How Test-Time Updates Become Behavior: Objectives, Training Scope, and Update Paths

Anonymous authors
Paper under double-blind review

Abstract

Can large language models turn information in deployment-time support into behavior that remains accessible after the support is removed? We study a Support-Derived Behavioral Objective (SDBO) that constructs one question–answer target from the support and a fixed schema, then takes four local stochastic-gradient steps. On 32 independent QAEdit confirmation episodes, SDBO improves joint attained success for Llama-3.1-8B-Instruct over raw-support training by 37.5 percentage points at both primary KL budgets. Paired analyses establish an objective-by-training-scope interaction; content controls and held-out formulations show that the effect depends on the correct support content and extends beyond the training wording. Two fixed update policies achieve positive native-query net gains and pass relative-preservation non-inferiority tests. To probe why scope matters, we decompose full-model updates and intervene on 32 separate fresh episodes. Scaling each full-update projection to the effective radius of its independently trained local counterpart leaves a 62.5-percentage-point behavioral gap at both policy points. Path comparisons further support a learning-rate-driven multistep contribution. The complete gain-and-preservation pattern is not reproduced on Qwen3-8B-Base, marking a model-setting boundary. These results show that final parameter support and update magnitude do not determine behavior: the process that forms the update also matters.

1 Introduction

Large language models increasingly encounter information only after pretraining: a corrected fact, a task rule, or context supplied by a user. Test-time training can absorb such information into parameters, but fitting the support does not ensure that the change becomes callable behavior after the support is removed. In an early facts-and-rules diagnostic, 5,700 of 5,760 records improve their proxy metric while none improves free-generation behavior. This gap motivates evaluating the behavior produced by an update rather than its training loss alone.

The distinction matters because the same information can define several update processes. A model can predict the support tokens, optimize a target derived from their meaning, or distribute the update over different parameter sets. Each process may reduce its own loss while forming a different parameter direction. We therefore treat support-removed behavior as the primary endpoint and vary both the training objective and the parameter scope that forms the update.

We study a simple rule built around a *Support-Derived Behavioral Objective* (SDBO; O_B). A deterministic transform maps the support and a fixed schema into one question–answer target. The model takes four stochastic-gradient steps on a local set of MLP projection weights, removes the support, and answers behavioral and preservation queries (Figure 1). Evaluation queries and outcomes never enter target construction or optimization.

On 32 independent QAEdit confirmation episodes, local Llama training with SDBO improves joint attained success over raw-support prediction by 37.5 percentage points at both primary KL budgets. The effect depends on training scope, correct support content, and model setting, while rephrased and independently constructed queries show access beyond the canonical training wording. Two fixed update policies retain positive query gains and pass relative-preservation non-inferiority (NI). Mechanism experiments then show that neither projecting a full-model update onto the same

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.

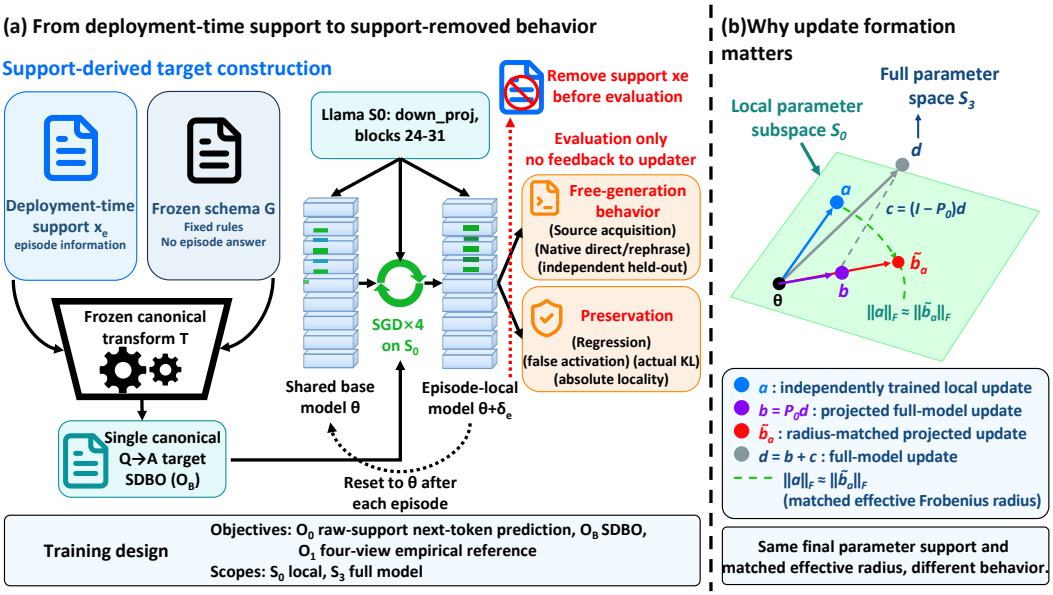


Figure 1: From deployment-time support to support-removed behavior. **(a)** A support-derived target drives a four-step local update before behavioral and preservation evaluation. **(b)** Independently trained, projected, and radius-matched updates motivate tests of whether final parameter support and effective radius determine behavior.

coordinates nor matching its effective Frobenius radius reproduces the independently trained local behavior. Together, these results connect an executable update rule to evidence that update formation has behavioral consequences.

Our contributions are:

- an independently confirmed, content-specific SDBO effect whose realization depends on the objective–scope recipe;
- fixed policies that achieve positive query gains and pass relative-preservation NI without outcome-based state selection; and
- update decomposition, direction–radius intervention, and path comparisons showing that training-time local restriction is not recovered by post-hoc projection and rescaling.

2 Related Work

Test-time training and fast adaptation. Test-time training derives a learning signal from the current test input and updates a model before prediction (Sun et al., 2020). Language-model variants fine-tune on retrieved neighbors (Hardt & Sun, 2024), adapt to target-domain text (Hu et al., 2025), use long context as training data (Bansal et al., 2026; Tandon et al., 2025), or expose projection matrices as online fast weights (Feng et al., 2026). We instead evaluate free-generation behavior after a sparse support input has left the context.

Per-context supervision and input parameterization. Programmatically generated question–answer pairs can train a model on one reading context before unseen questions (Banerjee et al., 2021); LIFT similarly compares synthetic-QA and raw-text fine-tuning after the original long input is removed (Mao et al., 2026). Building on this idea, we test how training scope and update formation affect support-removed behavior under a common information interface.

