A Window Is Not a Door: Mapping the Limits and Conditions of Knowledge Editing in Language Models

Anonymous authors
Paper under double-blind review

Abstract

Interpretability makes factual knowledge in language models observable, but observability does not imply editability. We identify a gap between locating factual signals and reliably changing knowledge, defined by control, transfer, and collateral damage. In the tested settings, localization-derived interventions move target answers without achieving selective control within the specified collateral budgets. Post-edit target hits likewise conceal already-correct answers, limited transfer, and damage to other knowledge. Reading out information and eliciting a target answer are therefore insufficient evidence of reliable knowledge editing. This gap is not uniform: the model's pre-edit support for the target answer helps organize where correction succeeds. Targets near the front of the base distribution, but not initially generated correctly, form a region of correction opportunities. Candidate rank predicts correction yield in independent fixed-budget tests, while far-rank successes reveal a model-dependent boundary rather than a hard cutoff. Context and weight interventions further expose distinct routes to changing answers, with different benefits and costs. Together, these findings shift the central question of knowledge editing from where knowledge resides to which facts can be changed, through which intervention, and at what cost. Interpretability provides a window into knowledge; a door to reliable editing depends on existing candidate support and the conditions of intervention.

1 Introduction

Language models answer a vast range of factual questions, yet they also confuse facts, recall them incorrectly, and retain outdated answers. When a model gives a wrong answer, we would like to correct it directly: the correction should survive a change in wording while leaving other knowledge intact. This would make model knowledge maintainable without retraining the entire model for each local revision. Knowledge editing pursues precisely such targeted, generalizable, and low-damage changes (Yao et al., 2023; Meng et al., 2022; 2023).

Interpretability offers an appealing starting point. Probes reveal information encoded in internal states (Belinkov, 2022); studies of factual recall trace how information contributes to an answer (Geva et al., 2023); and representation interventions can alter model outputs (Zou et al., 2023). These advances invite a natural question: **does locating a factual signal also locate an entry point for editing the fact?** Reading information and controlling it are distinct capabilities (Hewitt & Liang, 2019; Hase et al., 2023). A direction that reveals an answer need not support a stable correction while preserving other knowledge. Interpretability opens a window into knowledge; whether that window is also a door to editing remains an empirical question.

Existing knowledge editors change target answers and already evaluate efficacy, generalization, and locality (Meng et al., 2022; 2023). Subsequent work exposes failures to propagate the implications of an edit (Cohen et al., 2024), degradation of general abilities (Gu et al., 2024), and evaluation artifacts (Yang et al., 2025). These findings sharpen the question: **which edits remain effective when new correction, transfer, and collateral damage are considered together?** A target hit may preserve an already-correct answer, depend on the training prompt, or accompany damage to another relation.

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.

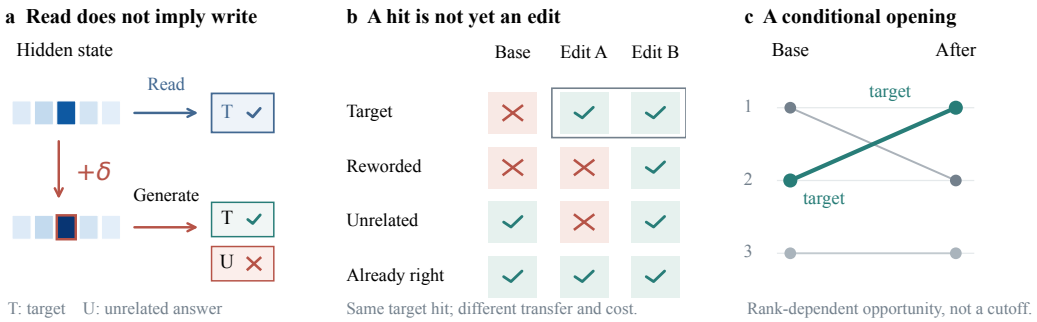

Figure 1: **Two barriers and a conditional opening.** Schematic examples, not measured activations or outcomes. (a) A target signal can be read out, yet intervention can change an unrelated answer as well. (b) Both edits hit the target; only Edit B also preserves the illustrated unrelated answer and succeeds on a reworded prompt. Unchanged correct answers are not new corrections. (c) A previously unselected candidate rises to rank one. Empirical tests and their boundaries follow in Sections 3–5.

We therefore distinguish target emission from new correction and specify whether transfer concerns new wording, unseen entities, or new relations.

Our central observation is that correction opportunities are associated with the base model's existing support for the target answer. An incorrect generation need not imply that the correct answer lacks support: it may already be competitive among the model's candidates without being generated correctly. Correction can then arise through a change in competition among existing candidates. Building on work relating edit outcomes to intrinsic features and knowledge familiarity (Soker et al., 2025; Pan et al., 2025), we test whether pre-edit candidate support identifies productive corrections under fixed interventions. This shifts the question from where a factual signal resides to which states of the target permit useful change.

Figure 1 summarizes the two barriers between observing knowledge and editing it reliably, and the conditional opening identified by our experiments. Across six models, none of 42 selected direct operator–model pairs combines positive selective control with the specified collateral budgets. A paired audit finds that 60–70 target hits per 100 facts contain only 7–22 new corrections, with further losses after preservation checks. Yet rank analyses across three models reveal productive correction opportunities among near-front candidates. Independent fixed-budget tests validate rank's predictive value, while far-rank successes establish a model-dependent opportunity distribution rather than a fixed writability threshold.

Context and training interventions probe the conditions of this opening. Explicit factual context raises target ranks and recovers answers without changing weights, following the in-context editing approach (Zheng et al., 2023). Increased local training exposure does not consistently overcome the joint control–cost barrier. Constrained low-rank updates (Hu et al., 2022) and an exploratory MoE masking trajectory reveal limited regions of low-damage editing. These complementary probes connect candidate support with the intervention used to change an answer.

Our contributions are threefold:

1. **An empirical gap between readability and selective control.** Direct interventions derived from localized signals distinguish moving target answers from preserving other knowledge while doing so.
2. **A characterization of apparent success versus reliable correction.** Paired output audits and transfer analyses separate target emission, new correction, and preservation, identifying what each success criterion establishes.
3. **Conditional routes to knowledge editing.** We characterize candidate support as a signal of correction opportunities, test its predictive value, and probe its conditions through context, local training, and architecture experiments.

A window is not a door: useful editing depends on which candidates can be corrected, through which intervention, and at what cost.

2 When Does an Edit Count?

A useful edit changes a target fact, carries that change into the intended test setting, and preserves knowledge outside its scope. We evaluate these requirements through three quantities: new correction, transfer, and collateral cost.

