or target context and requires a separately frozen record. The post-hoc WikiRecent KL and token-change diagnostics point in the same damage direction but reuse the same two probes, so they add metric breadth rather than probe breadth. Gate values should therefore be tied to an explicit deployment loss or tolerance and frozen per endpoint; the Control Map does not nominate five points as a universal calibration.

B EXECUTION-RESOLUTION AND EFFECT DETAILS

Proposition 2 (Working-basis estimability)**.** *Let $P_r = D_r(D_r^\top D_r)^\dagger D_r^\top$ be the orthogonal projector that averages policy coordinates within row r's executed-history classes. Let $U \in \mathbb{R}^{|\Pi|\times d}$ be a centered design basis, $\mathbf{1}^\top U = 0$, and let $\mu = \alpha\mathbf{1} + U\beta$ be the working vector of policy means. Define $I_U = U^\top(\sum_r w_r P_r)U$. A functional $a^\top\beta$ is structurally estimable when it has the same value for every two parameter pairs producing the same projected means $P_r\mu$ on all rows with $w_r > 0$. Then $a^\top\beta$ is structurally estimable if and only if $a \in \mathrm{range}(I_U)$.*

The proof appears in Appendix F.7.

B.1 COMPLETE OUTCOME-BLIND TRACE ATLAS

The atlas uses the fixed MAC-760M evaluator, 24 permutations of periods $\{1, 2, 4, 8\}$ over four ordered six-layer bands, and eight complete benchmark panels. If L is the tokenizer length of the write-enabled common prompt, 16-token native chunks, 512-token data segments, and 256 inserted memory tokens before each segment after the first give

$$H(L) = \left\lfloor \frac{L + 256\,\lfloor (L-1)/512 \rfloor}{16} \right\rfloor. \tag{25}$$

For period p, the phase-zero trace is $a_p(H) = (\mathbf{1}\{t \bmod p = 0\})_{t=1}^{H}$. Direct enumeration gives 1, 4, 12, and 24 effective schedule classes for $H = 0$, $H = 1$, $H \in \{2, 3\}$, and $H \geq 4$. Candidate and target tokens do not enter H because writes are disabled before they are scored.

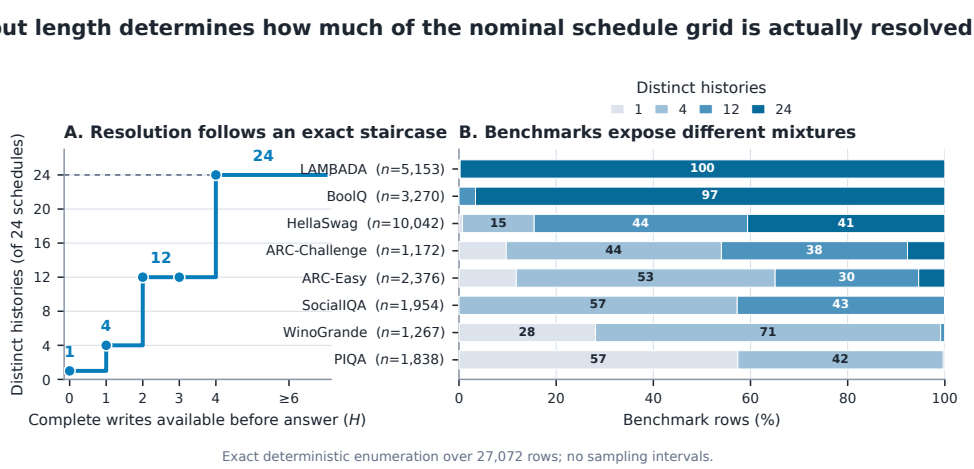


Figure 3: **Resolution depends on the pre-answer write horizon.** Exact enumeration yields 1, 4, 12, or 24 distinct histories as the number of complete writes available before an answer increases. The right panel shows the resulting mixture for each benchmark, with each bar normalized by that benchmark's row count. The calculation covers all 27,072 rows and is deterministic, so no sampling intervals apply.

The eight evaluation panels are SocialIQA (Sap et al., 2019), BoolQ (Clark et al., 2019), LAMBADA (Paperno et al., 2016), PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2020), HellaSwag (Zellers et al., 2019), and the ARC Easy and Challenge sets (Clark et al., 2018).

Table 6: Complete execution atlas. "Pair alias" is the mean fraction of 276 nominal schedule pairs with equal rowwise traces. "Full rows" is the percentage of rows that distinguish all 24 policies. Rank and condition number κ refer to the centered nine-dimensional band–period working basis and diagnose design support, not causal identification or sampling power.

Panel	Rows	Pair alias	Full rows	Rank/9	$\kappa(I_U)$
SocialIQA	1,954	14.31%	0.05% (1 row)	9	1954.000
BoolQ	3,270	0.15%	96.61%	9	1.035
LAMBADA	5,153	0.01%	99.77%	9	1.002
PIQA	1,838	66.60%	0.00%	6	—
WinoGrande	1,267	43.58%	0.00%	6	—
HellaSwag	10,042	5.84%	40.60%	9	2.445
ARC-Easy	2,376	24.65%	5.39%	9	16.375
ARC-Challenge	1,172	21.02%	7.68%	9	11.756

Across all 27,072 rows, the mean collision probability over a uniformly chosen row and unordered policy pair is $960{,}264/7{,}471{,}872 = 0.128517$. SocialIQA illustrates why global injectivity is insufficient. Let P_z project the 24-policy response vector onto values constant within row z's trace classes, let $R = 1954^{-1}\sum_z P_z$, and let U be the orthonormal centered nine-dimensional band–period contrast basis. The smallest eigenvalue of $U^\top RU$ is only $1/1954$. The atlas validator performs 184 deterministic checks without reading predictions, checkpoints, or scheduler state. Condition numbers are reported only for full-rank panels; a dash denotes rank deficiency. They describe support geometry in the declared working basis. The 27,072 rows are benchmark rows, not independent training trajectories or an experimental sample-size claim.

In a retrospective SocialIQA outcome check, six period-4/period-8 schedule swaps differ in declared pre-answer history on only the single fully resolving row, giving $s_Z = 1/1954$ for each pair. Their logits differ only on that row, and their predictions and accuracies are identical. These six swaps therefore do not support a broad schedule effect on SocialIQA; a separately named pair in the same panel is fully resolved.

B.2 COMPLETE-ORDER LOCALITY CORRECTION

The four-event editor grids index dose through query placement while terminal state follows the complete event sequence. Dose-indexed Latin-square cells can therefore differ in focal rank and event order. The corrected comparison holds editor, strength, seed/order replicate, family, and the complete ordered case IDs fixed. It covers 160 MEMIT, 480 ROME, and 960 LoRA cross-dose pairs. All 1,600 pairs have identical terminal parameter hashes and identical locality, proving that the earlier terminal dose–locality curve was a mixture of orders rather than a same-state dose effect. This exact equality says nothing about query-time acquisition, which is measured before the shared terminal state.

B.3 IMMEDIATE-VERSUS-DEFERRED EFFECT

The phase comparison holds selected objectives, total actions, optimizer steps, and terminal state fixed while moving the same writes across the focal query. Immediate and deferred policies each execute 772 selected actions and 3,088 optimizer steps per endpoint. The held-out immediate-minus-deferred means are $+0.119608$, $+0.059816$, $+0.432027$, and $+0.168832$ nats for GPT-2 Small, Medium, Large, and XL; their separate Student-t_3 intervals are $[+0.096519, +0.142698]$, $[+0.053095, +0.066537]$, $[+0.294741, +0.569313]$, and $[+0.140598, +0.197066]$. Each model has four LoRA-initialization units and 4/4 positive signs; under an independent and identically distributed (i.i.d.) symmetric-sign null, the exact one-sided probability is nevertheless $1/16$, so each frozen gate relies on its stated parametric interval.

The prospective Pythia-410M confirmation returns $+0.1592$ nats with Student-t_3 interval $[-0.4484, +0.7667]$ and only 3/4 positive signs. Exact execution checks pass, making this a scientific nonconfirmation rather than a missing result. The five studies share one constructed 128-row association panel; the four GPT-2 sizes are analyzed separately and do not become an $n = 16$ scale curve.

B.4 EQUAL-COUNT PLACEMENT RESPONSE

The placement study evaluates all 24 assignments of periods $\{1, 2, 4, 8\}$ to four ordered six-layer bands on the same 6,000 Single-NIAH rows. Accuracy ranges from 19.733% for schedule $(4, 1, 2, 8)$ to 34.233% for $(1, 8, 4, 2)$. The additive band–period description accounts for 95.322% of between-schedule variation; leave-one-schedule-out prediction has $R^2 = 0.863$ and root-mean-square error 1.592 percentage points. These are descriptive results for one checkpoint and evaluation set. The full surface is shown in Figure 2.

B.5 SELECTED-SETTING DOSE SUMMARY

Across 18 purposively and partly adaptively assembled LoRA/SGD settings, the empirical peak at doses $1/2/4/8$ occurs in $9/7/2/0$ settings, respectively, and the best-dose prediction specified before outcomes is confirmed in 7 of 18 settings. Each setting uses the same four-dose grid and four receiver seeds, but the models, checkpoints, writers, and timing classes differ. The counts are therefore a descriptive summary of the selected grid, not an estimate of population prevalence or a universal dose rule.

B.6 EFFECT RECORD AND UNCERTAINTY

For setting c and metric m, let $P_{c,m}$ be the evaluation distribution over examples r and exogenous randomness ξ, and let $Y^m_{r,c}(\pi, \xi)$ be the resulting score. Define $\mu^m_c(\pi) = \mathbb{E}_{(r,\xi)\sim P_{c,m}}[Y^m_{r,c}(\pi, \xi)]$. For a policy comparison q with sufficient support from the recorded-history design, the effect record is

$$\Delta^m_c(q) = \frac{2}{\|q\|_1} \sum_{\pi \in \Pi} q_\pi \mu^m_c(\pi), \qquad \mathsf{Eff}^m_c(q) = (\widehat{\Delta}^m_c, I^m_c, \mathcal{U}_c, \tau_c), \tag{26}$$

where I^m_c is the uncertainty summary, $\mathcal{U}_c$ the inference unit, and τ_c the analysis-timing label. The effect is evaluated only for a comparison supported by the recorded histories and does not replace the separate memory-quality checks. Unless stated otherwise, the reported 95% ranges apply the paired family- or block-level reweighting draws to one fixed evaluation set; they describe sensitivity to those units rather than population coverage.

C ABSOLUTE FEASIBILITY EVIDENCE

The following subsections retain complete gate values and chronology for the main-text feasibility claims. The summaries do not pool underlying cohorts or designs.

C.1 COMPLETE FEASIBILITY DECISION RULES

Let $I_A = [\ell_A, u_A]$, $I_R = [\ell_R, u_R]$, and $I_L = [\ell_L, u_L]$ be the finite-panel ranges for acquisition, retention, and locality success, respectively. For a valid observed range requiring evidence above threshold τ, define

$$\mathsf{S}_\tau([\ell, u]) = \begin{cases} \text{PASS}, & \ell > \tau, \\ \text{FAIL}, & u < \tau, \\ \text{INCONCLUSIVE}, & \ell \le \tau \le u, \\ \text{UNDEFINED}, & \text{the required record is absent or invalid.} \end{cases} \tag{27}$$

Let $\mathsf{S}^{\mathrm{pt}}_\tau(v)$ be PASS if $v \ge \tau$ and FAIL otherwise, with missing or invalid values UNDEFINED. For a point–range record, let $\mathsf{J}_\tau(v, [\ell, u])$ be PASS if $v \ge \tau$ and $\ell > \tau$, FAIL if $v < \tau$ and $u < \tau$, and INCONCLUSIVE otherwise; it is UNDEFINED if either input is absent or invalid. This four-state harmonization leaves every original study's pass predicate unchanged.