Behavioral model editing and preservation. ROME and MEMIT modify parameters to change factual behavior (Meng et al., 2022; 2023), while localization need not identify the most effective editing site (Hase et al., 2023) and AlphaEdit explicitly constrains preservation (Fang et al., 2025).

Selective parameter tuning can also outperform full fine-tuning under distribution shift (Lee et al., 2023); our comparison asks whether training-time restriction can be recovered after full-model optimization. Teacher-forced evaluation can overestimate model-editing performance (Yang et al., 2025). For test-time training, Song et al. (2026) distinguish proxy improvements from behavioral evidence for deployment-time memory after support removal.

3 Problem Setting and Behavioral Evaluation

3.1 Support-only test-time learning

An episode e contains support x_e, an acquisition probe P_e, native queries Q_e, independently constructed held-out queries H_e, and preservation items R_e. The updater may read only x_e and fixed global metadata $\mathcal{G}$:

$$\delta_e = \mathcal{U}(\theta; x_e, \mathcal{G}). \tag{1}$$

The probe, evaluation queries, their targets, and all evaluation outcomes are read only after the update. The support is then removed, the updated model is evaluated, and the shared base checkpoint θ is restored before the next episode. Episodes are the paired sampling units.

We compare raw-support next-token prediction (O_0), SDBO's single support-derived target (O_B), and an empirical reference (O_1) trained on four fixed support-derived acquisition formulations: reference-question, cloze, retrieval, and structured-record views. The main parameter scopes are a local scope S_0 and the full model S_3, with $S_0 \subset S_3$.

3.2 Behavioral endpoints

For query q with binary free-generation scorer s_e^q, define signed acquisition gain

$$G_e^q(\delta) = s_e^q(\theta + \delta) - s_e^q(\theta) \in \{-1, 0, 1\}. \tag{2}$$

Let $\bar{G}_e^Q$ average the native direct and rephrased gains, and let G_e^{src} be the gain on the acquisition probe. Behavioral success is

$$B_e(\delta) = \mathbf{1}\left[G_e^{\mathrm{src}}(\delta) \geq 1 \ \wedge \ \bar{G}_e^Q(\delta) \geq 0.25\right]. \tag{3}$$

We report direct, rephrased, and independent held-out endpoints separately.

Let $C_{\mathrm{reg},e}$ and $C_{\mathrm{FA},e}$ denote regression and false-activation rates. Absolute locality $L_e(\delta) = 1$ requires both rates to be at most 0.05 with valid denominators, and joint success is $J_e(\delta) = B_e(\delta)L_e(\delta)$. Actual KL $D_e(\delta)$ is the token-averaged full-vocabulary $D_{\mathrm{KL}}(p_\theta \| p_{\theta+\delta})$ over replay-prompt tokens.

3.3 Attained and fixed-policy estimands

For $X \in \{B, J\}$, attained success asks whether any of six prespecified configurations succeeds within actual-KL budget ϵ:

$$W_{\epsilon,e}^X = \mathbf{1}[\exists p \in \mathcal{P} : X_e(\delta_{e,p}) = 1, \ D_e(\delta_{e,p}) \leq \epsilon]. \tag{4}$$

Here $\mathcal{P} = \{.01, .04, .075, .15, 1.0, 3.0\}$. Each p indexes a separately calibrated four-step run from the shared base checkpoint, rather than an intermediate iterate of a single run. This estimand describes what the configuration grid reaches; it does not select a deployable state for a new episode. Fixed-policy analysis instead applies one preselected configuration π to every episode and compares objectives at that common operating rule. Its paired query net gain is

$$\Delta_Q^\pi = \frac{1}{N} \sum_e \left[\bar{G}_e^Q(\delta_{e,O_B}^\pi) - \bar{G}_e^Q(\delta_{e,O_0}^\pi)\right]. \tag{5}$$

The same updated state supplies the behavioral, preservation, and actual-KL outcomes. A policy passes preservation NI when the one-sided 95% upper bounds for both paired regression and false-activation differences are at most 0.05.

4 SUPPORT-DERIVED OBJECTIVE AND LOCAL TEST-TIME UPDATE

4.1 SUPPORT-DERIVED TARGET

A deterministic transform maps support and schema to one canonical query–target pair,

$$(v_e, y_e) = \mathcal{T}(x_e; \mathcal{G}). \tag{6}$$

For tokenized sequence $z_e = [v_e; y_e]$, a mask M_e selects the target and terminal EOS tokens. SDBO minimizes

$$\mathcal{L}_{\mathrm{SDBO}}(\theta; x_e) = -\frac{1}{|M_e|} \sum_{t \in M_e} \log p_\theta(z_{e,t} \mid z_{e,<t}). \tag{7}$$

Question and formatting tokens condition the prediction but do not contribute to the loss. By contrast, O_0 applies ordinary next-token prediction to the serialized support. Appendix C gives the template and complete objective definitions.

4.2 FOUR-STEP LOCAL OPTIMIZATION AND FIXED POLICIES

For Llama-3.1-8B-Instruct, S_0 contains the complete `down_proj` matrices in blocks 24–31 (zero indexed), totaling 469,762,048 parameters; S_3 contains all parameters. Let m denote the resident FP32 master weights for the trainable parameters, with all remaining parameters fixed. Starting from the shared base, the method performs four SGD steps on S_0,

$$m^{(j+1)} = m^{(j)} - \eta \nabla_m \mathcal{L}_{\mathrm{SDBO}}(m^{(j)}), \qquad j = 0, 1, 2, 3. \tag{8}$$

Forward and backward computation use BF16. Each objective–scope cell receives a learning rate calibrated separately on three pilot episodes using only actual KL. Calibration forms a monotone lower-learning-rate envelope of median actual KL and interpolates in log-KL/log-learning-rate space; confirmation behavior is never read.

Attained analyses use these six labeled configurations. Operational analyses use two fixed configurations, `targetkl_0.04` and `targetkl_0.075`, each mapping to an objective- and scope-specific learning rate. The label selects a fixed recipe; no per-episode KL adjustment or outcome-based state selection is performed.

5 EXPERIMENTAL SETUP

We evaluate Llama-3.1-8B-Instruct and Qwen3-8B-Base on natural QAEdit episodes stratified across NQ, SQA, and TQA. An initial 32-episode discovery set identifies the comparison of interest; primary method effects use a disjoint 32-episode confirmation set. Fixed-policy and decomposition analyses reuse confirmation32, whereas direction–radius and path tests use a separate fresh32 set. Secondary repetition and no-support controls use preallocated $n = 16$ subsets, and numerical traces use eight development episodes.