2.1 Changing an Answer, Correcting a Fact

Let $b_i \in \{0, 1\}$ indicate whether the base model's complete answer passes the fixed target scorer, and $e_i(a)$ the outcome after intervention a. On N facts, target-hit rate decomposes into retained correct answers and new corrections:

$$H(a) = \underbrace{\frac{1}{N}\sum_i b_i e_i(a)}_{\text{retained correctness}} + \underbrace{\frac{1}{N}\sum_i (1 - b_i) e_i(a)}_{C(a)\text{: new correction}}. \tag{1}$$

Net gain also subtracts regressions, where an initially correct answer becomes wrong:

$$G(a) = C(a) - \frac{1}{N}\sum_i b_i[1 - e_i(a)] = \frac{1}{N}\sum_i [e_i(a) - b_i]. \tag{2}$$

We report correction counts and the fraction of base-wrong facts corrected, with denominators explicit. Our correction and resurfacing tasks can contain base-correct targets; counterfactual replacement tasks specify alternative targets. Pre-existing target credit must therefore be measured within each task.

2.2 Where Should the Change Transfer?

Write a factual input as $x(u, r, t)$, with entity u, relation r, and wording t. We distinguish:

$$\begin{aligned} \text{Wording:} \quad & (u, r, t) \to (u, r, t'), \\ \text{Entity:} \quad & (u, r, t) \to (u', r, t'), \\ \text{Relation:} \quad & (u, r, t) \to (u', r', t'). \end{aligned} \tag{3}$$

Here primes denote held-out settings; in relation transfer, whether entity and wording also change is protocol-specific. Each test uses its own target answers. For each test population $\mathcal{D}_k$, transfer gain is measured separately:

$$G_k(a) = \frac{1}{|\mathcal{D}_k|}\sum_{i \in \mathcal{D}_k} [e_i(a) - b_i]. \tag{4}$$

Rewording a corrected fact tests persistence of that edit; correcting an unseen entity tests reuse of the intervention. These are different capabilities. Unmeasured transfer dimensions are left unreported, not assigned zero.

2.3 At What Collateral Cost?

We measure next-token distribution shift (KL), wrong-relation damage (WD), and unrelated-answer damage (UD). Prespecified thresholds define the admissible interventions:

$$\mathcal{A}_{\text{budget}} = \{a : \ D_{\text{KL}}(a) \le \tau_{\text{KL}}, \ \text{WD}(a) \le \tau_{\text{WD}}, \ \text{UD}(a) \le \tau_{\text{UD}}\}. \tag{5}$$

Correction studies ask whether $G_k(a) > 0$ within the specified budget. Direct-controller tests instead require positive selective control, with mean KL ≤ 1 and WD/UD ≤ 0.1. Later studies retain their absolute or base-relative gates. Budget passage denotes bounded aggregate damage, not zero change on every fact; partial checks remain identified as partial. Appendix Table A1 lists the experimental populations and comparisons. Each result retains its original protocol and selected checkpoints.

3 Readable, but Not Reliably Writable

From observing a signal to testing a control. Localization identifies internal states, sites, or directions associated with factual responses (Geva et al., 2021; Dai et al., 2022; Geva et al., 2023; Hernandez et al., 2024). The editing question is whether intervening on that structure changes the target selectively. We test seven operator families—readout, patching, path attribution, interaction groups, QKV routing, DAS, and CDAS—across Qwen3-8B, 14B, 32B, 30B-A3B, Gemma-31B, and Llama-3.3-70B. A separate geometric analysis constructs RepE read directions from correct/incorrect recall pools and compares them with fitted write directions. These experiments test a proposed control interface, rather than equating a readable signal with an editable fact.

Target movement and preservation do not coincide. None of the 42 selected direct operator–model pairs achieves both positive selective control and the specified collateral budgets: 26 stay within budget without positive control, eight achieve positive control outside budget, and eight achieve neither. Figure 2 makes the two failure modes visible. The result is also 0/21 among unflagged pairs; the remaining 21 pairs retain upstream measurement-NO-GO flags and are descriptive intervention outcomes, not valid localization evidence. Appendix Table A2 separates these branches, and Appendix Figure A1 reports target and damage rates by operator.

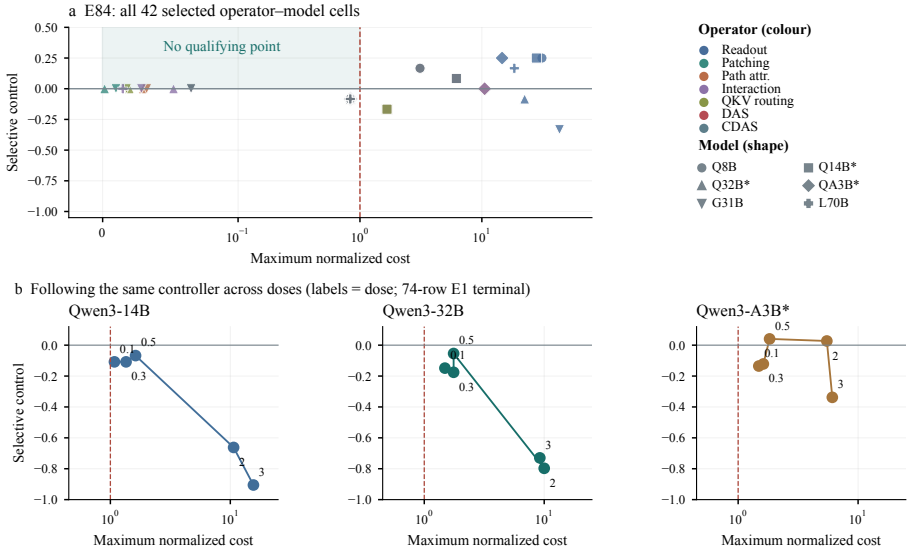

Figure 2: **Positive control and low collateral cost do not coincide in the selected direct interventions.** Top: 42 E84 cells, colored by operator and shaped by model. Bottom: five E1 doses per model. Cost is $\max(\mathrm{KL}, \mathrm{WD}/0.1, \mathrm{UD}/0.1)$. Cohorts differ; lines connect only within-model E1 doses. Three E84 Qwen branches are measurement-NO-GO (Appendix Table A2); they remain descriptive. *E1 A3B lacks a separate preregistration amendment.

Increasing strength does not remove the trade-off. The separate 74-row experiment follows five doses on each of three Qwen models. Its trajectories show how stronger intervention changes selective control and collateral cost together (Figure 2, bottom). Increasing dose can increase cost while control deteriorates; the small positive-control segment in the A3B trajectory remains outside budget. These within-model paths explain why target movement alone is insufficient and preserve the distinction between this population and the 42-pair comparison.