Let $\widehat{p}_{\mathrm{inst}}$ be the fraction of edit pairs meeting the declared immediate-learning condition, $\widehat{p}_{\mathrm{keep}}$ the fraction retained among those installed edits, and $\widehat{L}$ the point estimate of locality success. Using the status conjunction $\bigwedge_{\mathcal{S}}$ with precedence FAIL>UNDEFINED>INCONCLUSIVE>PASS, and retaining each study's thresholds, the reporting records are

$$\begin{aligned} S_A &= \mathsf{S}_0(I_A), \\ S_R &= \bigwedge_{\mathcal{S}} \left(\mathsf{S}_0(I_R), \mathsf{S}^{\mathrm{pt}}_{0.20}(\widehat{p}_{\mathrm{inst}}), \mathsf{S}^{\mathrm{pt}}_{0.50}(\widehat{p}_{\mathrm{keep}}) \right), \\ S_L &= \mathsf{J}_{-5}(\widehat{L}, I_L), \qquad S_{\mathrm{full}} = \bigwedge_{\mathcal{S}} (S_A, S_R, S_L). \end{aligned} \tag{28}$$

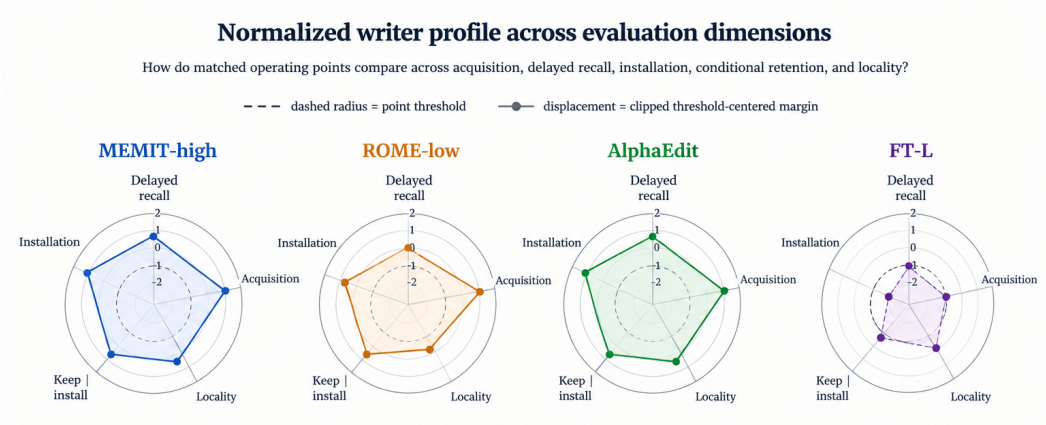
The retention-among-installs field is UNDEFINED when no installation occurs. Cost is reported separately and becomes a gate only under a declared budget.

Table 7: **Acquisition cannot stand in for feasibility.** Rows are assay-level summaries; underlying counts, horizons, cohorts, and inference units are not pooled. A full pass is the conjunction of all protocol-specific behavioral clauses; no common cost budget was declared.

Assay	Acquisition	Retention	Locality	Full gate
LoRA, six selected CounterFact settings	6/6 PASS	0/6 PASS	0/6 PASS	0/6 PASS
ROME, 13 CounterFact operating points	13/13 PASS	13/13 PASS	0/13 PASS	0/13 PASS
MEMIT-high, pilot-disjoint 160-family CounterFact point	PASS	PASS	PASS	PASS
ROME-low, same 160-family point	PASS	PASS	INCONCLUSIVE	INCONCLUSIVE
AlphaEdit, shared-eight-event CounterFact doses	4/4 PASS	4/4 PASS	4/4 PASS	4/4 PASS

C.2 MATCHED FACTUAL-RECALL STAGE-0 PROTOCOL AND FULL DECISION

Frozen protocol and collection boundary. Stage 0 was frozen before validation-run or study outcomes under protocol `NATURAL-FACTUAL-RECALL-STAGE0-REMOTE-BUNDLE-V2`. It crosses four fixed settings (Pythia-410M current, Pythia-410M step100000, GPT-2 Large, and GPT-2XL), four LoRA initialization seeds, three doses $\{1, 2, 4\}$, and four eight-family shards. The resulting 192 evaluation runs cover 32 CounterFact families and produce 1,536 family–model–seed–dose rows. Each row includes a direct prompt, two unseen paraphrases, terminal recall after all four writes, and ten unrelated-neighborhood prompts. At $d = 4$ the query is terminal, so continuous acquisition and retention coincide; $d = 1, 2$ include later writes. Timed and deferred policies execute the same four actions and 16 optimizer steps and must end in byte-identical mutable state.



Writer	Acquisition (nats) ↑	Delayed recall (nats) ↑	Installation ↑	Keep | install ↑	Locality (damage, pp) ↓
MEMIT-high	7.18 [6.21, 8.09]	7.27	0.38	1.00	1.09
ROME-low	6.38 [5.46, 7.23]	5.60	0.31	0.80	6.72
AlphaEdit	7.72 [6.62, 8.79]	7.72	0.56	1.00	1.56
FT-L	0.28 [0.14, 0.40]	-0.34	0.00	0.00	12.81

Figure 4: **Dose-one writer profiles differ across evaluation dimensions.** The threshold-centered display uses a shared eight-event CounterFact cohort with 16 disjoint family-pair blocks per writer (32 focal orientations and four seeds; 128 rows per writer). Outward is favorable, the dashed radius marks each axis's threshold, and the embedded table reports values on their native scales. Dose one indexes query-time acquisition; delayed recall and locality are measured after the shared eight-event sequence. Formal acquisition, delayed-recall, and locality decisions use recorded intervals; installation and keep-among-installs use point thresholds, and those intervals are not displayed on the spokes. The axes are separate requirements, so polygon area is not an overall utility or an editor-family ranking. FT-L denotes the constrained fine-tuning baseline. Because FT-L has no installations, its displayed 0.00 at Keep|install is a layout placeholder; the formal conditional-retention field is UNDEFINED.

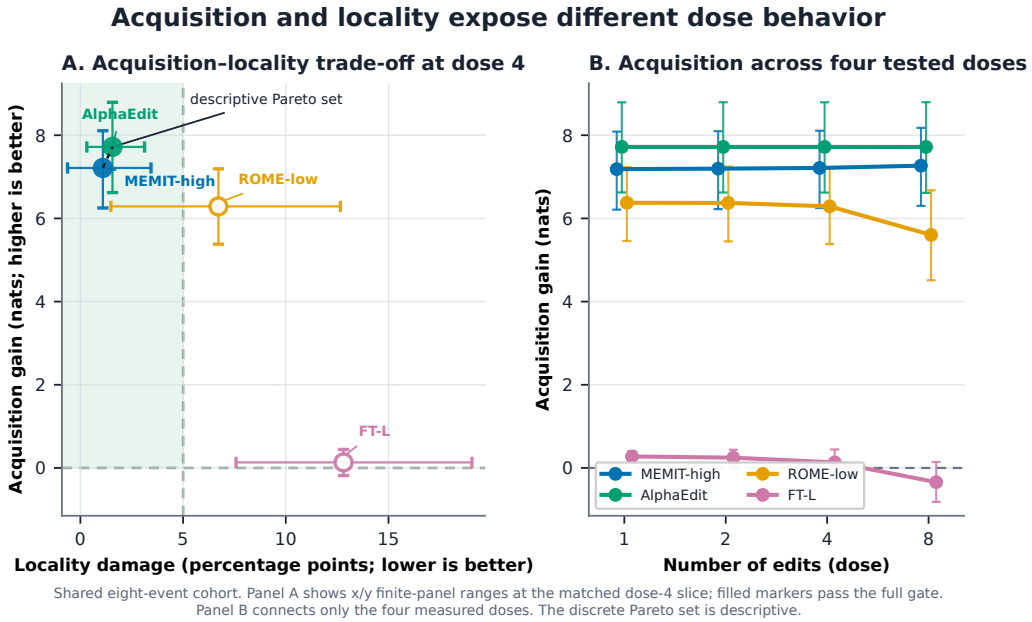


Figure 5: **Acquisition and locality expose different dose behavior on the shared eight-event cohort.** Panel A juxtaposes query-time acquisition with terminal locality at the matched dose-four slice and shows separate marginal finite-panel ranges; filled markers pass the full declared gate. Panel B shows query-time acquisition at the four measured doses $\{1, 2, 4, 8\}$ only. The Pareto annotation and connecting lines are descriptive: they neither identify a causal dose trade-off, interpolate between doses, nor define an uncertainty-adjusted frontier.

Before outcome access, a completion audit verified all 192 expected runs as unique and complete, with no retries, and checked all stored definitions. A pre-outcome correction normalized one overly restrictive launch-metadata field while preserving every scientific field, result path, output, and estimator. The corrected source and outcome bundle are content-bound in the evidence ledger.

Frozen rule. The family is the inferential unit; the four adapter seeds and two paraphrases are averaged within family, followed by 10,000 deterministic family-cluster bootstrap draws. At the pre-outcome-frozen predicted dose, acquisition requires a positive lower endpoint of the frozen 95% family-reweighting range for the mean two-paraphrase timed-minus-deferred target-log-probability contrast. Retention requires a positive lower endpoint for terminal-minus-reset target log probability, joint post-focal installation of at least 20%, and terminal retention among installed pairs of at least 50%. Locality requires a point estimate of at least -5 percentage points and a lower endpoint above -5 for terminal-minus-reset neighborhood success. Stage 1 is authorized only if all three clauses pass in at least two of four settings.

Table 8: Full Stage-0 decision at each pre-outcome-frozen predicted dose. "Call" is the frozen selected dose, "Peak" is the empirical maximizing dose, and "Keep|install" is retention among installed pairs. "Acq." and "Ret." denote acquisition and retention. Log-probability contrasts are in nats; locality is in percentage points.

Setting	Call	Peak	Acq. [95% range]	Ret. [95% range]	Install	Keep|install	Locality [95% range]
Pythia current	1	1	$+8.481$ $[+7.531, +9.473]$	-10.947 $[-13.035, -8.853]$	0.8125	0.0096	-37.03 $[-49.61, -24.53]$
Pythia step100000	1	2	$+6.666$ $[+5.781, +7.572]$	-12.029 $[-13.529, -10.609]$	0.3906	0.0000	-38.44 $[-48.83, -27.42]$
GPT-2 Large	4	4	$+4.463$ $[+3.841, +5.123]$	$+4.463$ $[+3.841, +5.123]$	0.0781	0.9000	-53.20 $[-62.81, -43.20]$
GPT-2XL	4	4	$+2.746$ $[+2.041, +3.477]$	$+2.746$ $[+2.041, +3.477]$	0.0547	0.7143	-41.95 $[-51.88, -32.03]$

Table 9: Mean acquisition contrast across the complete Stage-0 dose grid. These are descriptive point means; the frozen decisions use the predicted-dose family-bootstrap ranges in Table 8.