In the path analysis, Local-slow opens only S_0 but uses the learning rate calibrated for full-model training, enabling a same-rate comparison with Full-slow.

We compare correct SDBO with wrong-target, support-repetition, and no-support controls, and evaluate native direct, native rephrase, independent held-out, and preservation queries. All formal comparisons are episode paired. We use 10,000 bootstrap draws for two-sided 95% intervals and positive one-sided tests, with Holm correction within each stated family. NI uses one-sided 95% upper bounds. Source-composition intervals separately assess sensitivity over the NQ/SQA/TQA strata. Appendix D gives the complete procedures and numerical qualification rules.

6 BEHAVIORAL EFFECTIVENESS AND PRESERVATION

6.1 OBJECTIVE AND TRAINING SCOPE JOINTLY DETERMINE BEHAVIOR

On independent confirmation32, Llama/S_0 SDBO raises joint attained success from 6.25% to 43.75% at budget .05 and from 12.5% to 50.0% at budget .10 (Figure 2).

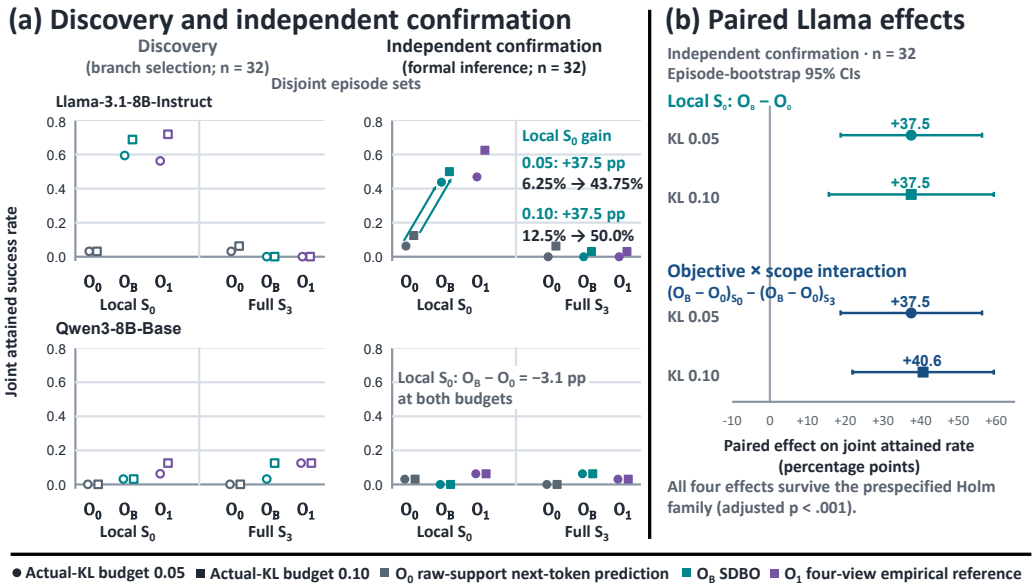


Figure 2: Objective and training scope jointly determine behavioral realization. **(a)** Joint attained rates on disjoint discovery and confirmation sets across objectives, scopes, models, and actual-KL budgets. **(b)** Episode-paired confirmation effects for the local SDBO gain and objective-by-scope interaction; bars are two-sided episode-bootstrap 95% confidence intervals.

Table 1: Primary paired confirmation estimands. Effects and episode-bootstrap 95% confidence intervals are in percentage points. Reported p values are Holm adjusted within two separate families: four method/scope effects and two fixed-policy gains.

Analysis	Budget/policy	Effect [95% CI]	Holm p	Paired outcome
Local $O_B - O_0$	.05	+37.5 [18.8,56.3]	$< .001$	joint success: 6.25% $\rightarrow$ 43.75%
Local $O_B - O_0$	.10	+37.5 [15.6,59.4]	$< .001$	joint success: 12.5% $\rightarrow$ 50.0%
Objective $\times$ scope	.05	+37.5 [18.8,56.3]	$< .001$	positive local-scope interaction
Objective $\times$ scope	.10	+40.6 [21.9,59.4]	$< .001$	positive local-scope interaction
Fixed-policy query gain	.04	+62.5 [46.9,78.1]	$< .001$	NI pass; 10/32 episodes over .05
Fixed-policy query gain	.075	+71.9 [54.7,85.9]	$< .001$	NI pass; 6/32 episodes over .10

Both paired gains are 37.5 percentage points. The objective-by-scope interactions are $+37.5$ and $+40.6$ points. These paired tests assess whether the SDBO gain changes with scope, and both survive the prespecified Holm family (Table 1); source-composition intervals remain positive.

The complete pattern is not reproduced on Qwen3-8B-Base: its local joint $O_B - O_0$ effect is -3.1 percentage points at both budgets, and all four Qwen scope–budget cells fail preservation NI. This establishes a model-configuration boundary. Appendix Table A6 reports the complete objective-level rates and multiplicity-adjusted results.

Table 2: Content specificity and cross-formulation transfer at actual-KL budgets .05/.10. All entries are attained-rate differences in proportions. For Panel A, identical budget-specific Holm-adjusted p values are reported once. Panels A–B use confirmation32; Panel C uses preallocated $n = 16$ subsets and reports control minus correct SDBO.

Panel	Model/control	Budget .05	Budget .10	Holm-adjusted p
A: correct–wrong joint	Llama	+.43750	+.50000	$< .001$ (both)
	Qwen	0	0	1 (both)
B: $O_B - O_0$ direct	Llama	+.53125	+.59375	–
B: rephrase	Llama	+.53125	+.59375	–
B: independent held-out	Llama	+.50000	+.56250	–
B: direct	Qwen	+.06250	+.09375	–
B: rephrase	Qwen	0	+.06250	–
B: independent held-out	Qwen	+.06250	+.09375	–
C: repetition–SDBO	Llama	-.25000	-.25000	–
C: no-support–SDBO	Llama	-.31250	-.31250	–

6.2 The gain is content-specific and transfers across formulations

Correct SDBO exceeds the wrong-target control by 43.75 and 50.0 percentage points in Llama joint attainment (Table 2). This contrast tests target-content specificity while holding the single-target interface fixed.