Reading and writing expose different geometries. On Qwen3-14B and 32B, the RepE direction has absolute cosine at most 0.03 with output-side fitted write directions or realized perturbations at the measured capture site. The 14B input-side cosine is approximately -0.13: alignment depends on which part of the intervention is compared. This is a geometric observation, not by itself a causal

explanation of editing failure. The matched-control experiment asks the stronger practical question: does using read-derived geometry improve a learned update?

Across five LoRA arms, three Qwen models, and three seeds, geometry is constructed on training data, checkpoints are selected on validation data, and outcomes are measured on 302 test entities. Matched random projections and an unmasked arm provide controls. The prespecified geometric advantage is unconfirmed in all six primary Qwen comparisons and both Gemma-E4B comparisons; their intervals still allow differences between arms. Together, the direct and learned tests show that a read-derived structure does not automatically supply a useful editing advantage. The next question is what remains when an update does produce the target answer.

4 The Answer Changes. Does the Edit Hold?

Most target hits are not new corrections. Rescoring the same outputs isolates the gap between target emission and correction. Qwen14B, 32B, and A3B emit the target on 60/100, 70/100, and 68/100 facts, but 38, 63, and 60 of those hits were already correct under the base model. New corrections are therefore 22/60, 7/31, and 8/35 among base-wrong facts. Adding the available per-row KL and unrelated-preservation checks leaves 6/60, 4/31, and 3/35 (Figure 3; Appendix Table A3). The final rung omits WD and is a partial preservation check. Holding outputs and the scorer fixed reveals how much apparent success comes from unchanged correct answers or corrections that exceed measured costs.

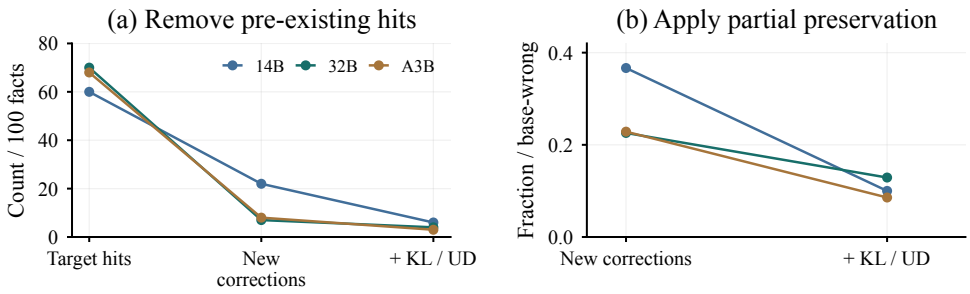

Figure 3: **Target hits exceed new, preserved corrections.** (a) Counts use the same 100 facts per model. (b) Rates condition on 60/31/35 base-wrong facts. The final check includes KL and unrelated preservation only; WD is omitted.

What transfers depends on what was learned. The synthetic capacity experiment separates training facts from held-out facts, rather than rewording the same facts. At its N800 final checkpoint, first-token matches reach 199/200 sampled training facts, but complete answers succeed on 74/200; complete answers on held-out facts succeed on 0/160 (Appendix I). Thus token-level fit, complete-answer acquisition, and unseen-fact performance are distinct outcomes. Because the synthetic targets are assigned to invented facts, this result measures that training channel; it does not establish that an edited fact fails under paraphrase. Conversely, the learned-update experiments in Section 5 do produce gains on unseen entities. Transfer must be judged on the stated population, not inferred from target hits or dismissed across all tasks.

A target hit can carry a substantial bill. In an exploratory target-directed LoRA overwrite, target control is $1/3$, mean KL is 3.669, wrong-relation damage is $11/12$, and unrelated damage is zero. The tested alternative relation is disrupted even though the unrelated arm is preserved. Our condensed ROME-style single-site implementation instead fails at target insertion, before post-insertion locality can be assessed. These configurations expose different failures; MEMIT, MEND, and SERAC remain related work, not official implementations evaluated here.

The complete answer matters. Headquarters cases show why semantic adjudication complements automatic scoring: “Astana, Kazakhstan” retains the correct place despite failing an exact “Astana”

match, whereas a matching place head can accompany a wrong country qualifier (Appendix Table A6). Scoring can therefore miss a correct answer or accept an incomplete indication of correctness.

The second wall is the gap between a target hit and a reliable correction. Paired scoring reveals incremental benefit, transfer tests specify its reach, and collateral checks expose its cost. The surviving correction opportunities motivate the next question: what distinguishes facts that can be corrected?

5 Where the Wall Gives Way

The preceding barriers do not eliminate correction opportunities. We now ask where those opportunities occur, whether they can be identified before editing, and which intervention conditions change them. Candidate rank connects the first two questions; context, training exposure, and masking probe the third.

5.1 The Answer Is Already a Candidate

Budget-admissible gains exist: learned updates yield positive net correction on unseen entities in 37/45 Qwen method–model–seed cells and 15/15 Gemma cells. A separate historical fact-level audit asks where correction occurs within the base answer distribution. Its unit is a fact, rather than a training run.

We measure support by the target answer's first-token rank before intervention, but judge correction by complete-answer generation. Thus even a rank-one first token can precede a wrong answer; candidate support does not establish complete factual knowledge. Rank 2–3 is productive in all three audited models: Qwen14B, 32B, and A3B correct 12/24, 3/12, and 7/14 base-wrong facts, respectively (Figure 4).

The boundary is graded, not absolute. Qwen14B also corrects 4/13 rank-11+ facts, compared with 0/6 and 0/4 for 32B and A3B. The full curves retain non-monotone rates, sample sizes, and uncertainty; per-fact KL exposes one component of cost. Near-front support identifies a recurring correction opportunity without defining a universal cutoff or a full-budget success rate.

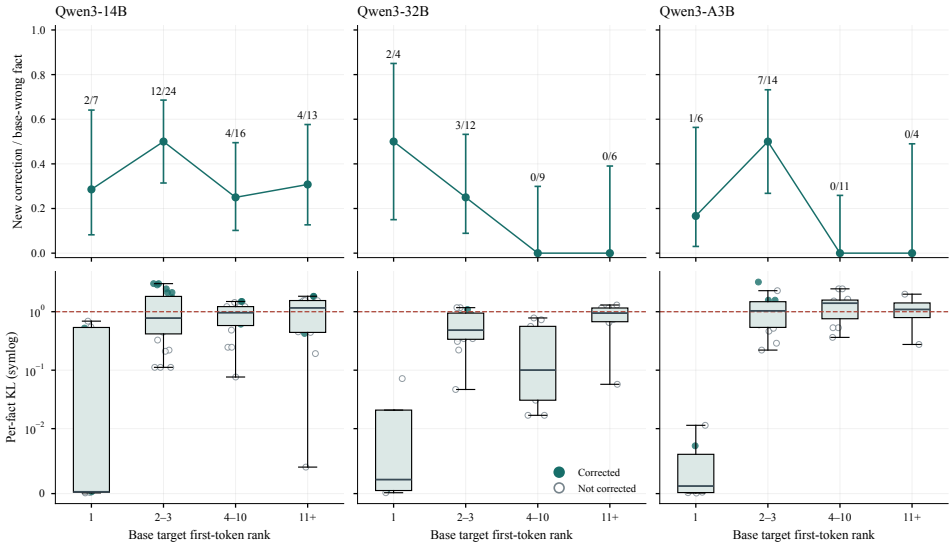

Figure 4: **Near-rank candidates yield corrections across all three models; far-rank successes persist in Qwen14B.** Top: rates and Wilson 95% intervals for 60/31/35 base-wrong facts. Bottom: matched per-fact KL, quartiles, 1.5-IQR whiskers and all observations; filled points indicate correction. Checkpoints are fixed. Qwen14B corrects 4/13 rank-11+ facts. KL is one collateral criterion.