Setting	d_1	d_2	d_4
Pythia current	$+8.481$	$+3.225$	-6.114
Pythia step100000	$+6.666$	$+6.674$	-7.497
GPT-2 Large	$+0.520$	$+1.340$	$+4.463$
GPT-2XL	$+0.342$	$+1.101$	$+2.746$

The acquisition clause passes 4/4 and the frozen peak calls match 3/4 numerical maxima. Nevertheless, retention fails 4/4, locality fails 4/4, and zero settings pass the conjunction. The immutable decision is therefore `STAGE1_STOP`; the planned follow-up stage was not run. This test supports setting-conditioned acquisition and localizes distinct write-quality failures, but it does not establish a safe writer, a population-level predictor accuracy, or an online controller.

C.3 Natural factual-recall family extension

The pre-outcome-frozen extension applies the Stage-0 estimator and gate to Qwen2.5-0.5B (Qwen Team, 2024) and Llama-3.2-1B (Meta AI, 2024a). It contains 96 unique first-attempt runs and 768 sufficient-statistic rows: two models, four LoRA initialization seeds, three doses, and 32 CounterFact families. A completion audit preceded outcome collection, and an independent replay reproduces the decision. Before outcome access, a collection correction removed an inherited cardinality assertion while retaining the exact grid, per-row validator, and decision rule; no outcome had been emitted.

Table 10: Family-extension acquisition surfaces and metrics at the frozen dose-1 call. A is acquisition, retention is a log-probability contrast in nats, and L is terminal-minus-reset neighborhood success in percentage points.

Model	$A(d_1)$ [95% range]	$A(d_2)$	$A(d_4)$	Retention (nats) / L (pp)
Qwen2.5-0.5B	6.145 $[4.737, 7.556]$	-2.212	-4.366	-7.029 / -30.625
Llama-3.2-1B	4.581 $[3.837, 5.358]$	0.333	-1.671	-4.143 / -38.359

Both frozen dose-1 calls equal the empirical finite-grid peak and both acquisition gates pass. Neither retention nor locality passes: retention ranges are $[-8.104, -5.961]$ and $[-5.504, -2.833]$, locality ranges are $[-41.328, -18.984]$ and $[-49.922, -25.938]$ percentage points, and retained fractions among installed facts are zero. Combining these two settings with the prior four yields 5/6 exact calls for the frozen dose-selection rule versus 3/6 for always-dose-1 (two wins, four ties, zero losses). At $n = 6$ this is a descriptive selected-setting tally with no significance or population claim.

C.4 Partially outcome-blind published-editor schedule grids

Evidence role. The analyses below reuse completed GPT-2XL Stage-0 outputs on the same 32 CounterFact families, four order replicates, two held-out paraphrases, terminal retention endpoint, ten neighborhood prompts, and original gate definitions. The analysis rules were hash-frozen after the $d = 4$ outcomes were known but before the completed $d = 1, 2$ outputs were opened. Thus they are partially outcome-blind extensions, not fully prospective peak tests. Empirical peaks and winner–runner-up contrasts are descriptive. Dose 8 is absent because the inherited Stage-0 family contains only four events. Acquisition is evaluated at the dose query, while retention and locality are evaluated at terminal after all four writes. Dose-indexed Latin-square cells can have different complete event orders and focal ranks. At $d = 4$ the query is terminal and continuous acquisition equals retention; $d = 1, 2$ include later writes.

Table 11: GPT-2XL published-editor schedule grids. A is acquisition at the dose query in nats and D is terminal locality damage after all four writes in percentage points (lower is safer). Each entry is A/D. Native strength values are not commensurate across editors.

Editor	Native strength	$d = 1$	$d = 2$	$d = 4$	Frozen A-response label
MEMIT	high	$7.185/1.875$	$7.196/1.875$	$7.205/1.875$	DOSE-INSENSITIVE
ROME	low	$6.376/5.391$	$6.513/5.469$	$6.854/6.328$	INCONCLUSIVE
ROME	default	$8.118/18.281$	$8.167/20.312$	$8.201/25.547$	DOSE-INSENSITIVE
ROME	high	$8.611/35.391$	$8.625/38.906$	$8.582/41.797$	DOSE-INSENSITIVE

For MEMIT-high, the max–min acquisition range is 0.019725 nats with a paired-family 95% empirical reweighting range $[0.004358, 0.040115]$, wholly below the frozen 0.5-nat practical-sensitivity threshold. Joint installation is $0.375/0.375/0.367$ and keep among installs is 1.0 at every dose; acquisition, retention, locality, and the full gate pass at all three doses. The nominal $d = 4$ minus $d = 2$ lead is only 0.008073 nats (95% bootstrap range $[-0.002786, 0.020460]$). This is classified `DOSE_INSENSITIVE`; it does not turn the known $d = 4$ maximum into a prospective prediction.

For ROME, the acquisition max–min ranges are 0.478 nats ($[0.210, 0.764]$) at low strength, 0.083 ($[0.020, 0.325]$) at default, and 0.043 ($[0.020, 0.358]$) at high. The frozen labels are respectively INCONCLUSIVE, `DOSE_INSENSITIVE`, and `DOSE_INSENSITIVE`. All nine cells pass acquisition and retention, but none certifies locality or the full conjunction. The low-strength locality- response range is below the frozen 5-point dose-sensitivity margin; uncertainty bands around the default/high 7.266/6.406-point ranges cross that margin, so those response classifications are inconclusive.

The combined pattern is not evidence that dose and writer quality are independent, nor that dose sensitivity diagnoses a weak writer: unsafe ROME settings can be acquisition-flat. Query-time acquisition and terminal quality both depend on the tested schedule/editor cell, but they are not a same-terminal-state comparison. The LoRA terminal-locality mixtures differ across dose labels, MEMIT-high is low-damage across its mixtures, and ROME's absolute locality depends strongly on native strength even where query-time acquisition is dose-insensitive. The disjoint 64-family MEMIT–ROME primary comparison remains INCONCLUSIVE, so these discovery-panel grids neither establish universal safety nor rank the methods causally.

The raw-row audit makes the temporal boundary exact. Holding editor, strength, seed, family, and complete ordered case IDs fixed, terminal model state and terminal locality are bit-identical across all 160 MEMIT and 480 ROME cross-dose pairs. Together with 960 LoRA pairs, the audit covers 1,600/1,600 pairs with zero mismatches. Therefore movement of terminal locality across dose-indexed cells is an order-mixture association, not an identified effect of moving the read-only query boundary.

All 384 MEMIT rows and 1,152 ROME rows completed successfully on the first attempt. Independent byte-level re-execution reproduced both frozen analysis records exactly.

C.5 PROSPECTIVE SHARED-EIGHT-EVENT MEMIT FACET

Dose 8 is undefined in the inherited four-event Stage-0 family. We therefore froze a separate structural extension before executing any of its outcomes: adjacent CounterFact families are paired into 16 disjoint blocks, both focal orientations are retained, and each focal is followed by a seed-dependent permutation of the other seven distinct events. All four doses share the identical eight-event sequence. The inference unit is the disjoint family-pair block; two focal orientations and four order replicates are averaged within block before 10,000 paired block-bootstrap draws.

Table 12: MEMIT-high on the separate shared-eight-event GPT-2XL facet. A is query-time acquisition in nats and D is terminal locality damage in percentage points. Brackets are pointwise 95% block-reweighting ranges; the max–min range and dose-eight-minus-dose-four comparison below use the frozen max-deviation band.

Dose	A [95% range]	D	Install	Retain/install	Full gate
$d = 1$	7.18483 [6.21130, 8.08745]	1.09375	0.375	1.000	PASS
$d = 2$	7.19700 [6.22382, 8.09705]	1.09375	0.375	1.000	PASS
$d = 4$	7.21356 [6.25123, 8.10839]	1.09375	0.375	1.000	PASS
$d = 8$	7.26836 [6.30097, 8.17701]	1.09375	0.375	1.000	PASS

The empirical maximum is dose 8. Its difference from dose 4 is $+0.05481$ nats with a max-deviation 95% range $[+0.00180, +0.10781]$. Thus the range remains above zero under the frozen joint-reweighting rule, but the estimate is small relative to the prespecified 0.5-nat practical threshold. The complete max–min range is 0.08353 nats with max-deviation range $[0.03053, 0.13653]$, yielding `DOSE_INSENSITIVE`. Terminal locality damage is 1.09375 points and terminal-minus-reset retention is 7.26836 nats at all four doses; every acquisition, retention, locality, and full gate passes.

The collector verifies exact equality across doses of the complete event order, reset/post-focal/terminal endpoint objects, retention, locality, recall, and terminal parameter hash for every seed and focal orientation. Only the read-only query checkpoint varies. This makes the small acquisition contrast internally interpretable within the eight-event construction, unlike the four-event schedule-mixture locality comparison.

Scope remains narrow. The structural extension was designed after the four-event `DOSE_INSENSITIVE` result was known, so it is prospective for its new executions but is not an independent pre-four-event confirmation. It is not pooled with Stage 0 and does not establish that dose insensitivity diagnoses writer quality.

C.6 Fresh published-editor confirmation

This pre-outcome-frozen test uses 64 CounterFact families excluded from the 430-case previously accessed discovery pool by both case ID and normalized subject. Target values may recur across panels; clean model reset for every family makes this compatible with the frozen design. The comparison holds GPT-2XL, the four sequential singleton edits and their order, two held-out paraphrases, ten neighborhood prompts, and the 10,000-draw paired family bootstrap fixed. The two operating points are sequential-singleton MEMIT-high (weight 6325) and ROME-low (clamp factor 2), selected from discovery before confirmation outcome access as the closest acquisition comparator. They are not commensurate strength labels.

Table 13: Fresh 64-family confirmation. A is post-fourth-edit minus reset target log probability; L is terminal-minus-reset neighborhood success, so $D = -L$ is locality damage. All brackets are pointwise 95% family-reweighting ranges; editor contrasts are formed within family before reweighting.

Point	A [95% range]	D	L range	Install	Keep | install
MEMIT-high	7.980 [7.272, 8.651]	3.750	$[-7.969, -0.781]$	0.508	1.000
ROME-low	7.522 [6.685, 8.264]	7.148	$[-11.523, -3.086]$	0.457	0.966

Both points pass acquisition and retention but are inconclusive on the frozen locality threshold, so neither full conjunction passes. The paired MEMIT-minus-ROME acquisition contrast is $+0.459$ nats (95% range $[-0.093, +1.030]$). The primary locality-success contrast is $+3.398$ percentage points (95% range $[-0.392, +7.070]$). Because that range crosses zero, the pre-outcome expectation `MEMIT_SAFER_REPLICATES` is not confirmed and the frozen taxonomy is INCONCLUSIVE; no retry or redesigned panel was used.

Here the post-fourth-edit acquisition query is terminal, so continuous retention duplicates acquisition rather than testing delay; installation and retained-among-installs remain separate components.

Two read-only neighborhood diagnostics already emitted by the unchanged runtime were secondary. MEMIT changes the greedy next token on 6.5625% of queries versus 14.6875% for ROME, a paired difference of -8.125 percentage points (95% range $[-12.931, -3.750]$). Mean full-vocabulary Kullback–Leibler (KL) divergence is 0.1076 versus 0.2136, but the paired difference -0.1060 has 95% range $[-0.2707, +0.0180]$. The greedy-change range lies below zero, but this diagnostic does not replace the inconclusive primary locality comparison or establish general safety.