The SDBO advantage over raw-support training is positive on native direct, native rephrased, and independently constructed held-out queries: rephrasing tests access beyond the canonical wording, while independently constructed evaluation queries test transfer beyond the target-generation template. Repetition and no-support controls are also below correct SDBO on their preallocated subsets, separating target construction from token repetition and update-only effects.

6.3 Fixed policies achieve gain with relative preservation

We next apply the two prespecified policies uniformly to confirmation32, without outcome-dependent state selection; query gain, preservation, and the KL tail therefore describe the same deployed rule.

The two fixed policies yield native-query $O_B - O_0$ net gains of $+62.5$ and $+71.9$ percentage points, both significant after correcting across policies. Both pass regression and false-activation NI relative to O_0. This population comparison coexists with absolute-locality pass rates of $68.8\%/65.6\%$ and actual-KL exceedances in 10/32 and 6/32 SDBO episodes (Figure 3). The former tests average preservation relative to raw-support training; the latter two describe absolute per-episode locality and perturbation magnitude. Complete objective-level outcomes and NI bounds appear in Appendix Table A7.

7 Why Training Scope Matters

The confirmed SDBO effect is strong under S_0 and weak under S_3. Because $S_0 \subset S_3$, parameter-set inclusion alone cannot explain this interaction: the full model can represent every local-coordinate update, yet its optimization path forms a different update. We next test whether the behavior can be recovered from the local component of a completed full-model update, then separate effective radius, direction, and multistep path.

7.1 Training-time restriction is not post-hoc projection

Let d_e be a full-model update and let P_0 retain coordinates in S_0. We decompose

$$b_e = P_0 d_e, \qquad c_e = (I - P_0) d_e, \qquad d_e = b_e + c_e, \tag{9}$$

and compare the projected update b_e with an update a_e independently trained in S_0. The six constructed states—base, independent local, original full, projected full, complement, and recomposed

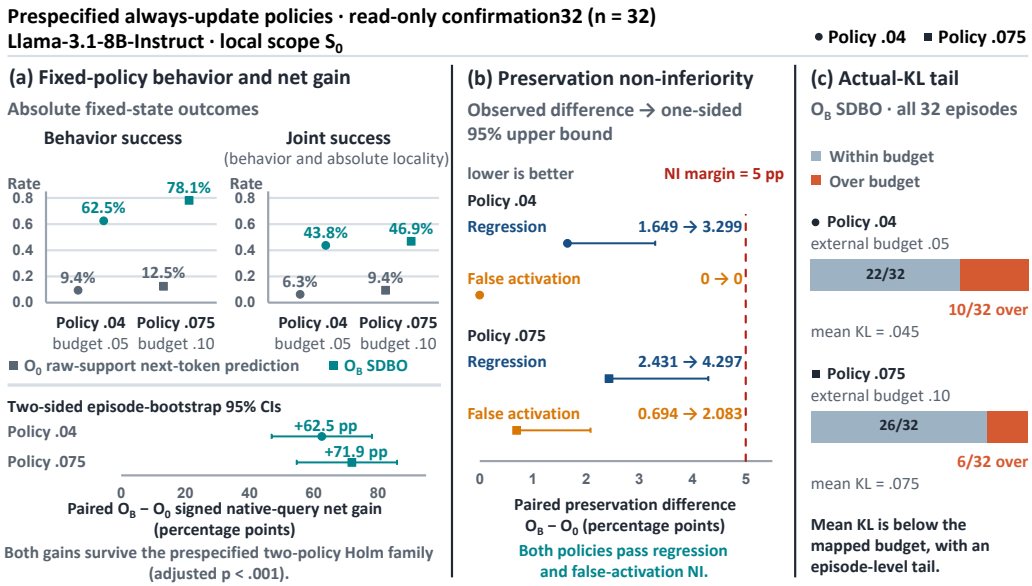


Figure 3: Fixed-policy behavior and preservation on confirmation32. **(a)** Absolute behavior and joint success, with paired SDBO–raw-support native-query net gains. **(b)** Paired preservation differences and one-sided 95% upper bounds against the .05 NI margin. **(c)** SDBO actual-KL exceedance counts under external budgets .05/.10.

full—separate the local coordinates of d_e from its complement. Both a_e and b_e end in the same parameter support, but only a_e was restricted to that support throughout optimization. Recomposition reproduces the original full update in all 64 episode–policy cases and has zero maximum probe-logit error.

The projected update has zero behavior at both fixed policies. Removing the complement sharply reduces perturbation without restoring behavior, whereas independent local training exceeds the projection by 62.5 and 78.1 percentage points (Figure 4). Thus, selecting local coordinates after full-model training is not interchangeable with training in those coordinates.

7.2 Effective direction after radius matching

Write $a_e = r_{a,e}u_{a,e}$ and $b_e = r_{b,e}u_{b,e}$. To compare directions at the independent update's radius, define the arithmetic target

$$\widehat{b}_{a,e} = \frac{r_{a,e}}{r_{b,e}} b_e \tag{10}$$

and denote its realized BF16 update by $\widetilde{b}_{a,e}$. Qualification and behavior are computed from $\widetilde{b}_{a,e}$, requiring nonzero source norm, scale at most 2048, relative-radius error at most .10, and source-direction cosine at least .99. All 64 fresh episode–policy cases qualify.

As a reverse check, the arithmetic target $\widehat{a}_{b,e} = (r_{b,e}/r_{a,e})a_e$ yields the realized update $\widetilde{a}_{b,e}$. None of its 64 cases meets the radius criterion, so this comparison remains intention to treat.

On fresh32, independent local training exceeds the realized radius-matched projection by 62.5 percentage points at both policies, with 95% intervals [43.8,78.1] and [46.9,78.1] points and Holm $p < .001$. Increasing radius along the projected direction therefore does not close the behavioral gap. The scale-b contrasts are 0 and $+9.4$ percentage points and do not meet the prespecified two-policy criterion. Effective Frobenius radius is matched, while actual KL remains lower for the scaled projection.