5.2 Can We Tell Before We Try?

If candidate support identifies an opportunity, it should help select facts before their edited outcomes are known. A separate fixed-budget test evaluates this prediction, connecting our observation to prior work on intrinsic features and knowledge familiarity (Soker et al., 2025; Pan et al., 2025).

With three previously selected shared adapters held fixed, we select the 200 eligible entities with lowest base rank and compare them with rank-blind and answer-frequency-matched subsets. Ranks are measured before edited outputs are opened. The 10,000 subsets characterize selection variability within each fixed output pool, not independent training replications.

Rank selects more new corrections in all three relation–model comparisons (Table 1), including against frequency matching. Each policy satisfies its prespecified per-seed cost gates. Because adapter training and the 200-entity budget are held constant, this tests the predictive usefulness of candidate support rather than an advantage from additional training.

Relation	Model	Eligible	Blind	Matched	Rank
Work–country	Qwen14B	347	38.63	50.06	**63.67**
Work–country	Gemma-E4B	380	25.91	38.58	**45.33**
Work–language	Qwen14B	486	24.93	30.42	**52.67**

Table 1: **Base rank identifies more corrections under the same 200-entity budget.** Entries are mean new corrections over three fixed shared adapters. Blind and answer-frequency-matched policies use 10,000 subsets. Eligible counts are base-wrong entities. Each policy satisfies its study's cost gates; adapter training cost is unchanged.

5.3 Context Changes Candidate Support

Candidate support is not fixed across inputs. Explicit factual context raises the target first token into the top three on 229/243 initially misanswered Qwen14B facts, compared with 74/243 without context; complete target emission occurs on 196/243. Figure 5(a–c) pairs rank movement with the separately measured generation outcome.

This inference-time route supplies information through context rather than weight updates (Zheng et al., 2023). The paired observations show that candidate promotion accompanies answer recovery, while the difference between top-three support and complete emission shows that promotion alone does not guarantee recovery. They motivate, rather than establish, a causal account of context through rank.

5.4 Does More Local Training Open the Door?

Increasing exposure tests a different possibility: can more local training improve control without exceeding collateral budgets? Across 23 available endpoints at 10, 50, 200, and 1,000 paraphrases per fact, control and cost remain in tension (Figure 5, d–f). A target-control increase is useful only if its endpoint also passes the original damage gates.

These are within-procedure scaling tests on a separate 74-fact population: 14B/A3B use 300-step full-weight training, whereas 32B uses one-epoch LoRA. Appendix Figure A2 adds unrelated damage; larger-scale and constrained arms retain their own protocols. The sweeps do not show a reliable escape through increased exposure in the tested settings, nor constitute a matched comparison with context.

5.5 Selective Editing in an MoE Masking Probe

Changing where an update acts provides another possible opening. In the exploratory A3B masking trajectory, 17/20 checkpoints satisfy incremental-damage budgets relative to base; peak target control reaches 26/44 with KL approximately 0.13. These checkpoints trace one optimization path, not independent replications.

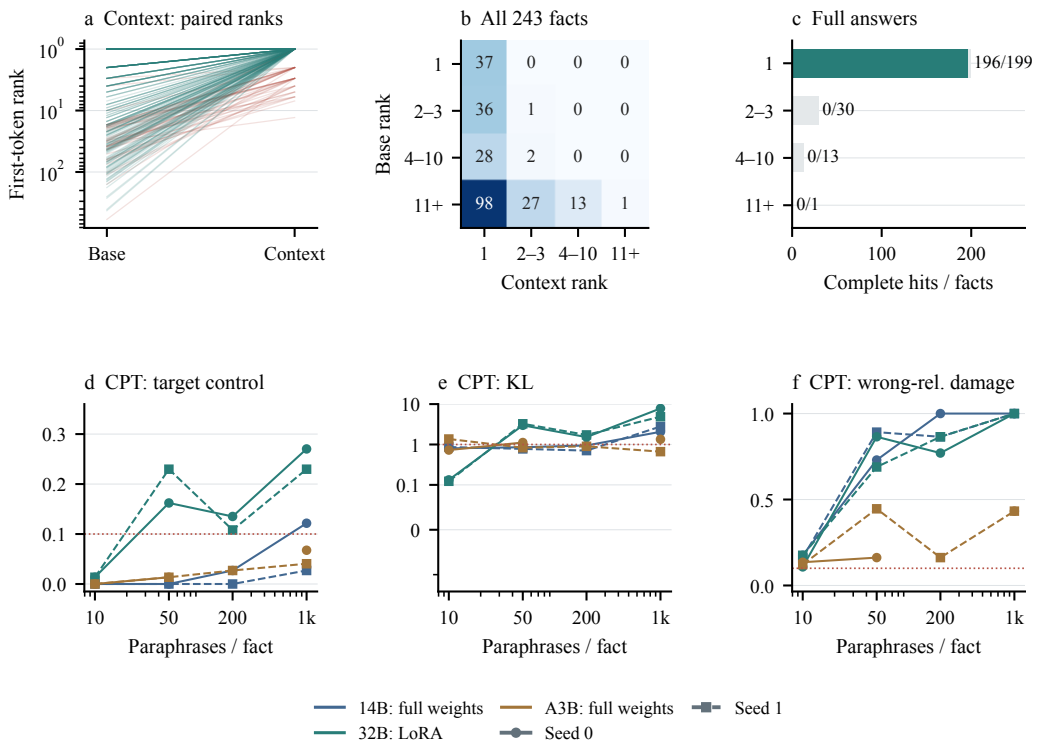

Figure 5: **Context shifts candidate support; increased local training retains substantial costs.** (a–c) 243 Qwen14B facts: context moves 229 targets into the first-token top three; 196 emit completely. Teal/red trajectories indicate complete emission/no emission in the separate generation assay. (d–f) All 23 available CPT endpoints on a separate 74-fact set; lines connect each model/seed across exposure levels. 14B/A3B use 300-step full-weight training; 32B uses one-epoch LoRA. Dotted lines are TC=0.1, KL=1 and raw WD=0.1 references, not a recomputed joint gate: 32B uses delta-damage gates. The missing A3B seed-0 endpoint at 200 is a gap. These are complementary probes, not a matched channel comparison.