C.7 Disjoint pre-outcome-frozen 160-family extension

Chronology and frozen panel. The earlier disjoint 64-family MEMIT-high–ROME-low comparison was INCONCLUSIVE: its paired locality-success contrast was $+3.3984$ percentage points with 95% range $[-0.3916, +7.0703]$, and MEMIT did not clear its own strict locality range. That completed panel was used only as an external pilot for sample size. Its family-level paired standard deviation was 15.1734 points. Assuming independent and identically distributed (i.i.d.) families, a normal-approximation sizing heuristic gives approximately 157 families for 80% two-sided power at the observed effect; the design rounded upward to 160 for exact eight-family shards and rank balance. This estimate may be optimistic because it is based on the pilot's observed effect.

The primary analysis uses only 160 new, disjoint CounterFact families and does not pool the pilot. Their 640 case IDs and normalized subjects have zero overlap with all 686 previously accessed discovery and confirmation cases. The fixed comparison is GPT-2XL at $d = 4$, sequential-singleton MEMIT-high ("mom2 update weight" 6325) versus the discovery-selected, acquisition-nearest ROME-low (clamp factor 2). The protocol retains two held-out paraphrases, ten neighborhood prompts, and four order replicates, averaged within family. The family is the reweighting unit; ranges use 10,000 paired family-bootstrap draws from NumPy's PCG64 generator with frozen seed 2026100601.

Table 14: Disjoint new-160-family confirmation. A is query-time acquisition in nats, $D = -L$ is terminal locality damage in percentage points, and brackets are frozen 95% empirical family-reweighting ranges.

Editor	A [95% range]	D (L range)	Install	Retain/install	Full gate
MEMIT-high	8.3016 $[7.8805, 8.7210]$	2.0781 $[-3.2031, -1.1250]$	0.4781	1.0000	PASS
ROME-low	7.5166 $[7.0202, 7.9752]$	4.6406 $[-6.4219, -3.0781]$	0.4172	0.9925	INCONCLUSIVE: locality

The pre-outcome-frozen primary MEMIT-minus-ROME terminal locality-success difference is $+2.5625$ percentage points with 95% range $[+1.0625, +4.2656]$. It lies above zero, while MEMIT's own locality range clears the frozen five-point gate and ROME's does not. MEMIT passes the acquisition, retention, locality, and full gates; ROME passes acquisition and retention, while its locality range crosses the threshold, making locality and the full conjunction INCONCLUSIVE. The frozen `MEMIT_SAFER_REPLICATES` taxonomy depends on MEMIT's absolute pass and on the paired locality range lying above zero; it does not establish that ROME is unsafe beyond this frozen rule and panel. Acquisition is not forced into an equivalence margin: the paired MEMIT-minus-ROME estimate is $+0.7850$ nats with range $[+0.4416, +1.1441]$.

Because the dose-four query is terminal, continuous retention equals acquisition and is not independent; installation and retained-among-installs remain separate gates.

Secondary diagnostics. On preexisting read-only neighborhood outputs, MEMIT has lower terminal full-vocabulary KL than ROME (paired difference -0.094463, 95% range $[-0.151609, -0.044748]$) and a lower greedy-next-token change fraction (-0.111094, $[-0.135156, -0.088750]$). These diagnostics are secondary and do not replace the pre-outcome-frozen locality-success comparison.

Execution integrity and scope. The outcome-blind completion audit found 152 completed runs and eight that failed during model loading, before any update call. A pre-outcome recovery executed exactly those eight logical coordinates with byte-identical worker, model, strength, dose, and estimator. The completion audit therefore covers all 160 logical runs and 1,280 rows; independent replay reproduces the analysis byte-for-byte.

This result meets the frozen relative-locality criterion on one GPT-2XL revision, one CounterFact protocol, one dose, and two selected native-strength editor points. It does not pool the 64-family pilot, erase that pilot's INCONCLUSIVE result, establish an architecture-wide ranking, or validate an online controller. The pilot-powered sample size and editor selection remain material caveats.

C.8 ADAPTIVE ROME-LOW EIGHT-EVENT FACET

Design and chronology. This secondary facet was designed after the four-event ROME grid and the MEMIT-high eight-event result were known. It therefore tests a newly frozen ROME-low execution on a new structural extension; it is not an independent pre-MEMIT comparison. Sixteen disjoint CounterFact family pairs yield 32 focal orientations. For each focal orientation and order seed, the four doses share the exact eight-event sequence, giving 16 runs over the complete grid and 512 rows. The inferential block is the disjoint family pair after averaging its two orientations and four seeds. Max-deviation ranges use 10,000 identical paired block-bootstrap draws with PCG64 seed 2026100501.

Table 15: Separately frozen ROME-low shared-eight-event facet. A, R, and L denote acquisition, retention, and locality; A is in nats and $D = -L$ is terminal locality damage in percentage points. The locality endpoint and gate values are identical at all four doses because the full ordered event sequence and terminal state are identical. "Install / retain" reports installation and retention among installs.

Dose	A	95% range for A	D	Install / retain	Gates
1	6.37607	$[5.45820, 7.23267]$	6.71875	0.3125 / 0.8000	A, R PASS; L, full INCONCLUSIVE
2	6.37283	$[5.44759, 7.24550]$	6.71875	0.3125 / 0.8000	A, R PASS; L, full INCONCLUSIVE
4	6.28910	$[5.38021, 7.19332]$	6.71875	0.3125 / 0.8000	A, R PASS; L, full INCONCLUSIVE
8	5.60322	$[4.51207, 6.67574]$	6.71875	0.3125 / 0.8000	A, R PASS; L, full INCONCLUSIVE

The frozen taxonomy is INCONCLUSIVE. The observed acquisition range is 0.77284 nats, but its max-deviation 95% range $[0.26121, 1.28448]$ crosses the 0.5-nat practical threshold. Dose 1 is the descriptive peak, but its $d_1 - d_2 = +0.00324$-nat contrast has max-deviation range $[-0.50840, +0.51488]$ and is unresolved. The larger descriptive fall to dose 8 thus cannot satisfy the pre-outcome-frozen `DOSE_SENSITIVE` conjunction, which requires both a range whose lower endpoint exceeds the practical threshold and a winner contrast whose range excludes zero.

Terminal-minus-reset retention is 5.60322 nats. The locality-success range is $[-12.65625, -1.48437]$ points at every dose. Acquisition and retention pass at every dose, whereas locality and the full conjunction are inconclusive throughout because the range crosses the five-point threshold. For every seed and focal orientation, collection verifies exact cross-dose equality of the event order, reset/post-focal/terminal endpoints, retention, locality, recall, and terminal parameter hash. Only acquisition at the query checkpoint varies. Independent reanalysis reproduces the frozen analysis byte-for-byte.

This facet is not pooled with either the four-event editor grid or the MEMIT-high eight-event facet. It does not establish agreement between ROME and MEMIT. Because one editor/strength's unresolved dose surface and terminal gate status do not identify a general relationship between schedule sensitivity and writer quality, it is not a writer-quality diagnostic.

C.9 POST-HOC RELATION-CLUSTER SENSITIVITY

The frozen new-160 primary remains the paired family bootstrap: MEMIT-high minus ROME-low terminal locality success is $+2.5625$ percentage points with a 95% family-reweighting range $[+1.0625, +4.265625]$. After outcomes were known, we audited whether repeated CounterFact relations dominate that range. The 160 focal families contain 33 observed relations; 30 relations repeat, three are singletons, 157/160 families lie in a repeated relation, and the largest relation contains 12 families.

Table 16: Post-hoc relation-aware sensitivity. Both ranges use a two-sided Student-t critical value with 32 degrees of freedom. CR1 denotes the cluster-robust finite-sample correction. Neither range replaces the frozen family-bootstrap primary; pp denotes percentage points.

Sensitivity target	Estimate (pp)	SE (pp)	95% range (pp)
Family-weighted point, relation-cluster CR1	$+2.5625$	0.9353	$[+0.6574, +4.4676]$
Equal weight over 33 observed relation means	$+2.3621$	0.9405	$[+0.4463, +4.2779]$

The CR1 check preserves the family-weighted point estimate and changes only the uncertainty calculation. The equal-relation check changes the descriptive target to equal weight over the 33 observed relation means. Both post-hoc ranges lie above zero on this fixed GPT-2XL CounterFact panel, so repeated focal relations do not explain away the observed advantage under these two checks. They do not create a pre-outcome-frozen relation-population claim, establish a universal editor ranking, or generalize beyond the selected model, benchmark, strengths, and endpoint. The frozen paired-family bootstrap remains the sole frozen decision summary.

C.10 Additional-editor dose breadth

Frozen question and implementation check. On the inherited same-sequence eight-event GPT-2XL/CounterFact construction, we prospectively predicted SAFE+INSENSITIVE for AlphaEdit (Fang et al., 2025) and FAILING+SENSITIVE for EasyEdit's native constrained fine-tuning baseline (FT-L). The first AlphaEdit implementation reset a module-global cache on every singleton and therefore did not execute faithful sequential AlphaEdit. Those runs were stopped and excluded before outcome access. The corrected implementation resets the cache only at sequence boundaries and carries it across writes; it followed a pre-outcome implementation check but retained the frozen predictions and estimator. FT-L was unaffected. Both valid cohorts use exact per-dose replay and the inherited acquisition, retention, and locality gates.

Table 17: Additional-editor breadth on the shared eight-event construction. "Acquisition range" is the max–min difference across doses; brackets are its max-deviation-adjusted 95% reweighting band. "Safe" denotes this CounterFact neighborhood protocol only; pp denotes percentage points.

Writer	Frozen prediction	Realized class	Acquisition range (nats)	Locality damage (pp)	Full-pass doses
AlphaEdit	safe + insensitive	safe + insensitive	0.00122 $[0, 0.01301]$	1.5625	$4/4$
FT-L	failing + sensitive	failing + inconclusive	0.61832 $[0.08915, 1.14749]$	12.8125	$0/4$

AlphaEdit's dose-specific acquisition values at $d = 1, 2, 4, 8$ are 7.719544, 7.719592, 7.719317, and 7.718369 nats. Its joint-install rate is 0.5625, retained-per-install is 1.0, and every dose clears the full gate. FT-L's corresponding acquisition values are 0.276407, 0.246447, 0.133071, and -0.341910 nats. It has zero joint installs and no full-pass dose. FT-L's max-deviation range straddles the frozen 0.5-nat margin, so its categorical sensitivity decision is INCONCLUSIVE; independently, zero joint installations and zero full-pass doses make its frozen behavioral conjunction fail.

The joint signature is therefore `INCONCLUSIVE_NOT_FALSIFIED`, not a confirmation. AlphaEdit realizes its positive prediction; FT-L does not confirm its predicted sensitivity class. No tested published editor conclusively occupies the unsafe-and-dose-sensitive quadrant, and no result licenses a general safety claim or causal dose–quality law.

Across the shared-eight-event MEMIT, ROME, AlphaEdit, and FT-L facets, the dose-eight query is terminal: continuous retention there equals acquisition and is zero-lag, whereas doses one, two, and four include intervening edits.