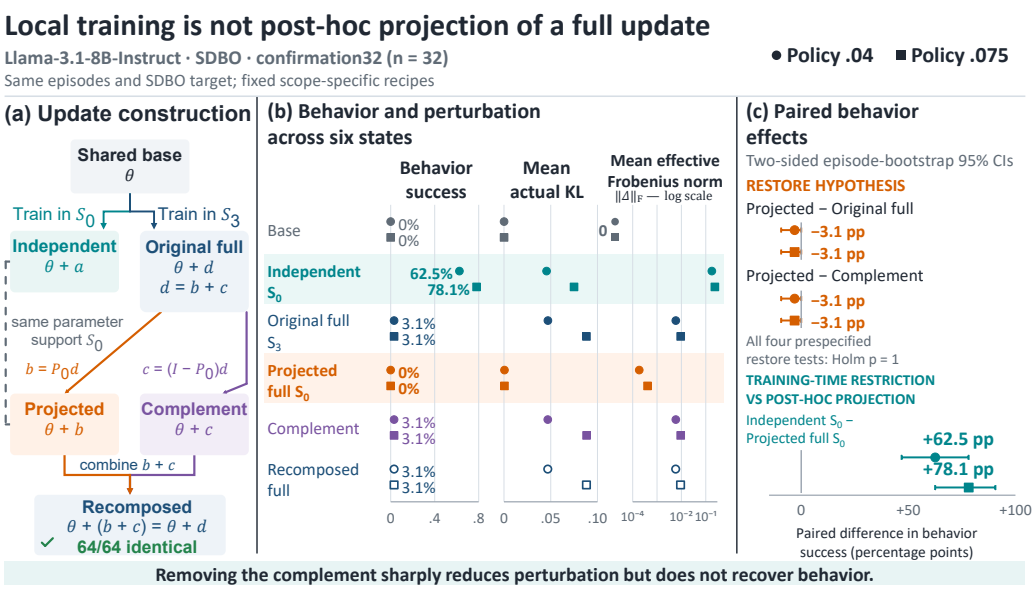


Figure 4: Local training versus post-hoc projection on confirmation32. **(a)** Decomposition of the same full-model update. **(b)** Behavior, actual KL, and effective Frobenius norm across constructed states. **(c)** Paired behavior differences with two-sided episode-bootstrap 95% confidence intervals. Complete state metrics are in Appendix Table A8.

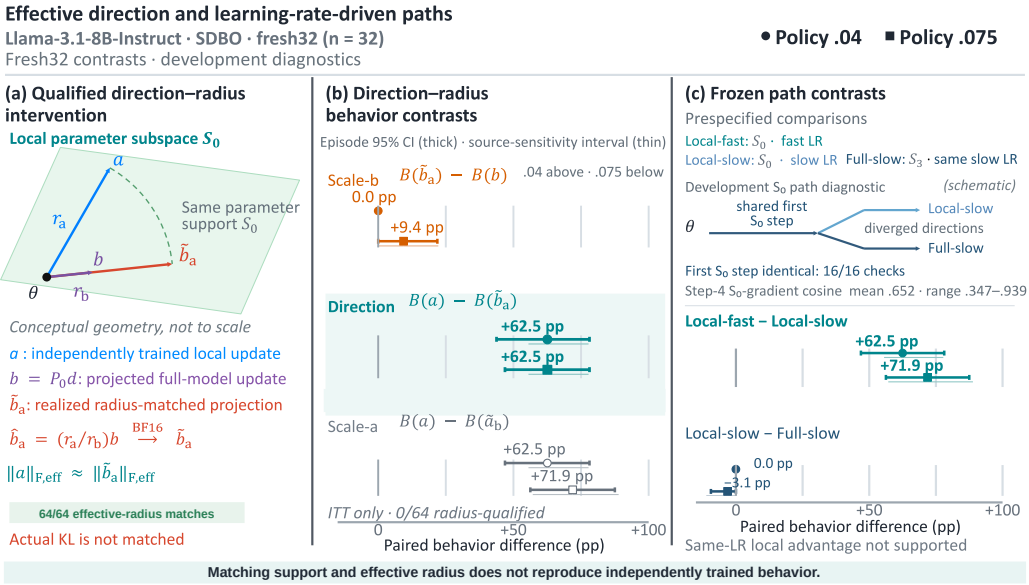


Figure 5: Direction–radius and path comparisons on fresh32, with separate development diagnostics. **(a)** Effective-radius matching after BF16 loading. **(b)** Paired scale, direction, and reverse-scale behavior contrasts; the reverse construction is numerically unqualified and shown as intention to treat. **(c)** Learning-rate and same-rate scope contrasts. Thick bars denote episode-bootstrap 95% intervals; thin bars denote source-composition sensitivity intervals.

7.3 The final direction reflects a multistep path

We compare Local-fast, which trains S_0 at the local-method learning rate; Local-slow, which trains the same scope at the full-model rate; and Full-slow, which trains S_3 at that rate. Local-fast minus

Local-slow isolates the learning-rate recipe within a fixed parameter scope. Local-slow minus Full-slow tests whether local scope remains advantageous when the learning rate is shared.

Equal-rate local and full paths have identical initial local gradients and first local FP32 master updates in all 16 development episode–policy checks. Their later local-gradient directions diverge because the surrounding parameter states differ. On fresh32, Local-fast exceeds Local-slow by 62.5 and 71.9 percentage points, with 95% intervals [46.9,78.1] and [56.3,87.5] points. Local-slow minus Full-slow is 0 and -3.1 percentage points, providing no positive same-rate local-scope advantage. Together with the matched-radius direction contrast and same-rate gradient trace, these comparisons support a learning-rate-driven multistep contribution to the final direction (Figure 5).

8 Discussion and Conclusion

A larger parameter set permits more updates, but a fixed optimization recipe need not produce the same effective direction within a local subspace. Projection and radius matching show that final parameter support and norm are insufficient to explain the observed behavior. The path comparisons further connect the separation to the learning-rate-dependent sequence of parameter states at which gradients are evaluated.

The fixed-policy result converts attained performance into an executable rule: one configuration is chosen before confirmation and applied uniformly, so gain, preservation, and KL belong to the same updated state.

The Llama result is supported by content controls, held-out formulations, fixed-policy evaluation, and fresh mechanism tests; Qwen3-8B-Base provides a contrasting model setting.

A single support-derived target and four local update steps create independently confirmed, content-specific behavior with positive fixed-policy gain and relative preservation. This behavior is not reproduced by projecting and rescaling a full-model update. Whether a test-time update becomes behavior depends on how the update is formed during training.

AI use statement

Generative AI tools were used for grammar and spelling checks, language polishing, related-work retrieval, and programming assistance for a subset of the experiments. All AI-assisted work was reviewed by the authors, who take responsibility for the final content of this work.