The different pattern in dense-model probes motivates an architectural hypothesis, but size, factual support, intervention site, and optimization also differ. The result identifies a bounded low-damage region under a particular masked update, rather than establishing general MoE superiority.

Together, these experiments locate conditional openings: candidate support predicts productive attempts, while context, exposure, and update location probe how those opportunities can change. Useful editing depends on both the target's initial support and the intervention's cost.

6 Discussion and Conclusion

Observability and editability are different capabilities. A readable signal identifies information available in a representation; a useful edit must additionally control which answers change and which remain intact. Our findings connect this distinction to the target's initial state: candidate support predicts correction opportunities under fixed interventions, while transfer and collateral tests establish their value. The graded rank pattern points to conditions for useful change, rather than an impossibility of editing or a universal writability threshold.

Implications for editing, context, and decoding. For editing, pre-edit support offers a signal for allocating attempts, followed by checks of new correction, transfer, and damage. Context offers another way to alter candidate competition without updating weights. For decoding, a supported but ungenerated answer suggests an opportunity to improve selection. Yet identifying the correct candidate remains a separate challenge: measuring its rank when the target is known does not supply a decoder that can recognize it at inference time. The distinction echoes RL pass@k analyses

of selection versus output coverage (Yue et al., 2025); our factual-correction results motivate this connection without establishing a shared mechanism.

What would distinguish the explanations? Three comparisons are decisive: rank versus other familiarity predictors on held-out relations; matched context interventions that separate rank promotion from other effects; and dense–MoE updates matched for factual support, site, and optimization. These would test whether candidate support adds explanatory power, mediates recovery, and accounts for apparent architectural differences. Current experiments establish predictive utility and intervention-specific observations, leaving these causal questions open.

A window is not a door. Observing factual information, emitting a target, and reliably correcting knowledge are distinct achievements. Our experiments locate conditional openings by asking which candidates can be corrected, how far the change transfers, and what it costs to preserve other knowledge.

7 Related Work

From localization to editing. Probing distinguishes recoverable information from its use in prediction (Belinkov, 2022; Hewitt & Liang, 2019; Elazar & Goldberg, 2018); mediation and causal abstraction test internal variables through interventions (Vig et al., 2020; Geiger et al., 2021). Latent-knowledge and factual-recall studies identify candidate structures (Burns et al., 2023; Geva et al., 2021; Dai et al., 2022; Geva et al., 2023; Hernandez et al., 2024), while steering methods use internal directions to alter outputs (Turner et al., 2023; Subramani et al., 2022; Rimsky et al., 2024; Zou et al., 2023; Li et al., 2023). Hase et al. already examine whether localization informs editing (Hase et al., 2023). We extend this question to selective control under joint collateral budgets and test whether read-derived geometry improves learned updates over matched controls. The contribution concerns the conditions of control, not the first observation of a read–write gap.

Editing success and reliable correction. ROME and MEMIT modify factual associations (Meng et al., 2022; 2023); MEND learns edits, whereas SERAC uses explicit memory and scope (Mitchell et al., 2022a;b). LoRA parameterizes low-rank updates (Hu et al., 2022), with surveys and frameworks organizing these approaches (Yao et al., 2023; Wang et al., 2024). Our implemented baseline scope is specified in Section 4.

Evaluation already extends beyond target hits: MQuAKE and RippleEdits test factual consequences (Zhong et al., 2023; Cohen et al., 2024); other work studies specificity, general-ability damage, forgetting, overfitting, and hallucination correction (Hoelscher-Obermaier et al., 2023; Gu et al., 2024; Gupta et al., 2024; Zhang et al., 2025; Huang et al., 2025). Mirage revisits evaluation under less controlled conditions (Yang et al., 2025). Our complementary audit holds outputs fixed, separates retained correctness from new correction, and measures how preservation checks change the apparent gain.

Prior knowledge and correction opportunities. Intrinsic-feature predictors, the Knowledge Spectrum, and studies of perplexing knowledge link editing outcomes to pre-edit properties (Soker et al., 2025; Pan et al., 2025; Ge et al., 2024). We examine first-token candidate rank against complete-answer correction, then test its selection value with fixed adapters and equal entity budgets. Frequency-matched controls show that the observed advantage extends beyond answer frequency; they do not establish superiority over every familiarity measure. This links a pre-edit signal to actionable correction opportunities while retaining model-dependent exceptions.

Context and the boundaries of post-training. In-context editing and retrieval-versus-fine-tuning studies examine alternative knowledge channels (Zheng et al., 2023; Ovadia et al., 2023). Work on unfamiliar facts studies fine-tuning and hallucination (Gekhman et al., 2024); broader critiques question when editing is appropriate (Pinter & Elhadad, 2023). RL pass@k analysis separates improvements in selecting successful outputs from expansion of their coverage (Yue et al., 2025). Our paired context measurements and local-training sweeps bring candidate support and collateral cost into this discussion. They concern factual answers under different intervention protocols, rather than a matched channel ranking or a common mechanism already established across reasoning and editing.

Reproducibility Statement

The source map links experiment-specific protocols, selected checkpoints, raw outputs, and SHA-256 references. All 54 recovered E84 artifacts match their recorded hashes. Measurement validity flags and later protocol extensions are disclosed with the relevant experiments. The paired metric audit was recomputed from original per-row checkpoint files. Row-level reanalyses and results drawn from historical summaries are identified separately.

AI Use Statement

Generative AI tools assisted with literature discovery and review, research discussion and interpretation of results, code and figure preparation, and manuscript outlining, drafting, translation, and language editing. The authors are responsible for the final text, analyses, and claims.

References

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

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*, 2024.

Damai Dai, Li Dong, Yaru Hao, Zhifang Liu, Zhifang Sui, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of ACL*, 2022.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of EMNLP*, 2018.

Huaizhi Ge et al. How well can knowledge edit methods edit perplexing knowledge? *arXiv preprint arXiv:2406.17253*, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Zorik Gekhman, Gal Yona, Roee Aharoni, et al. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of EMNLP*, 2024.

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

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of EMNLP*, 2023.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

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

Evan Hernandez, Zexuan Li, Jacob Geiger, and Jacob Andreas. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP*, 2019.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, and Ali Payani. Can knowledge editing really correct hallucinations? In *International Conference on Learning Representations*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

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

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

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

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

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in LLMs. *arXiv preprint arXiv:2312.05934*, 2023.

Tsung-Hsuan Pan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Diagnosing model editing via knowledge spectrum. *arXiv preprint arXiv:2509.17482*, 2025.