D ENDPOINT TRANSPORT

The GPT-2XL and Llama results below preserve endpoint-specific estimands, pointwise reweighting ranges, absolute gates, execution boundaries, and timing status. In tables, "WikiData-CF" denotes WikiData-CounterFact.

D.1 FORMAL COMPARISON ACROSS SETTINGS

Fix a reference contrast q_0. For each endpoint–bundle cell $j = (c_j, e_j) \in \mathcal{J}$, a prespecified correspondence $\iota_j : \mathrm{supp}(q_0) \to \Pi_{c_j}$ induces the cell contrast

$$q_j(\pi) = \sum_{\pi_0 : \iota_j(\pi_0) = \pi} q_0(\pi_0). \tag{29}$$

The comparison is UNDEFINED unless each q_j is nonzero and preserves the declared positive-versus-negative orientation. For the endpoint-specific metric m_j, let $\widehat{\Delta}_j^{m_j}(q_j) = 2\|q_j\|_1^{-1} \sum_{\pi \in \Pi_{c_j}} q_{j,\pi} \widehat{\mu}_j^{m_j}(\pi)$ have valid range $I_j = [\ell_j, u_j]$, and set

$$\rho_j = \begin{cases} +, & \ell_j > 0, \\ -, & u_j < 0, \\ ?, & 0 \in I_j, \end{cases} \qquad \boldsymbol{\sigma}_{\mathcal{J}}^{\mathbf{m}}(q_0; \{\iota_j\}_{j \in \mathcal{J}}) = (\rho_j)_{j \in \mathcal{J}}. \tag{30}$$

A ? is unresolved, not a tie; endpoint-specific metrics are not pooled into a common utility. A missing or invalid range makes ρ_j undefined.

Remark 2 (Off-support comparison). Let $\mathcal{H}$ be the evaluated setting–policy–endpoint support and $x^\dagger \notin \mathcal{H}$. Over unrestricted bounded response surfaces whose outcome range contains at least two values, the record on $\mathcal{H}$ does not identify the response at $x^\dagger$: two surfaces can agree everywhere on $\mathcal{H}$ and differ at $x^\dagger$. A source ordering therefore needs a stated invariance condition or new target evidence. Likewise, resolved signs $\rho_j = +$ and $\rho_{j'} = -$ reject a constant-sign hypothesis for the mapped contrast, while unresolved cells remain explicit. Because the measurements are endpoint-specific, this is not a cross-instrument utility ranking.

The construction is proved in Appendix F.9.

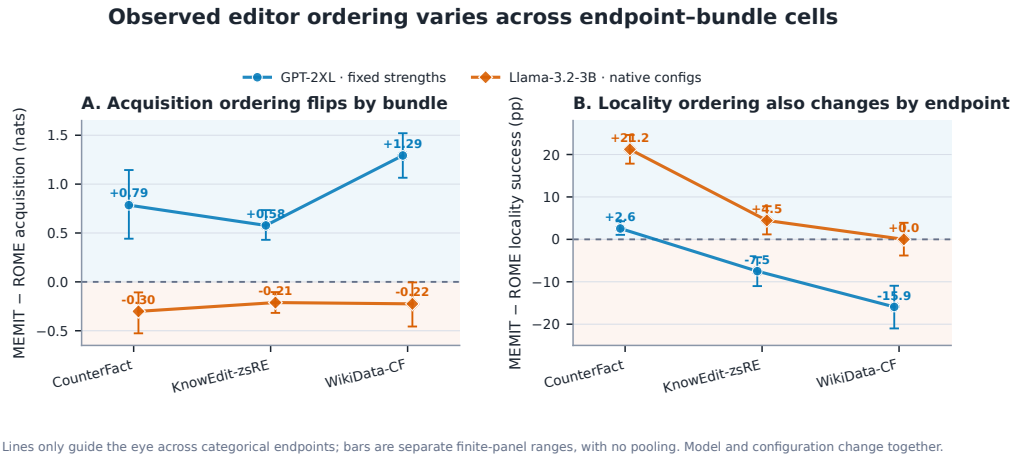


Figure 6: **Observed editor contrasts vary across evaluated endpoint–bundle cells.** Values are MEMIT-minus-ROME contrasts with separate pointwise paired-family ranges (160 families for CounterFact and zsRE; 128 for WikiData-CounterFact). Connecting lines only guide the eye across categorical endpoints. Because model, configuration, benchmark, and probe change together, the trajectories document heterogeneity rather than isolate its cause.

D.2 PROSPECTIVE WIKIRECENT FOUR-RECORD EXECUTION

This study executes the complete audit after freezing the panel, policies, trace compiler, estimands, family-reweighting draws, gates, and transport rule. It is a knowledge-insertion study: each focal item has a new target, canonical update prompt, one source-provided rephrase, and two relation-specificity probes, but no old focal target. It is therefore not pooled with the three replacement-task endpoint studies above.

Frozen design and resolution. The deterministic panel contains 64 families of four distinct edits. Four cyclic Latin orders place each focal edit at every rank. MEMIT-high (`mom2_update_weight` $= 6325$) and ROME-low (`clamp_norm_factor` $= 2$) each execute four singleton GPT-2XL outer calls, yielding $64 \times 4 \times 2 = 512$ policy rows and 256 paired family–order units. Both held-out preflight examples passed before evaluation outcomes were opened. Every executed row matched its precomputed score-ancestral outer-call certificate, so the planned MEMIT-minus-ROME contrast has $s_Z = 1$ for the canonical and rephrase scores.

Before outcomes, the allocator read only KnowEdit WikiRecent revision `6d218881e60c` and GPT-2XL tokenizer revision `15ea56dee5df`; it used no model scores or editor outputs. Schema, probe, and token constraints preceded exclusion of prior-panel overlap in normalized subject, update prompt, new target, or source-record hash. It then sorted by SHA-256(namespace|source index) and greedily selected the first 256 rows with globally unique subjects, target strings and token sequences, and all update, rephrase, and locality prompts. A four-item preflight panel was disjoint from the evaluation panel under the same identifiers. The two policies were inherited from earlier GPT-2XL work and fixed before WikiRecent outcomes: this confirms the workflow, not blinded policy discovery. The reviewer artifact binds the full revisions by hash.

Table 18: **Prospectively frozen effect and transport record.** Entries are mean dose-four-minus-reset per-target-token log probabilities in nats. Brackets are 95% empirical family-reweighting ranges from the fixed 64-family panel, not population confidence intervals.

Context	MEMIT-high	ROME-low	Mean paired difference	Effect status
Canonical prompt	4.1261	3.9119	$+0.2142$ $[+0.1079, +0.3369]$	PASS
One rephrase	2.9104	2.1925	$+0.7178$ $[+0.5530, +0.8903]$	PASS

The pre-outcome-frozen transport predicate preserves only the resolved mean direction under identity policy correspondence. Both ranges lie above zero, so it passes. Descriptive post-hoc heterogeneity checks show that on canonical prompts 35/64 family means and 115/256 order-pair contrasts are positive (141 are negative), the family median is $+0.0057$, and the focal-rank-four mean is -0.0158 nats. On rephrases 58/64 family means are positive, but only 33/64 are positive in both contexts. This is aggregate-direction transport, not family-, prompt-, order-, or rank-wise dominance.

Absolute gates. Table 19 keeps the relative effect separate from each policy's acquisition, installation, retention-among-installs, and shared locality records. Every non-locality component passes its pre-outcome-frozen rule. ROME's rephrase installation point is 20.70%, just above the 20% threshold; its family-reweighting range $[11.72\%, 30.47\%]$ crosses that threshold because the frozen installation gate is point-estimate based.

Table 19: **Absolute feasibility fails despite transported mean acquisition.** Acquisition is in nats; installation and retained-among-installs are proportions; locality is terminal-minus-reset relation-specificity success in percentage points. All entries are prospective.

Context	Policy	Acquisition	Install	Retain/inst.	Locality	Full
Canonical	MEMIT	4.1261	.8867	.9912	-43.36 $[-52.73, -33.98]$	FAIL
Canonical	ROME	3.9119	.8047	.9563	-41.41 $[-50.78, -32.23]$	FAIL
Rephrase	MEMIT	2.9104	.3828	1.0000	-43.36 $[-52.73, -33.98]$	FAIL
Rephrase	ROME	2.1925	.2070	.9811	-41.41 $[-50.78, -32.23]$	FAIL

Locality here is deliberately narrow. Each of two source-supplied relation-specificity probes contributes an indicator of whether the source-correct answer has lower token-average negative

log-likelihood (NLL); their mean is the row score in both prompt contexts. The large negative changes cross the pre-outcome-frozen -5-point boundary for both policies. Consequently all four policy–context feasibility conjunctions, the designated MEMIT target-feasibility claim, and EvidenceComplete are FAIL even though mean direction transport is PASS.

On the same two probes and 256 family–order rows per editor, post-hoc terminal full-vocabulary Kullback–Leibler (KL) divergence is 2.2944/2.4154 and greedy next-token change is 73.63%/77.73% for MEMIT/ROME. Neither metric was a frozen gate; because both reuse the same probes, they add no probe breadth and support no population claim.

Continuous acquisition and terminal-minus-reset retention are numerically the same dose-four checkpoint and provide only one continuous confirmation. The prefix-retention summary also contains a zero-lag stratum when the focal edit is fourth. Restricting descriptively to focal ranks 1–3, canonical retention is 169/171 for MEMIT and 140/149 for ROME; rephrase retention is 73/73 and 37/38. The rank-four zero-lag counts for MEMIT and ROME are 56/56 and 57/57 on canonical prompts and 25/25 and 15/15 on rephrases. Costs are recorded, but no common budget or scalarization was declared and no cost gate is active.

Outcome-access and provenance boundary. The source manifest and freeze preceded both preflight runs. Their complete two-of-two decision had already bound the terminal, output, and collection hashes when a relative-path publication call failed. An independently authorized absolute-path publication-only recovery published those unchanged bytes; neither preflight run was repeated or recollected. The evaluation then proceeded. The scheduler submitted 63 of 64 runs; the remaining run's metadata lookup timed out before submission. Exact name/history and result-root checks found it absent before any evaluation-output access, after which an independently authorized submission-only recovery created the initial attempt. All 64 runs then reached the exact terminal state at the original scheduler version and attempt number, with no restarts, exit code 0, and no terminating signal before the one-shot collection.

The manuscript evidence binding preserves the original partial-submission receipt, recovery authorization and receipt, collection authorization, completion audit, output manifest, collected-row digest, and final record. This matters because the final record itself binds the terminal and output chain but does not directly hash the earlier submission incident. A detached assembly produced a byte-identical pretty-printed four-record file (SHA-256 prefix `87bb2a42`); the assembler's printed compact-canonical record digest has a different prefix (`087fa824`). They are distinct hashes of distinct serializations, not conflicting results.

A clean checkout of the freeze commit passes all 15 prelaunch tests. Post-result closure, independent raw-row recomputation, and detached assembly also pass. Remote receipt/completion wrappers are absent from the local flattened artifact, so their hashes and chronology rely on the frozen collector and captured scheduler records; neither detached assembly alone nor the reviewer capsule independently reproduces the scheduler.