Reproducibility statement

Appendices B–D describe the episode sets, scoring rules, update configurations, statistical procedures, and numerical qualification criteria. Appendix E reports detailed results for the confirmation, preservation, decomposition, direction, and path analyses.

References

Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. Self-supervised test-time learning for reading comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1200–1211, 2021.

Rachit Bansal, Aston Zhang, Rishabh Tiwari, Lovish Madaan, Venkata Sai Surya Subramanyam Duvvuri, Devvrit Khatri, David Brandfonbrener, David Alvarez-Melis, Prajjwal Bhargava, Mihir Kale, and Samy Jelassi. Let's (not) just put things in context: Test-time training for long-context LLMs. In *International Conference on Learning Representations*, 2026.

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.

Guhao Feng, Shengjie Luo, Kai Hua, Ge Zhang, Wenhao Huang, Di He, and Tianle Cai. In-place test-time training. In *International Conference on Learning Representations*, 2026.

Moritz Hardt and Yu Sun. Test-time training on nearest neighbors for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CNL2bku4ra.

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. arXiv:2301.04213.

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*, pp. 24823–24849, 2025.

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=APuPRxjHvZ.

Yansheng Mao, Yufei Xu, Jiaqi Li, Fanxu Meng, Haotong Yang, Zilong Zheng, Xiyuan Wang, and Muhan Zhang. LIFT: A novel framework for enhancing long-context understanding of LLMs via long input fine-tuning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2502.14644.

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

Xiangchen Song, Zhenhao Chen, Lingjing Kong, Shaoan Xie, Xinshuai Dong, Guangyi Chen, and Kun Zhang. Beyond perplexity: A behavioral evaluation framework for deployment-memory claims in LLM test-time training. *arXiv preprint arXiv:2607.00368*, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

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.

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 (Volume 1: Long Papers)*, pp. 15336–15354, 2025. doi: 10.18653/v1/2025.acl-long.745.

A Early Proxy–Behavior Diagnostic

The initial diagnostic asked whether fitting support text was sufficient to change support-removed free generation. In a Qwen3-8B-Base facts-and-rules matrix, 5,700 of 5,760 condition-level records improved the training proxy and none improved the behavioral endpoint. In a smaller nonce-fact study, all 504 query-family behavior gains were also zero despite lower support and answer negative log-likelihood. These observations motivated the behavioral endpoints and support-derived target studied in the main experiments; they are not part of the confirmation sample.

B Data and Scoring

B.1 Episode sets

Table A1 separates discovery, confirmation, reuse analyses, numerical development, and the fresh mechanism sample. Reusing an episode set for paired analyses does not create additional independent observations.

Fresh32 was selected before outcome evaluation. Starting from 1,890 candidates, exclusions for prior update or outcome exposure and duplicate knowledge, subject, or relation exposure left 919 candidates. Selection used fixed source quotas, maximized relation coverage, and resolved remaining ties deterministically. Wrong-target controls exchange answers within source while approximately matching token length. Repetition duplicates support tokens without a canonical target, and no-support removes the episode value. Secondary repetition and no-support analyses use preallocated 16-episode subsets.

B.2 Generation, behavior, and preservation

A separate frozen alias-aware word-boundary audit checks scorer robustness without replacing the primary normalized-containment endpoint.

Each episode uses 12 general locality items sampled from a 64-item pool (three QA, three mathematics, two instruction-following, two safety, and two formatting items), four same-domain negatives, and five native locality items for natural QA. Base behavior is zero by the relative-acquisition definition; this is distinct from the base model's absolute QA accuracy. Teacher forcing is diagnostic only.

Table A1: Episode sets and their roles. Confirmation32 and fresh32 are disjoint from the sets used to select the corresponding comparisons.

Set	Composition	Use	Inferential role
Discovery32	32 natural QAEdit episodes	Select the objective–scope comparison	exploratory
Confirmation32	32 disjoint episodes; NQ 9/SQA 14/TQA 9	Method, interaction, content, and NI tests	formal confirmation
Confirmation reuse	confirmation32	Fixed policies and update decomposition	paired reuse
Development8	8 episodes $\times$ 2 policies	Numerical and trajectory diagnostics	development only
Fresh32	32 episodes; NQ 9/SQA 14/TQA 9; 18 relations	Direction–radius and path tests	fresh formal sample

Table A2: Generation and scoring definitions used throughout the paper.

Item	Definition
Generation	Greedy decoding with at most 64 new tokens and fixed tokenizer, prompt format, EOS, and stopping rules.
Scorer	Primary binary match after lowercasing, accent folding, and removing nonalphanumeric characters: a normalized target or fixed alias must occur in the normalized generation.
Source gain	Acquisition-probe success after the update minus success at the base model.
Native query gain	Mean signed success gain over direct and rephrased native queries; regressions remain negative.
Behavior B	Source gain ≥ 1 and mean native-query gain $\geq .25$.
Held-out behavior	Source gain ≥ 1 and independently constructed held-out gain ≥ 1.
Regression	Correct-to-wrong fraction among locality items answered correctly by the base model.
False activation	Erroneous activation rate on same-domain negatives.
Absolute locality L	Regression $\leq .05$, false activation $\leq .05$, and valid denominators.
Joint J	$B \times L$ at the same update state.
Actual KL	Token-averaged full-vocabulary $D_{\mathrm{KL}}(p_\theta \| p_{\theta+\delta})$ on replay prompts.
Preservation NI	Paired SDBO-minus-raw-support regression and false-activation differences; both one-sided 95% upper bounds $\leq .05$.

C Objectives, Parameter Scopes, and Calibration

C.1 Objectives and target construction

For QAEdit, the transform parses support of the form

```
Fact update: the {relation} of ``{subject}'' is ``{value}.''
This replaces the previous entry.
```

and emits

```
Entity: {subject}
Requested field: {relation}
Updated value:
```

with `value` as the target. The parser uses only the support and a relation schema. Question and formatting tokens are masked from the loss; target and EOS tokens remain supervised. Llama uses its chat template, whereas Qwen3-8B-Base uses the raw-prompt route associated with its base-model setting.

The four O_1 formulations present the same support-derived answer as a reference question, cloze, retrieval, or structured-record view. Evaluation queries are not used to construct these views.

Table A3: Training objectives and content controls.