Yuval Pinter and Michael Elhadad. Emptying the ocean with a spoon: Should we edit models? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Erik Rapoport, et al. Steering Llama 2 via contrastive activation addition. In *Proceedings of ACL*, 2024.

Yanay Soker, Martin Tutek, and Yonatan Belinkov. Predicting success of model editing via intrinsic features. In *Conference on Language Modeling*, 2025.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In *Findings of ACL*, 2022.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. In *Advances in Neural Information Processing Systems*, 2020.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Huang, Zekun Xu, Siyuan Wang, Shumin Deng, Jiacheng Li, and Huajun Chen. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*, 2024.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. The mirage of model editing: Revisiting evaluation in the wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Yang Yue et al. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, and Pengjie Ren. Uncovering overfitting in large language model editing. In *International Conference on Learning Representations*, 2025.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Supplementary Evidence Guide

This appendix gives model–operator outcomes, expanded-support dose ladders, threshold checks, implementation details, installation examples, capacity controls, complete CPT results, and scoring cases. Each experiment retains its own population and gate definitions. Measurement-NO-GO flags, the later E1 A3B extension, partial preservation checks, and unmatched training procedures delimit the corresponding comparisons. The final section presents an exploratory representation analysis without paired editing outcomes.

Table A1: **Experimental setup.** Each row reports its own evaluation unit and comparison design; sample sizes are not pooled across studies.

Experiment	Models	Evaluation set	Comparison
Direct control	Six models	42 operator–model pairs	7 operators/model
Dose sweep	Qwen (3)	74 facts/model	5 doses
Metric audit	Qwen (3)	100 facts/model	Paired scoring
Rank analysis	Qwen (3)	60/31/35 base-wrong facts	4 rank bins
Learned geometry	Qwen (3), Gemma	302 entities/model	5 arms $\times$ 3 seeds
Rank selection	Qwen14B, Gemma	200 entities/policy	Rank / blind / matched
Context	Qwen14B	243 facts	With / without context
Local training	Qwen (3)	74 facts/model	4 scales $\times$ 2 seeds

Qwen (3): Qwen3-14B, 32B, and 30B-A3B; Gemma: Gemma-E4B. Rank selection uses two relations with the model coverage in Table 1. Local training has 23/24 available endpoints. Measurement flags, checkpoint choices, and gate definitions are reported with each study.

B Exact Counts for Direct Control and Paired Scoring

Model	Audit status	Within budget	Positive SC	Both
Qwen3-8B	Unflagged	4/7	3/7	0/7
Qwen3-14B*	NO-GO*	0/7	3/7	0/7
Qwen3-32B*	NO-GO*	6/7	0/7	0/7
Qwen3-A3B*	NO-GO*	4/7	1/7	0/7
Gemma-31B	Unflagged	6/7	0/7	0/7
Llama-70B	Unflagged	6/7	1/7	0/7

Table A2: 42 archived selected direct operators. Within-budget=26/42, positive-SC=8/42, both=0/42. *Original measurement-NO-GO branch; retained as flagged descriptive results, not valid localization evidence.

Model	Target emitted	Base already right	New / base wrong	+ partial gate
Qwen3-14B	60/100	38/60	22/60	6/60
Qwen3-32B	70/100	63/70	7/31	4/31
Qwen3-A3B	68/100	60/68	8/35	3/35

Table A3: Exact counts from three D-3 metric-audit checkpoints. Denominators differ explicitly: 100 all facts, target hits, then base-wrong facts. Last column omits WD and is not full joint admission. These are internal comparisons, not estimates of published-method inflation.

C Operator-Level Outcomes

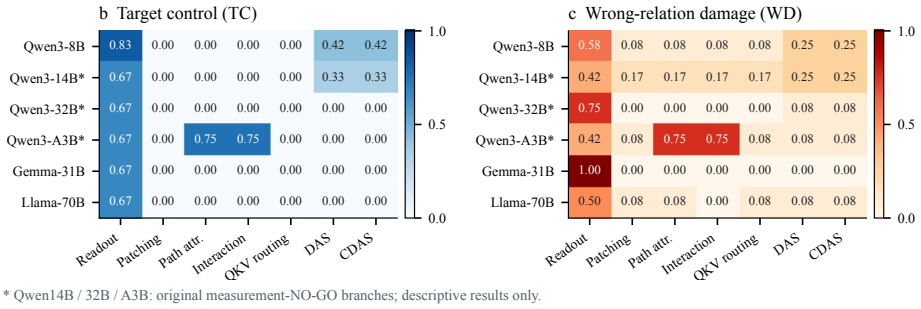

Figure A1: **Target control and wrong-relation damage across seven operators and six models.** Cells report the frozen behavioral rates. *Qwen14B/32B/A3B carry upstream measurement-NO-GO flags and remain descriptive results. Table A2 reports joint outcomes; the main-text control–cost figure also includes the separate E1 dose experiment.

D Extended-Support Ladders

Qwen3-14B, frozen protocol cell, n=74:

dose	SC (net/74)	KL	WD	UD	TC
0.1	$-8/74$	0.07	$8/74$	$0/74$	$0/74$
0.3	$-8/74$	0.16	$10/74$	$0/74$	$2/74$
0.5	$-5/74$	0.61	$12/74$	$0/74$	$7/74$
2.0	$-49/74$	10.65	$52/74$	$0/74$	$3/74$
3.0	$-67/74$	15.59	$67/74$	$0/74$	$0/74$

The frozen-support (n=12) dose-0.5 window that did *not* survive expansion:

Model (dose 0.5)	SC	KL	WD	UD	TC
Qwen3-14B	2/12 (0.17)	0.79	1/12	0/12	3/12
Qwen3-32B	2/12 (0.17)	0.72	1/12	0/12	3/12
Qwen3-30B-A3B	2/12 (0.17)	0.81	0/12	0/12	2/12
Llama-3.3-70B	2/12 (0.17)	0.79	0/12	0/12	2/12

Qwen3-32B tells the same story (dose 0.5: SC $-4/74$, CI $[-0.16, 0.05]$, KL 0.80), and Qwen3-30B-A3B completes the triangle with the only positive point estimate (dose 0.5: SC $+3/74$, CI $[-0.08, 0.16]$, KL 1.83 — out of the box on KL regardless of the interval). The full n=74 ladders:

Model	dose	SC (net/74)	KL	WD	UD	TC
Qwen3-32B	0.1	-11	0.02	11	0	0
	0.3	-13	0.12	13	0	0
	0.5	-4	0.80	13	0	9
	2.0	-48	10.27	50	0	2
	3.0	-59	13.58	62	0	0
Qwen3-30B-A3B	0.1	-10	0.01	11	0	1
	0.3	-9	0.31	12	0	3
	0.5	$+3$	1.83	11	0	14
	2.0	$+2$	5.49	15	0	17
	3.0	-25	6.10	43	0	18

A3B was a later extension without a separate preregistration amendment. The model-dependent trajectories do not isolate an architectural cause.