D.3 GPT-2XL ENDPOINT-NATIVE TRANSPORT

Three studies separately froze a fixed GPT-2XL operating point and tested it with endpoint-native estimands. They are three fixed settings, not exchangeable replicates: CounterFact uses two unseen rephrases and ten neighborhood prompts; KnowEdit-zsRE uses one rephrase and two relation-specificity prompts; and KnowEdit WikiData-CounterFact uses one rephrase and the first two canonical source-ordered relation probes. Dataset and locality instrument therefore vary jointly. All empirical reweighting ranges are computed separately by endpoint; no pooled interaction is estimated, and the contrasts cannot identify a causal dataset or instrument effect.

Frozen operating point. All three studies use the public Hugging Face GPT-2XL revision `15ea56dee5df4983c59b2538573817e1667135e2`, MEMIT-high (`mom2_update_weight=6325`), ROME-low (`clamp_norm_factor=2`), four sequential singleton edits, and query/terminal dose $d = 4$. The inferential family counts are 160 for CounterFact, 160 for zsRE, and 128 for WikiData-CounterFact.

All runs use EasyEdit revision `14cea8245f06715684592ab55184939b99d70784`, with `subject_last`, value-loss layer 47, value learning rate 0.5, config-

uration field `v_num_grad_steps=20`, value weight decay 0.5, KL factor 0.0625, `transformer.h.{}.attn`, `transformer.h.{}.mlp`, and `transformer.h.{}.mlp.c_proj` modules, `transformer.h.{}` layer modules, `transformer.ln_f`, `transformer.wte`, Wikipedia moment statistics with 100,000 samples in float32 from the frozen statistics directory, device 0, and model parallelism disabled. MEMIT-high edits layers 13–17 with all-layer selection, moment adjustment enabled, moment-update weight 6325, and clamp factor 0.75. ROME-low edits layer 17 with moment adjustment disabled, clamp factor 2, fp16 disabled, and context-template length/count pairs $(5, 10)$ and $(10, 10)$.

Estimands, reweighting ranges, and gates. At reset and terminal, each relation probe is paired with a canonical source answer and the focal edited target. We prepend one space to each continuation, tokenize it without special tokens, sum its autoregressive token negative log-likelihood, and divide by the number of continuation tokens. A probe succeeds exactly when the source-correct answer has lower token-average negative log-likelihood than the edited target. The compiler deterministically keeps the first nonempty source-answer string (or the first string in the first list); it neither maximizes nor marginalizes over aliases. Probe indicators are averaged within row, then the four Latin-order/seed replicates are averaged within family at reset and terminal. Locality is $L = 100(\text{terminal success} - \text{reset success})$.

Each endpoint uses a paired family bootstrap: the same resampled family indices are used for both editors, with 10,000 draws and linearly interpolated 2.5th and 97.5th percentiles. CounterFact uses NumPy PCG64 seed `2026100601`; zsRE and WikiData-CounterFact use seed `2026091605`. The acquisition gate requires its range's lower endpoint to exceed zero. Retention requires its range's lower endpoint to exceed zero, joint install $\geq 20\%$, and retained per install $\geq 50\%$; locality requires its point estimate to be at least -5 points and its range's lower endpoint to exceed -5.

The per-editor 95% inputs to those absolute gates are as follows. On CounterFact, MEMIT has acquisition/retention ranges $[7.880546, 8.721021]$, locality $[-3.203125, -1.125000]$, joint install $[0.401563, 0.554688]$, and retained per install $[1, 1]$; ROME has $[7.020214, 7.975242]$, $[-6.421875, -3.078125]$, $[0.346875, 0.487500]$, and $[0.975709, 1]$. On zsRE, the corresponding MEMIT ranges are $[4.121185, 4.816444]$, $[-37.343750, -25.390625]$, $[0.509375, 0.657813]$, and $[0.991279, 1]$; ROME's are $[3.555082, 4.225588]$, $[-29.140625, -18.671875]$, $[0.351523, 0.485938]$, and $[0.963454, 0.996032]$. On WikiData-CounterFact, MEMIT has $[3.632744, 4.373639]$, $[-75.878906, -63.183594]$, $[0.072266, 0.183594]$, and $[0.947368, 1]$; ROME has $[2.376302, 3.042119]$, $[-60.839844, -46.579590]$, $[0.031250, 0.101563]$, and $[0.884615, 1]$.

Table 20: Fixed-strength GPT-2XL results under separately frozen endpoint-native instruments. Acquisition A is in nats and damage D in percentage points. The final column is the within-endpoint MEMIT-minus-ROME locality-success contrast with a paired-family 95% empirical reweighting range.

Endpoint	Editor	A	D	MEMIT–ROME locality success
CounterFact	MEMIT-high	8.3016	2.0781	$+2.5625$ $[+1.0625, +4.2656]$
CounterFact	ROME-low	7.5166	4.6406	—
KnowEdit-zsRE	MEMIT-high	4.457329	31.328125	-7.5000 $[-11.015625, -4.218750]$
KnowEdit-zsRE	ROME-low	3.880083	23.828125	—
KnowEdit-WikiData-CF	MEMIT-high	3.997292	69.726562	-15.9180 $[-20.9961, -10.9375]$
KnowEdit-WikiData-CF	ROME-low	2.704324	53.808594	—

Results. MEMIT's acquisition advantage is positive in all three settings: $+0.7850266$ nats $[+0.4416111, +1.1441018]$ on CounterFact, $+0.5772456$ $[+0.4301146, +0.7341698]$ on zsRE, and $+1.2929676$ $[+1.0650381, +1.5207201]$ on WikiData-CounterFact. The locality-success contrasts instead are $+2.5625$, -7.5000, and -15.9180 points, respectively. Thus the locality ranking reverses on two endpoints, whereas the acquisition ranking does not.

The third study froze `WIKICOUNTERFACT_MEMIT_LOCALITY_ADVANTAGE` before preflight or study outcome access; its realized taxonomy is `WIKICOUNTERFACT_REVERSED`. The zsRE replay analogously returns `ZSRE_REVERSED`. These outcomes are reported under their frozen estimators rather than a post-outcome replacement metric.

On zsRE, both editors fail the absolute locality gate and therefore the behavioral acquisition/retention/locality conjunction. MEMIT installs 58.4375% of focal cases and retains 99.7326% per install; ROME installs 41.7188% and retains 98.1273% per install. On WikiData-CounterFact, both acquisition gates pass, but both retention gates fail because joint install is 12.50% for MEMIT and 6.25% for ROME, below the frozen 20% floor. Retained per install is 98.4375% and 96.8750%, respectively, but cannot rescue that failed conjunct. Both the locality gate and the behavioral acquisition/retention/locality conjunction fail. The relative ROME advantage therefore does not make ROME safe. In all three panels, the dose-four query is terminal, so continuous acquisition and retention are numerically identical by construction rather than independent confirmations.

Execution and evidence authority. Across the three cohorts, a retrospective, outcome-value-blind source-and-receipt replay reproduces all 448 compiled-cell digests and verifies their receipt bindings. Direct comparison of full request objects gives 1,792 matched family–seed units, 3,584 ROME/MEMIT score-ancestral traces, 14,336 score-ancestral singleton calls, and another 14,336 post-reset calls for terminal validation. Receipts bind compiled-cell, worker, source-manifest, and runtime hashes. Completion audits establish exact accepted-cell closure: CounterFact includes eight pre-outcome byte-identical replacements after model-loading failures (Appendix C.7); the KnowEdit cohorts had none. This is source-and-receipt-bound outer-call reconstruction, not an observed per-call ledger. The runs did not record per-event parameter hashes, random-number-generator snapshots, or EasyEdit-internal optimization/solve steps, and the configured step count above is not an observed internal-step ledger. Raw receipt bodies remain private, so the sanitized reviewer capsule exposes aggregate counts and the certificate digest but does not independently re-project those receipts.

All 160/160 zsRE and 128/128 WikiData-CounterFact planned runs completed without retry or replacement, producing 1,280/1,280 and 1,024/1,024 expected editor–family–order rows. Independent replay reconstructs closure, estimates, ranges, and the `ZSRE_REVERSED` and `WIKICOUNTERFACT_REVERSED` taxonomies. The pinned zsRE file contains 1,301 rows although its source card reports 1,230; the pinned WikiData-CounterFact file contains 839 although its card reports 885. Each complete pinned file passed its schema and hash audit and was used without silent truncation.

For WikiData-CounterFact, the outcome-blind eligible pool contained 589 globally disjoint rows. The design froze 128 families (512 update records), the largest launch-compatible power of two, without outcome-driven resizing. Selected families have zero normalized subject, update-prompt, new-target, or source-record overlap with the prior panels; 69 old-target strings repeat. Source manifests, completion audits, collected rows, analyses, and replays are bound in the immutable evidence ledger (SHA-256 `57a4df961e7b6a9090ce813a266f4eb94a306ff6ae45c68b351c9792dd5dfdaf`). The three positive acquisition contrasts and $(+2.5625, -7.5000, -15.9180)$ locality pattern are evidence from three fixed bundles, not a prevalence estimate or a causal endpoint effect.

D.4 Pre-outcome-frozen Llama-3.2-3B recurrence test

Frozen question and model bundle. Before preflight or outcomes, we froze the question of whether GPT-2XL's MEMIT-minus-ROME locality signature $(+, -, -)$ would recur, together with the three panels, primary binary-locality estimator, paired family bootstrap, prediction, and interpretation branches. The bundle uses `unsloth/Llama-3.2-3B` revision `d4446454d87d51aa42e1fb174f25acc5f8762331`, shipped-native ROME `clamp_norm_factor=4`, and MEMIT `mom2_update_weight=15000`, without endpoint-specific strength selection. Model and native configuration therefore change jointly; this is not an isolated model-effect test.

Execution and authority. All 448 predeclared runs completed on their first attempt, without retry or failure, yielding 3,584 rows. Before collection, a completion audit certified 448/448 logical identities; the collector checked every embedded source, plan, result, receipt, and completion binding. Independent reanalysis is byte-identical, and all 15 protocol tests pass. Source-manifest and analysis SHA-256 prefixes are `a832811e85d0` and `122aeca1c2fb`; the machine record retains both complete digests.

Later outer-call reconstruction. A post-outcome structural projection that excludes outcome fields validates 3,584/3,584 declared atomic outer-call rows. At the dose-four/terminal score cut, all 1,792 paired MEMIT–ROME traces separate ($s_Z = 1$); reset is aliased and the balanced pre-focal mixture has $s_Z = 3/4$. EasyEdit internals and a cross-bundle correspondence record remain unavailable. At $s_Z = 1$ the bounded-effect certificate spans the full outcome range, so this rules out only declared outer-call aliasing. The result is a target-resolution diagnostic, not a complete formal transport lift.

Table 21: **Frozen Llama-3.2-3B endpoint results.** Editor-level signed locality (M: MEMIT; R: ROME) and paired M-minus-R contrasts. L is terminal-minus-reset success (points); A is acquisition (nats). Brackets are pointwise 95% empirical family-reweighting ranges (10,000 paired draws); ordering uses paired contrasts.