Condition	Views	Training signal or role
O_0	support	Next-token prediction on serialized support.
O_B	1	Canonical support-derived question–answer target.
O_1	4	Four fixed support-derived acquisition formulations.
Wrong target	1	Within-source answer derangement.
Repetition	–	Repeated support without the behavioral target.
No support	–	Episode value removed.

Table A4: Learning rates for the two primary policy labels. Local-slow opens only S_0 but uses the Llama O_B/S_3 rate in the corresponding column.

		Llama-3.1-8B-Instruct		Qwen3-8B-Base	
Scope	Objective	.04	.075	.04	.075
S_0	O_0	1.5203×10^{-3}	2.3564×10^{-3}	2.6874×10^{-4}	4.0141×10^{-4}
S_0	O_B	9.8893×10^{-4}	1.5362×10^{-3}	1.2755×10^{-4}	1.9567×10^{-4}
S_0	O_1	9.3895×10^{-4}	1.6372×10^{-3}	1.4452×10^{-4}	1.9259×10^{-4}
S_3	O_0	5.1411×10^{-5}	8.7284×10^{-5}	4.4809×10^{-5}	6.3598×10^{-5}
S_3	O_B	3.0988×10^{-6}	5.6895×10^{-6}	5.2079×10^{-5}	8.0907×10^{-5}
S_3	O_1	3.0154×10^{-6}	5.8257×10^{-6}	4.4341×10^{-5}	6.2645×10^{-5}

C.2 Parameter scopes and optimization

Llama S_0 contains the eight complete `down_proj.weight` tensors in blocks 24–31, totaling 469,762,048 parameters. S_3 contains 291 tensors and 8,030,261,248 parameters. Qwen3-8B-Base uses the corresponding `mlp.down_proj.weight` tensors in blocks 27–35. All conditions use four SGD steps with zero momentum, weight decay, and dropout, and no gradient clipping. Forward/backward computation uses BF16, with resident FP32 master updates for trainable parameters.

Each model–objective–scope cell is calibrated separately on three pilot episodes. Calibration evaluates an initial learning-rate grid, forms a monotone lower-learning-rate envelope of median actual KL, and interpolates in log-KL/log-learning-rate space. It does not read behavioral outcomes. Table A4 reports the two fixed-policy rates used in the main operational and mechanism analyses.

D Statistical Inference and Numerical Qualification

D.1 Episode-paired inference

For an episode-level paired difference d_e, each of 10,000 bootstrap draws resamples N episode indices with replacement and averages their differences. All conditions belonging to an episode move together. Two-sided intervals use the 2.5th and 97.5th percentiles. Positive one-sided tests use

$$p_+ = \frac{1 + \sum_{b=1}^{10000} \mathbf{1}[\bar{d}^{*(b)} \le 0]}{10001}, \tag{11}$$

followed by Holm correction within the families in Table A5. Objective-by-scope interaction is

$$\mathcal{I}_\epsilon = \frac{1}{N}\sum_e \left[(W^J_{\epsilon,e}(O_B,S_0) - W^J_{\epsilon,e}(O_0,S_0)) - (W^J_{\epsilon,e}(O_B,S_3) - W^J_{\epsilon,e}(O_0,S_3))\right]. \tag{12}$$

Source-composition sensitivity resamples the three predefined NQ/SQA/TQA aggregate strata and pools their episode sums and counts. It measures sensitivity to source composition; it is not used as the primary confidence interval. For the qualified fresh32 direction effect, the estimates are .5556 for NQ ($n = 9$), .5714 for SQA ($n = 14$), and .7778 for TQA ($n = 9$). Both policy effects remain positive after leaving out any one source.

Table A5: Multiplicity families used for formal tests.

Family	Members	Direction
Method/scope	2 Llama local gains + 2 interactions	Positive joint gain and interaction
Content	2 models × 2 budgets	Correct-target minus wrong-target joint attainment
Decomposition	2 restore contrasts × 2 policies	Projected minus original full/complement
Fixed policy	2 policies	SDBO-minus-raw-support native-query net gain
Direction/radius	3 contrasts × 2 policies	Scale-b, direction, and reverse scale-a
Path	2 contrasts × 2 policies	Local-fast minus Local-slow; Local-slow minus Full-slow

D.2 Preservation NI and radius matching

For preservation outcome $c \in \{\mathrm{reg}, \mathrm{FA}\}$ and policy π, NI evaluates the paired mean

$$d_c^\pi = \frac{1}{N} \sum_e [C_{c,e}(\delta^\pi_{O_B,e}) - C_{c,e}(\delta^\pi_{O_0,e})]. \tag{13}$$

The upper bound is the 95th percentile of the paired bootstrap distribution, and a policy passes when both upper bounds are at most .05.

The cross-radius construction requires source effective norm above 10^{-12}, scale at most 2048, BF16 realized relative-radius error at most .10, and cosine with the source direction at least .99. An unqualified construction remains in the intention-to-treat behavior contrast but does not support a matched-radius interpretation. A two-policy mechanism chain additionally requires positive estimates at both policies, at least one Holm-significant family member, and at least 90% numerical qualification at both policies.

E Complete Results

E.1 Objective, scope, and fixed-policy results

The source-composition intervals for the two Llama local gains are [.2222,.4444] and [.3333,.4286]; the interaction intervals are [.22222,.44444] and [.33333,.44444]. Llama/S_0 raw-behavior attainment under SDBO is .625/.750 and held-out attainment is .625/.750 at budgets .05/.10. Correct-target minus wrong-target joint effects are +.43750 and +.50000 (episode intervals [.28125,.62500] and [.34375,.65625]); the corresponding Qwen effects are zero. Complete cross-formulation and secondary-control effects appear in Table 2.

Llama/S_3 passes relative NI at both policies: its regression differences are .01172/.01519 with uppers .02604/.03255, and false-activation differences are zero. The four Qwen scope–budget cells fail NI: the regression upper bounds are .20528/.43907 for S_0 at budgets .05/.10 and .29900/.42755 for S_3 at the same budgets. Qwen/S_3 at the higher budget also has false-activation difference .02083 with upper .06250.

E.2 Update decomposition

All 64 recomposed episode–policy states match the original full update under the state-identity check, with zero maximum probe-logit error. Projected-minus-original-full and projected-minus-complement behavior are $-.03125$ at both policies (Holm $p = 1$ for all four tests). Independent-minus-projected behavior is +.62500/+.78125, with held-out differences +.71875/+.81250.