E Failure-Mode Decomposition

We decompose the method-axis failures of the two boundary-negative models — both released with full grid access (Qwen3-8B and Gemma-31B: 7 cell families $\times$ 5 layers $\times$ 2 λ settings $= 70$ points each) — into the constraint each point violates. (Families here are E85 grid initialization labels — optimized cells — not the E84 operator points of the method axis.) On Qwen3-8B every family fails by *both* modes: between 50% and 70% of points fail to move the answer (SC < 0.1), while 90–100% of points violate the distribution constraint (KL > 1.0) and 60–100% violate a collateral constraint (WD or UD > 0.1). Zero points enter the box in any family.

Qwen3-8B cell family	points	SC< 0.1	KL> 1	damage> 0.1	enters
r16 CDAS	10	6	10	10	0
r16 DAS	10	5	10	9	0
r16 J-lens (src)	10	6	10	9	0
r16 orthogonal (random)	10	7	9	9	0
r2 J-lens (src)	10	6	9	7	0
r2 orthogonal (random)	10	6	10	6	0
r64 orthogonal (random)	10	5	10	7	0

Gemma-31B shows the same zero-admissibility with a shifted mode mix — near-total failure to move the answer (SC< 0.1 at 80–100% of points) combined with universal distribution violation, and somewhat less collateral damage:

Gemma-31B cell family	points	SC< 0.1	KL> 1	damage> 0.1	enters
r16 CDAS	10	10	10	6	0
r16 DAS	10	10	10	6	0
r16 J-lens (src)	10	10	10	7	0
r16 orthogonal (random)	10	10	10	7	0
r2 J-lens (src)	10	8	10	6	0
r2 orthogonal (random)	10	10	9	5	0
r64 orthogonal (random)	10	10	10	5	0

The tabulated failures apply to this optimized-cell grid, separately from the selected direct-operator E84 matrix.

F Threshold Sensitivity

The admissibility box's thresholds are preregistered but arbitrary; a reviewer should ask whether any verdict flips under reasonable alternative thresholds. We re-evaluate every grid point of every model under (i) independent perturbation of each single constraint over $\pm 50\%$ and (ii) simultaneous scaling of all five thresholds by a global factor $s \in [0.5, 1.5]$. A *verdict* is model-level: "admits an in-box controller" iff at least one grid point satisfies all five constraints. Caveat: the four IN verdicts below rest on the $n{=}12$ ladders — precisely the evidence class that the preregistered $n{=}74$ expansion (Appendix D) dissolved. We keep the table to document where the apparent windows sat, not as evidence of admissibility.

Model	evidence	baseline	τ_{sc}	τ_{kl}	τ_{d}	τ_{tc}	stability r^*
Qwen3-8B	full grid	OUT	–	–	–	–	$[-0.50, 0.50]$
Gemma-31B	full grid	OUT	–	–	–	–	$[-0.50, 0.50]$
Qwen3-14B	mini ladder	IN	–	F(-50%)	F(-50%)	–	$[-0.15, 0.50]$
Qwen3-32B	mini ladder	IN	–	F(-50%)	F(-50%)	–	$[-0.15, 0.50]$
Qwen3-30B-A3B	mini ladder	IN	–	F(-50%)	–	–	$[-0.15, 0.50]$
Llama-3.3-70B	mini ladder	IN	–	F(-50%)	–	–	$[-0.20, 0.50]$

Each threshold column reports whether the model-level verdict flips when that constraint alone moves by $\pm 50\%$ (F(-50%) = flips under tightening; F($+50\%$) = flips under relaxation; – = ro-

Prompt	Base (wrong)	Installed (correct)	KL	Collateral
Where was Marie Curie born?	Poland	**Warsaw**	0.58	none†
When was Kia founded?	1973	**1944**	0.004	none†
In which year did Hitachi start?	1905	**1910**	0.001	none†
In which language was G-Taste originally made?	Korean	**Japanese**	0.90	none†
A catalog lists the work Tarrath Usk. What is its country of origin?	United States	**Finland**‡	0.88	none†

Table A4: Training-time installation of facts the base model gets wrong or imprecise (32B, box-constrained LoRA). KL = next-token KL(base‖installed). †Wrong and unrelated arms preserved. ‡Invented entity (synthetic arm). All rows from in-box checkpoints.

bust). r^* is the interval of simultaneous proportional perturbation of all five thresholds over which the verdict is unchanged ($r = +0.5$: every constraint relaxed by 50%; $r = -0.5$: every constraint tightened by 50%). Negative verdicts (full 70-point grids) are invariant across the entire sweep. Positive verdicts rest on the five-dose ladders: they are robust to arbitrary relaxation and to $\sim$15% simultaneous tightening, and the binding constraint under tightening is always τ_{kl} — consistent with the threshold-marginal window reported in the historical mini-frontier analysis.

G Baseline Implementation and Optimization Details

The condensed ROME-style implementation uses Qwen3-14B, covariance estimated from 200 pool prompts, a subject-last-token key, and a closed-form rank-one update. Three tested instantiations span learning rates 0.5–1.0 and 100–300 optimization steps with three KL weights. Target NLL remains above approximately 2.2 nats and insertion was not established. This tests the recorded single-site implementation, not an official full-protocol reproduction.

The target-directed LoRA endpoint described in Section 4 has KL 3.669, target control 1/3, wrong-relation damage 11/12 and unrelated damage 0. Earlier penalty runs contained an inference-mode gradient bug; the corrected sweep uses a live KL penalty. The historical corrected sweep records KL 0.94 at penalty 10 with target control 0; its first positive-control epoch has KL 1.15.

H Installation Examples and Architecture Probe

The examples describe the archived 32B installation checkpoints. The A3B masking sweep separately records 17/20 delta-budget checkpoints and peak control 26/44 at KL approximately0.13. Correlated checkpoints and unmatched dense/MoE settings do not yield an independent architectural effect estimate.

I Synthetic Content and Capacity Controls

A historical synthetic-relation control assigns two-digit answers to real entities ($n = 48$). Across doses 0.1–2.0, outputs become digits-only on 40–47/48 rows, while assigned-prefix matches remain 0–2/48. At dose 0.5, SC is$-4/48$ and KL 3.78. Surface-format steering and correct assigned content are separate outcomes.

The D-5 capacity experiment trains on synthetic facts at sizes 50/200/800/3200. Its 80/20 split separates facts, rather than paraphrases of the same fact. The historical trained-row first-token accuracy is 96–100%; at N800, complete generation succeeds on 74/200 sampled trained facts and 0/160 held-out facts. These results separate first-token scoring, full generation, and unseen-fact transfer. They do not measure same-fact paraphrase robustness.