Endpoint	Families	Editor-level L [95% range]	Paired ΔL [95% range]	Paired ΔA [95% range]
CounterFact	160	M: -7.9531 $[-10.7188, -5.4688]$ R: -29.1719 $[-33.4688, -25.0313]$	$+21.2188$ $[+17.8438, +24.6094]$	-0.3017 $[-0.5264, -0.1076]$
KnowEdit-zsRE	160	M: -43.0469 $[-49.2969, -36.8750]$ R: -47.5000 $[-53.7500, -41.3281]$	$+4.4531$ $[+1.1719, +7.8125]$	-0.2116 $[-0.3173, -0.1046]$
KnowEdit WikiData-CounterFact	128	M: -70.5078 $[-75.7813, -65.1367]$ R: -70.5078 $[-75.4883, -65.4297]$	0.0000 $[-3.8086, +3.9063]$	-0.2247 $[-0.4566, -0.0052]$

Frozen decision. The predicted locality signature was $(+, -, -)$ and the realized one is $(+, +, ?)$. WikiData is unresolved, so the primary call is `LLAMA_PATTERN_INCONCLUSIVE`. Excluding whole families whose new-target first subtoken appears in their update prompt leaves 160/152/124 CounterFact/zsRE/WikiData-CounterFact families and gives the same signature. All six method–endpoint points pass acquisition and adjacent terminal retention but fail locality and the full conjunction; retention here duplicates acquisition. Thus positive zsRE defeats exact recurrence, while unresolved WikiData retains the INCONCLUSIVE pattern call. Joint model and native-configuration changes preclude architecture-specific attribution.

D.5 One field-complete adapter-relative assembly

Table 22 retrospectively assembles the fixed GPT-2XL comparison while retaining each endpoint's timing label. "Field-complete" means that typed fields are populated, not that EvidenceComplete passes.

Table 22: **Field-complete four-record assembly.** Resolution is conditional on the outer-call adapter; cost has no declared gate; H is effect-decision status.

Record	Bound evidence	Status
Resolution	Outer-call reconstruction gives $s_Z = 1$ for 640 CounterFact, 640 zsRE, and 512 WikiData-CF family–order pairs.	conditional PASS; diagnostic
Effect	Source acquisition/locality contrasts are $+0.7850$ $[+0.4416, +1.1441]$ nats and $+2.5625$ $[+1.0625, +4.2656]$ points; endpoint signs are $(+, +, +)/(+, -, -)$.	Acq. H: UNDEFINED (no directional gate); locality H: PASS/FAIL/FAIL
Feasibility	CounterFact: MEMIT passes and ROME is INCONCLUSIVE; both fail locality on zsRE and retention/locality on WikiData-CF. Costs are ungated.	source mixed; targets FAIL
Transport	The correspondence fixes GPT-2XL, method identities, strengths, four-call template, dose, and orientation while endpoint data/instruments change. Locality reverses to -7.5000 $[-11.0156, -4.2188]$ and -15.9180 $[-20.9961, -10.9375]$.	FAIL

Timing and integrity labels remain endpoint-specific. The diagnostic, retrospective reconstruction binds the worker, equal requests, operator identities, and four synchronous calls—not a per-call ledger, EasyEdit internals, random-number-generator snapshots, event hashes, or complete state. Transport and the claim ladder fail; MEMIT fails target locality twice. Endpoint and locality instrument co-vary, so no dataset-only causal effect is identified.

D.6 Post-hoc across-endpoint sensitivity for the Llama signs

The frozen primary uses separate pointwise 95% paired-family ranges. Post-hoc Bonferroni ranges use endpoint-specific reweighting percentiles 0.8333 and 99.1667 (10,000 PCG64 draws, seed 2026091905), without pooling. They test simultaneous sign stability, not population coverage.

Table 23: **Bonferroni sensitivity.** MEMIT-minus-ROME 98.333% marginal ranges span three endpoints; no population coverage is claimed.

Endpoint	Locality success (points)	Acquisition (nats)
CounterFact	$+21.2188$ $[+17.2031, +25.4480]$	-0.3017 $[-0.5827, -0.0730]$
KnowEdit-zsRE	$+4.4531$ $[+0.3906, +8.5156]$	-0.2116 $[-0.3405, -0.0808]$
WikiData-CounterFact	0.0000 $[-4.6875, +4.8828]$	-0.2247 $[-0.5071, +0.0441]$

Adjusted locality remains $(+, +, ?)$: WikiData is unresolved and positive zsRE defeats frozen $(+, -, -)$ recurrence. Acquisition becomes $(-, -, ?)$, so $(-, -, -)$ is only pointwise. The endpointwise rules were prespecified; Bonferroni is a post-hoc simultaneous-stability test.

D.7 Post-hoc cluster-pairs sensitivity

We resample paired family differences by relation cluster: 33 native CounterFact IDs and 105 zsRE/27 WikiData-CounterFact exact subject-redacted templates. KnowEdit lacks semantic IDs, so templates may split paraphrases.

Table 24: **Post-hoc cluster-pairs sensitivity.** MEMIT-minus-ROME points have Bonferroni-indexed 98.33% ranges within each model–metric family; CounterFact uses native IDs and KnowEdit exact templates.

Bundle	Endpoint	Acquisition (nats)	Locality (points)
GPT-2XL fixed	CounterFact	$+0.7850$ $[+0.3779, +1.3489]$	$+2.5625$ $[+0.5719, +4.9598]$
	zsRE	$+0.5772$ $[+0.4149, +0.7682]$	-7.5000 $[-12.0295, -3.2389]$
	WikiData-CF	$+1.2930$ $[+1.0491, +1.5279]$	-15.9180 $[-23.4375, -10.1958]$
Llama native	CounterFact	-0.3017 $[-0.6196, -0.0818]$	$+21.2188$ $[+16.9091, +25.5673]$
	zsRE	-0.2116 $[-0.3327, -0.0931]$	$+4.4531$ $[+0.4580, +8.7248]$
	WikiData-CF	-0.2247 $[-0.4933, +0.1092]$	0.0000 $[-5.4314, +6.5476]$

All six GPT-2XL ranges retain their sign pair $(+, +, +)/(+, -, -)$. Llama zsRE locality remains positive, WikiData remains unresolved, and acquisition is $(-, -, ?)$. This checks exact relation-form dependence, not arbitrary semantic-relation or model-population inference.

D.8 Summary of invalid shortcuts

Table 25: **Four unsupported shortcuts and evidence-based alternatives.**

Unsupported shortcut	Relevant evidence	Evaluation alternative
Equal update counts imply the same pre-answer history	24 schedule labels produce 1, 4, 12, or 24 history classes	Record the ordered updates before each answer
Immediate learning implies durable, local learning	LoRA: acquisition passes 6/6; retention and locality pass 0/6	Test acquisition, retention, and locality separately
A method ordering carries over to a new setting	Observed GPT-2XL and Llama-3.2-3B orderings differ across evaluated bundles; the cross-model mapping is incomplete	Evaluate the same methods directly in the new setting
An average improvement implies usability	WikiRecent mean improvement passes while locality fails	Require every prespecified behavioral check to pass

E Reviewer artifact and provenance boundary

The reviewer artifact is released only after automated checks pass. It contains the paper bindings, tracing code, display files rendered from hash-bound aggregate projections, sanitized WikiRecent sufficient statistics, validators, and upstream notices. Its manifest checks hashes, anonymity, declared license labels, PDF and font inventory, bindings, and tests. It omits legacy family-level measurements, raw or restricted data, models, private receipts, randomized policies, and third-party implementation details or histories.

Scope of the combined framework. The framework builds on treatment versions and exposure mappings to distinguish an assigned plan from the updates actually completed; program slicing and provenance to record operations that may affect an answer; model-editing metrics to assess

acquisition, retention, and locality; and transportability work to separate evidence in the original and new settings. Our contribution is to connect these ingredients in one evaluation record. We do not claim to originate the underlying concepts, automatically discover every causal path, or identify causal effects in an unevaluated setting.

F Complete proofs of the formal results

This section supplies the complete proofs referenced in the main text and by the associated robustness discussions. The arguments below state all measure-theoretic and linear-algebraic steps used by the named results. Throughout, total variation uses the convention in Definition 2.

F.1 Deterministic history-resolution formula

Derivation of Equation 3. For row r, let $C_{r,1}, \dots, C_{r,K_r}$ be the history-equivalence classes, so $(D_r^\top q)_k = \sum_{\pi \in C_{r,k}} q_\pi$. Under the positive and negative contrast mixtures,

$$P^{q^\pm}_{R,Z}(r,k) = \frac{2w_r}{\|q\|_1} \sum_{\pi \in C_{r,k}} q^\pm_\pi.$$

Since $q = q^+ - q^-$, subtraction gives

$$P^{q^+}_{R,Z}(r,k) - P^{q^-}_{R,Z}(r,k) = \frac{2w_r}{\|q\|_1}(D_r^\top q)_k.$$

The discrete total-variation identity therefore yields

$$\begin{aligned} \mathrm{TV}(P^{q^+}_{R,Z}, P^{q^-}_{R,Z}) &= \frac{1}{2} \sum_r \sum_{k=1}^{K_r} \left| \frac{2w_r}{\|q\|_1}(D_r^\top q)_k \right| \\ &= \frac{\sum_r w_r \|D_r^\top q\|_1}{\|q\|_1}, \end{aligned}$$

which is Equation 3. □

F.2 Count factorization

Proof of Lemma 1. First suppose that $F(\Gamma) = Q(\nu(\Gamma))$ for every finite word. For any $a, b \in \mathcal{A}$, the words ab and ba have the same count vector. Hence

$$T_b \circ T_a = F(ab) = Q(\nu(ab)) = Q(\nu(ba)) = F(ba) = T_a \circ T_b.$$

Thus the maps commute pairwise.

Conversely, suppose that the maps commute pairwise. Fix an ordering $\mathcal{A} = \{a_1, \dots, a_m\}$. For $n = (n_1, \dots, n_m) \in \mathbb{N}_0^m$, define

$$Q(n) := T_{a_m}^{n_m} \circ \cdots \circ T_{a_1}^{n_1},$$

where the zeroth power is the identity map. Any finite word Γ can be transformed into the canonical word $a_1^{n_1} \cdots a_m^{n_m}$, with $n = \nu(\Gamma)$, by finitely many adjacent swaps. Each adjacent swap preserves the composed map because the corresponding two maps commute. Therefore $F(\Gamma) = Q(\nu(\Gamma))$ for every finite word. This constructs the required factorization and proves sufficiency. □

Restricted-catalogue factorization. Let $\mathcal{W}$ be any restricted catalogue of words. If a factorization $F|_{\mathcal{W}} = Q \circ \nu$ exists, then $F(\Gamma) = F(\Gamma')$ whenever $\Gamma, \Gamma' \in \mathcal{W}$ have the same count vector. Conversely, if this fiberwise equality holds, define $Q(n)$ for $n \in \nu(\mathcal{W})$ to be $F(\Gamma)$ for any $\Gamma \in \mathcal{W}$ with $\nu(\Gamma) = n$. The assumed equality makes this definition independent of the chosen word. Extending Q arbitrarily outside $\nu(\mathcal{W})$ gives the desired factorization on $\mathcal{W}$. □

The proof is map-valued: equality holds for every starting state in $\mathcal{X}$. On a restricted word catalogue, factorization is equivalent only to equality of the maps induced by admitted words sharing a count vector; pairwise commutativity remains sufficient but is no longer necessary.