E.3 Direction–radius and path results

The projected-b reload passes all 16 development identity checks. The primary scaled-b construction qualifies in 16/16 development and 64/64 fresh episode–policy cases, with maximum development relative-radius error .000548. The reverse construction qualifies in 0/16 development and 0/64 fresh cases and has maximum error .953098. In the same-learning-rate path diagnostic, the initial local gradients and first FP32 local updates agree in all 16 development episode–policy checks. At step four, the local-gradient cosine ranges .3465–.9394 with mean .6520.

Table A6: Complete confirmation32 objective–scope results. Panel A gives joint attained rates and fixed-state NI status. Panel B gives paired SDBO–raw-support effects and the two Llama interactions.

Panel A: joint attained rates and preservation

Model	Scope	Budget	O_0	O_B	O_1	Fixed NI
Llama	S_0	.05	.06250	**.43750**	.46875	pass
Llama	S_0	.10	.12500	**.50000**	.62500	pass
Llama	S_3	.05	0	0	0	pass
Llama	S_3	.10	.06250	.03125	.03125	pass
Qwen	S_0	.05	.03125	0	.06250	fail
Qwen	S_0	.10	.03125	0	.06250	fail
Qwen	S_3	.05	0	.06250	.03125	fail
Qwen	S_3	.10	0	.06250	.03125	fail

Panel B: paired effects

Model/effect	Scope	Budget	Estimate	Episode 95% CI	Holm p
Llama: $O_B - O_0$	S_0	.05	**+.37500**	[.18750,.56250]	.00039996
	S_0	.10	**+.37500**	[.15625,.59375]	.00059994
	S_3	.05	0	–	–
	S_3	.10	-.03125	–	–
Qwen: $O_B - O_0$	S_0	.05	-.03125	[-.09375,0]	–
	S_0	.10	-.03125	[-.09375,0]	–
	S_3	.05	+.06250	–	–
	S_3	.10	+.06250	–	–
Llama: objective × scope	–	.05	**+.37500**	[.18750,.56250]	.00059994
	–	.10	**+.40625**	[.21875,.59375]	.00059994

Table A7: Complete Llama/S_0 fixed-policy results on confirmation32. Policy labels .04/.075 map to external KL budgets .05/.10.

Panel A: absolute outcomes; query gain is relative to base

Policy	Objective	Behavior	Joint	Held-out	Query gain	Mean KL	Locality
.04	O_0	.09375	.06250	.12500	.06250	.02563	.81250
.04	O_B	**.62500**	**.43750**	**.71875**	**.68750**	.04546	.68750
.04	O_1	.96875	.65625	.87500	.93750	.04740	.65625
.075	O_0	.12500	.09375	.15625	.12500	.04511	.81250
.075	O_B	**.78125**	**.46875**	**.81250**	**.84375**	.07494	.65625
.075	O_1	1.00000	.59375	1.00000	1.00000	.08178	.59375

Panel B: paired native-query gain and SDBO KL tail

Policy	$O_B - O_0$	Episode 95% CI	Holm p	KL over budget
.04	**+.62500**	[.46875,.78125]	.00019998	10/32
.075	**+.71875**	[.546875,.859375]	.00019998	6/32

Panel C: paired preservation NI; parentheses give one-sided 95% upper bounds

Policy	Regression Δ (upper)	False activation Δ (upper)	NI
.04	.01649 (.03299)	0 (0)	pass
.075	.02431 (.04297)	.00694 (.02083)	pass

Table A8: Six-state decomposition on confirmation32. Each cell reports policies .04/.075.

Panel A: behavioral outcomes

State	Behavior	Joint	Held-out
Base	0/0	0/0	0/0
Independent S_0	**.62500/.78125**	**.43750/.46875**	**.71875/.81250**
Original full S_3	.03125/.03125	0/.03125	0/.06250
Projected local	0/0	0/0	0/0
Complement	.03125/.03125	0/.03125	0/.03125
Recomposed full	.03125/.03125	0/.03125	0/.06250

Panel B: perturbation and locality

State	Locality	Mean KL	Mean $\|\Delta\|_F$
Base	1/1	0/0	0/0
Independent S_0	.68750/.65625	.04546/.07494	.17681/.23565
Original full S_3	.71875/.65625	.04685/.08817	.005864/.009290
Projected local	.84375/.87500	.000413/.000421	.000186/.000406
Complement	.68750/.65625	.04683/.08833	.005860/.009280
Recomposed full	.71875/.65625	.04685/.08817	.005864/.009290

Table A9: Eight evaluated states on fresh32. Each cell reports policies .04/.075.

State	Behavior	Joint	Held-out	Mean KL	Locality
Base	0/0	0/0	0/0	0/0	1/1
O_0 local	0/0	0/0	.03125/.06250	.024729/.051540	.93750/.96875
Independent S_0 a	**.62500/.71875**	**.46875/.46875**	**.62500/.68750**	.055493/.079794	.75000/.68750
Projected local b	0/0	0/0	0/0	.000418/.000416	.96875/.96875
Scaled $b \to r_a$	0/.09375	0/.06250	.03125/.12500	.006478/.013618	.90625/.87500
Scaled $a \to r_b$	0/0	0/0	0/0	.000274/.000355	.96875/.96875
Full S_3	0/.03125	0/0	0/.06250	.054438/.108557	.81250/.62500
Local-slow	0/0	0/0	0/0	.000426/.000423	.96875/.96875

Table A10: Prespecified direction–radius and path contrasts on fresh32. Effects are behavior-rate differences evaluated on realized BF16 updates. Reverse scale-a is intention to treat: 0/64 constructions satisfy the radius criterion.

Family	Contrast	Policy	Effect (episode 95% CI)	Holm p
Direction/radius	Scale-b: $B(\widetilde{b}_a) - B(b)$	.04	0 [0,0]	1
		.075	+.09375 [0,.21875]	.0877912
	Direction: $B(a) - B(\widetilde{b}_a)$	.04	+.62500 [.43750,.78125]	.00059994
		.075	+.62500 [.46875,.78125]	.00059994
	Reverse scale-a: $B(a) - B(\widetilde{a}_b)$	.04	+.62500 [.46875,.78125]	.00059994
		.075	+.71875 [.56250,.87500]	.00059994
Path	Local-fast minus Local-slow	.04	+.62500 [.46875,.78125]	.00039996
		.075	+.71875 [.56250,.87500]	.00039996
	Local-slow minus Full-slow	.04	0 [0,0]	1
		.075	-.03125 [-.09375,0]	1