J Local Training: Plain and Constrained Exposure

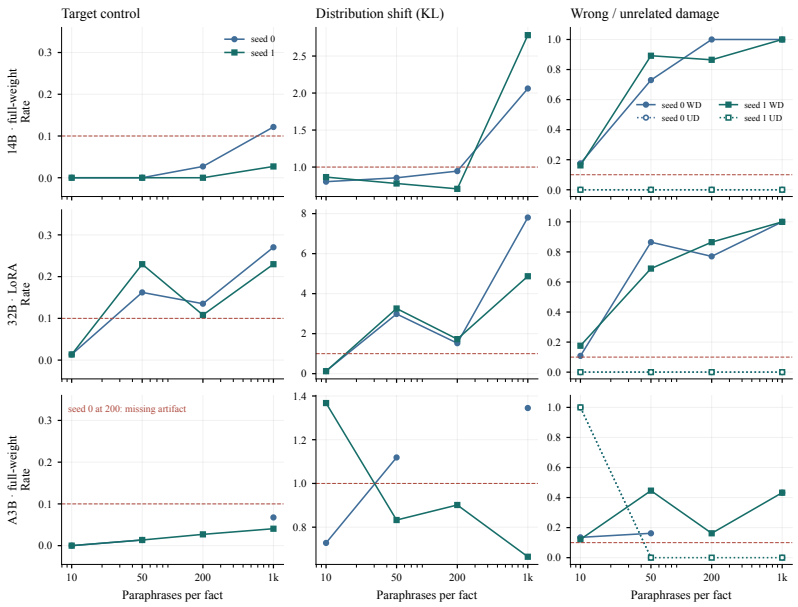

Figure A2: **Local-data scaling leaves control and costs in tension.** The 23 endpoints span two seeds and a 74-row terminal. Qwen14B/A3B use full-weight 300-step training; 32B uses one-epoch LoRA. A3B n200 seed0 is missing. Dashed lines show TC/KL/raw-damage thresholds; 32B gates delta-damage. These protocols do not isolate architecture.

strength	λ	TC (≥ 0.1)	KL (≤ 1)	WD	UD	SC	in-box
10	0.3	0.068	2.71	0.18	0	-0.11	no
10	1.0	0.081	4.80	0.27	0	-0.19	no
50	0.3	0.068	4.17	0.22	0	-0.15	no
50	1.0	0.000	2.69	0.19	0	-0.19	no
200	0.3	0.027–0.054	1.79–5.54	0.20	0–1.0	≤ -0.15	no
200	1.0	0.000–0.068	1.17–6.47	0.27–0.99	0	≤ -0.20	no
1000	0.3	0.000–0.041	9.75–14.33	0.31–1.0	0–1.0	≤ -0.27	no
1000	1.0	0.027–0.081	3.40–3.60	0.32–0.39	0–1.0	≤ -0.30	no
10000	1.0	0.068	4.34	0.23	0	-0.16	no

Table A5: Historical box-constrained CPT terminal summary for Qwen14B, full-weight 300-step runs. Ranges represent the archived seed summaries. All listed terminal endpoints fail joint admission. The table reports the constrained arm separately from the plain-exposure plot; finite failures do not establish a structural impossibility.

K Qualitative Scoring Cases

Entity	Base output	Edited output	Full / head	Archived adjudication
XDS Astana Development Team	Astana	Astana, Kazakhstan	Fail / pass	Qualifier consistent; base already right
S.P. Cailungo	Rome	Cailungo, Italy	Fail / pass	Wrong country qualifier
F.C. Penafiel	Penafiel	Penafiel, Brazil	Fail / pass	Wrong country qualifier

Table A6: Three archived post-result HQ examples; full-string fail versus lexical-head pass. Cases do not modify frozen scores or estimate overall semantic net gain.

L Exploratory Representation View

Figure A3 preserves the real three-dimensional representation view. It describes paired historical context records, not before/after weight editing. The archived outcome identifiers do not overlap these activation identifiers; no correction or safety labels can therefore be assigned. This visualization does not establish an editing mechanism.

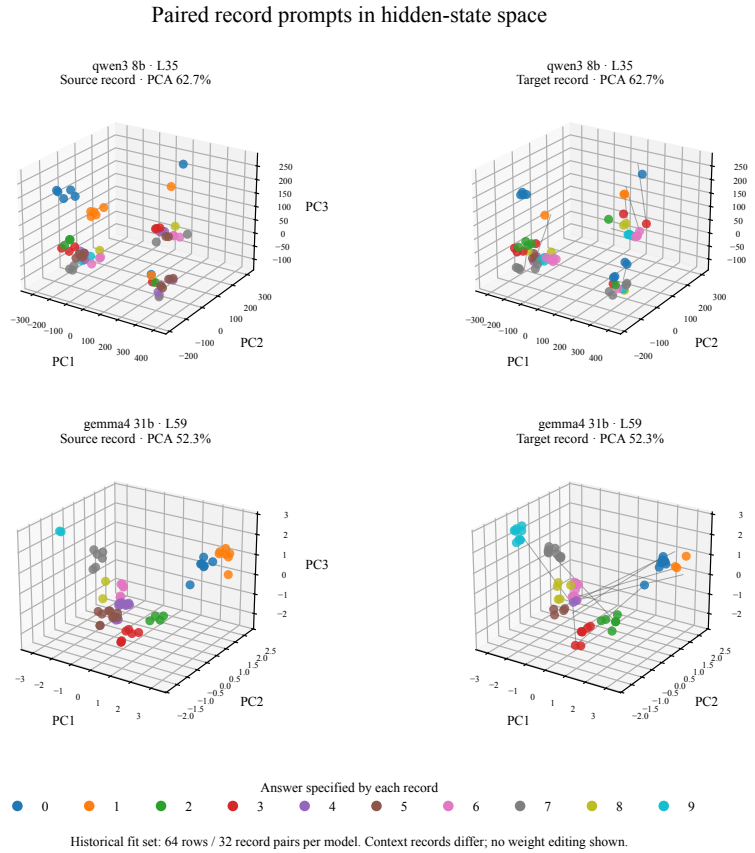

Figure A3: Actual last-prompt-token decoder states: Qwen3-8B L35 and Gemma4-31B L59, 64 historical fit queries / 32 base entities each. Source and target change record identifiers and values. Colours denote specified answer digits, not correctness. Joint source/target PCA per model retains 62.7%/52.3%; coordinates are not aligned across models. Eight paired lines selected by ID hash. This is neither weight-edit before/after nor evidence of low-damage correction. Same-experiment outcome IDs overlap 0/64. Four-condition effect-labelled figure remains pending.