F.3 HISTORY-RESOLUTION CERTIFICATE

We first record the elementary form of total-variation contraction used below. If P, Q are probability measures and f is measurable with $0 \le f \le 1$, the layer-cake representation gives

$$\int f \,\mathrm{d}(P - Q) = \int_0^1 \{P(f > t) - Q(f > t)\} \,\mathrm{d}t,$$

and hence

$$\left| \int f \,\mathrm{d}(P - Q) \right| \le \mathrm{TV}(P, Q). \tag{31}$$

Proof of Theorem 1. Write $X = (R, Z)$ and $P_+ = P^{q^+}_{R,Z}$, $P_- = P^{q^-}_{R,Z}$. Score sufficiency supplies a single Markov kernel $L(\mathrm{d}y \mid x)$ for every policy in the contrast support. Mixture linearity therefore gives the two joint laws $P_+(\mathrm{d}x)L(\mathrm{d}y \mid x)$ and $P_-(\mathrm{d}x)L(\mathrm{d}y \mid x)$, and their output marginals are $Q_+ = P_+L$ and $Q_- = P_-L$.

For every measurable output event B, the function $f_B(x) = L(B \mid x)$ lies in $[0, 1]$. Equation 31 therefore implies

$$|Q_+(B) - Q_-(B)| = \left| \int f_B \,\mathrm{d}(P_+ - P_-) \right| \le \mathrm{TV}(P_+, P_-).$$

Taking the supremum over B proves the Markov-contraction inequality

$$\mathrm{TV}(Q_+, Q_-) \le \mathrm{TV}(P_+, P_-) = s_Z(q).$$

Because $Y_e \in [a, b]$, the rescaled function $g(y) = (y - a)/(b - a)$ takes values in $[0, 1]$. Applying Equation 31 to Q_+, Q_- and g gives the following; the additive constant a cancels because both measures have total mass one:

$$\begin{aligned} |\Delta^{\mathrm{exec}}_e(q)| &= (b - a) \left| \int g \,\mathrm{d}(Q_+ - Q_-) \right| \\ &\le (b - a)\,\mathrm{TV}(Q_+, Q_-) \\ &\le (b - a) s_Z(q). \end{aligned}$$

Finally, for a measurable summary h, every measurable set C in its codomain satisfies

$$|h_\# P_+(C) - h_\# P_-(C)| = |P_+(h^{-1}(C)) - P_-(h^{-1}(C))| \le \mathrm{TV}(P_+, P_-).$$

Taking the supremum over C proves $\mathrm{TV}(h_\# P_+, h_\# P_-) \le s_Z(q)$. □

Equality condition for the summary contraction. Let $\mu = P_+ - P_-$ and let A_+ be a positive set in a Hahn decomposition of μ. Because μ has total mass zero, $\mu(A_+) = \mathrm{TV}(P_+, P_-)$. If the sign partition is measurable through h, there is a measurable set C in the summary space such that $h^{-1}(C) = A_+$ up to a $|\mu|$-null set. Consequently,

$$h_\# P_+(C) - h_\# P_-(C) = \mu(A_+) = \mathrm{TV}(P_+, P_-).$$

The reverse inequality is the data-processing bound already proved, so equality holds. If (R, Z) is recoverable almost surely from $h(R, Z)$ by a measurable inverse, then every measurable sign set is measurable through h up to a null set, which gives the stated sufficient special case. □

The common-initial-state, coupled-randomness, and complete-history conditions in the main text are operational conditions intended to justify the common kernel L. The proof itself requires that factorization; if an omitted path makes the conditional outcome law depend on the policy after conditioning on (R, Z), the premise fails.

F.4 RESOLUTION FLOOR

Proof of Corollary 1. Theorem 1 and the asserted lower bound give

$$\delta \le |\Delta^{\mathrm{exec}}_e(q)| \le (b - a) s_Z(q).$$

Since $b - a > 0$, division by $b - a$ yields $s_Z(q) \ge \delta/(b - a)$. The contrapositive gives the stated pre-outcome certificate. □

F.5 CONDITIONAL-KERNEL MISMATCH

Proof of Proposition 1. Write $P_+ = P_X^{q^+}$ and $P_- = P_X^{q^-}$, and define

$$Q_+ = P_+K_+, \qquad Q_- = P_-K_-, \qquad Q_+^0 = P_+L, \qquad Q_-^0 = P_-L.$$

For every measurable output event B,

$$\begin{aligned} |Q_+(B) - Q_+^0(B)| &= \left| \int \{K_+(B \mid x) - L(B \mid x)\} \, \mathrm{d}P_+(x) \right| \\ &\leq \int \mathrm{TV}(K_+(\cdot \mid x), L(\cdot \mid x)) \, \mathrm{d}P_+(x) = \epsilon_+. \end{aligned}$$

Taking the supremum over B gives $\mathrm{TV}(Q_+, Q_+^0) \leq \epsilon_+$. The same argument gives $\mathrm{TV}(Q_-^0, Q_-) \leq \epsilon_-$. Markov contraction under the common kernel L gives

$$\mathrm{TV}(Q_+^0, Q_-^0) \leq \mathrm{TV}(P_+, P_-) = s_Z(q).$$

The triangle inequality now yields

$$\mathrm{TV}(Q_+, Q_-) \leq \epsilon_+ + s_Z(q) + \epsilon_-.$$

Total variation between probability measures is at most one, proving the first claimed bound after taking the minimum with one. Applying Equation 31 to $(Y_e - a)/(b - a) \in [0, 1]$ proves the corresponding mean-score bound. □

F.6 SCHEDULE-CLASS STAIRCASE

Proof of Equation 6 and the pairwise-resolution identity. For $H = 0$, every $a_p(H)$ is the empty sequence, so all $4!$ assignments have one common history. For $H = 1$, $a_1(H) = (1)$ while $a_2(H) = a_4(H) = a_8(H) = (0)$. A history is therefore determined only by the band receiving period one, giving four classes.

For $H \in \{2, 3\}$, the sequences for periods one and two are distinct, while $a_4(H) = a_8(H)$ is the all-zero sequence. Choosing the band receiving period one and then the band receiving period two determines a history; swapping periods four and eight does not change it. Hence there are $4 \cdot 3 = 12$ classes.

For every $H \geq 4$, the entries at opportunities $1, 2, 4$ distinguish all four periods: period one is active at opportunity one, period two is first active at two, period four is first active at four, and period eight is inactive at all three. Thus the four period sequences are distinct, so their assignments to four labeled bands give $4! = 24$ histories.

For a deterministic pair $q = \mathbf{e}_s - \mathbf{e}_{s'}$, the two joint laws put the same mass w_r on row r. If the two rowwise histories coincide, that row contributes zero to their total variation. If they differ, the two distinct atoms each contribute w_r to the ℓ_1 difference and hence contribute w_r after the factor one half. Summing over rows gives

$$s_Z(q) = \sum_r w_r \mathbf{1}\{z_{r,s,e} \neq z_{r,s',e}\} = 1 - \sum_r w_r \mathbf{1}\{z_{r,s,e} = z_{r,s',e}\}.$$

□

F.7 WORKING-BASIS ESTIMABILITY

Proof of Proposition 2. Each row of D_r contains exactly one entry equal to 1, and each of its K_r columns is nonempty. Consequently, $D_r^\top D_r$ is diagonal with positive diagonal entries and $P_r = D_r(D_r^\top D_r)^\dagger D_r^\top$ is the orthogonal projector onto $\mathrm{col}(D_r)$. Because $\mathbf{1} = D_r \mathbf{1}_{K_r}$, it also satisfies $P_r\mathbf{1} = \mathbf{1}$.

Define the observation operator

$$\mathcal{O}v = (\sqrt{w_r} P_r v)_{r:w_r>0}.$$

Two parameter pairs are observationally equivalent exactly when their difference $(\delta\alpha, h)$ satisfies $\mathcal{O}(\delta\alpha\mathbf{1} + Uh) = 0$. Since every P_r is symmetric and idempotent,

$$\begin{aligned}\|\mathcal{O}(\delta\alpha\mathbf{1} + Uh)\|^2 &= \sum_r w_r \|P_r(\delta\alpha\mathbf{1} + Uh)\|^2 \\ &= |\Pi|(\delta\alpha)^2 + 2\delta\alpha \sum_r w_r \mathbf{1}^\top P_r U h + h^\top U^\top \left(\sum_r w_r P_r\right) U h \\ &= |\Pi|(\delta\alpha)^2 + h^\top I_U h.\end{aligned}$$

The cross term vanishes because $\mathbf{1}^\top P_r U = (P_r\mathbf{1})^\top U = \mathbf{1}^\top U = 0$. Both remaining terms are nonnegative. Thus an invisible perturbation must have $\delta\alpha = 0$ and $h^\top I_U h = 0$. Since I_U is symmetric positive semidefinite, the latter condition is equivalent to $h \in \ker(I_U)$.

The functional $a^\top\beta$ is structurally estimable exactly when $a^\top h = 0$ for every invisible h, that is, exactly when $a \in (\ker I_U)^\perp$. For a symmetric matrix, $(\ker I_U)^\perp = \mathrm{range}(I_U)$, which proves the claim. □

F.8 Dimension of the band–period working basis

Proof of the nine-dimensional affine-span claim. The difference of any two 4×4 permutation matrices has zero row sums and zero column sums. The vector space

$$\mathcal{V} = \{H \in \mathbb{R}^{4\times 4} : H\mathbf{1} = 0,\ \mathbf{1}^\top H = 0\}$$

has dimension nine: its upper-left 3×3 block is free, the last entry in each of the first three rows and columns is then forced, and the bottom-right entry is forced consistently.

For $i, j \in \{1, 2, 3\}$, define $B_{ij} = E_{ij} - E_{i4} - E_{4j} + E_{44}$. These nine matrices form a basis of $\mathcal{V}$. To express B_{ij} as a difference of permutation matrices, choose any bijection between the two rows outside $\{i, 4\}$ and the two columns outside $\{j, 4\}$. Let one permutation use the assignments $i \mapsto j$ and $4 \mapsto 4$, and let the other use $i \mapsto 4$ and $4 \mapsto j$, with the chosen remaining assignments shared. Their matrix difference is exactly B_{ij}. Hence differences of permutation matrices span $\mathcal{V}$. The affine hull of the 24 permutation matrices therefore has dimension nine. □

F.9 Off-support non-identification

Construction for Remark 2. Let the bounded outcome range contain distinct values y_0 and y_1, and fix any response surface on the evaluated support $\mathcal{H}$. Extend it to two bounded response surfaces f_0 and f_1 that agree with the fixed surface on $\mathcal{H}$, agree with each other at every other point except $x^\dagger$, and satisfy $f_0(x^\dagger) = y_0$ and $f_1(x^\dagger) = y_1$. The two surfaces induce the same complete record on $\mathcal{H}$ but different responses at $x^\dagger$. The off-support response is therefore not identified without an additional restriction or target evidence.

If a constant-sign hypothesis held on all mapped cells, every resolved cell would have that same sign. Observing one resolved positive sign and one resolved negative sign therefore contradicts the hypothesis. A cell whose interval contains zero supplies neither sign and does not by itself establish or refute constancy.

For the two-coordinate illustration in Equation 21, the unobserved value Y_{11} may be any element of $[a, b]$. Subtracting the fixed quantity $Y_{10} + Y_{01} - Y_{00}$ maps that interval exactly to the interval displayed in the equation. Likewise, if the product in Equation 22 is negative, the two within-context utility differences have opposite signs. Hence neither of the two constant policy choices maximizes utility in both contexts within the stated two-policy set. □