WHEN SPARSE ROUTING LOOKS ADDITIVE: HIDDEN PAIRWISE STRUCTURE CHANGES ROUTE CHOICES

Anonymous authors
Paper under double-blind review

ABSTRACT

An additive fit to mean sparse-route loss need not preserve token-local effects or valid route choices. We audit temporal preservation and choice validity separately using complete four-band intervention surfaces and exact functional ANOVA. Beyond guaranteed energy attenuation under averaging, pairwise effects retain less normalized energy than unary effects on average. In a fresh 40-source, two-model study, the mean within-surface pairwise share of nonadditive gold-NLL energy is 86.72%; retaining pairwise terms reduces pooled mean hindsight regret versus unary by 77.5%, equally weighting separately evaluated NLL/KL endpoints. On 160 further documents, degree-two projection fitted to the same 70 feasible native Quest routes as unary reduces mean native-reference KL hindsight regret by 59.9% under fixed Llama/H100 execution. A prespecified support-construction intervention on 80 new documents narrows the retention gap at fixed K and mean overlap. With per-document calibration, two separate 48-document eight-band studies confirm held-configuration prediction at cohort-mean and document levels, including 28.69% lower mean document NLL RMSE versus unary in the prespecified Llama-interleaved condition. Better prediction can still yield a worse route. These findings motivate evaluating approximation-induced choices directly and reporting support construction when comparing preservation gaps.

1 INTRODUCTION

Sparse attention reduces long-context inference cost by selecting pages, blocks, chunks, or tokens across layers (Tang et al., 2024; Yuan et al., 2025; Lu et al., 2025; Gao et al., 2025; Hu et al., 2026b). When several routing decisions change together, do their next-token effects combine additively?

Cross-layer systems choose where route indices are computed or reused (Deshmukh et al., 2025; Bai et al., 2026; Sun et al., 2026; Zan et al., 2026). An additive sequence-level surrogate can appear to justify treating band choices independently, yet select a nonoptimal route while explaining most surface energy (Figure 1). Moreover, an interaction can help at one token and hurt at another, remaining locally material but disappearing in the mean. We find substantial pairwise structure in these loss surfaces. In separate prespecified tests, per-document pairwise calibration improves unseen-configuration prediction, and aligned native fits reduce mean hindsight choice regret. We test these consequences alongside order-dependent temporal preservation.

We enumerate four-band route surfaces under fixed teacher-forced histories, allowing internal activations to differ. Exact functional ANOVA (fANOVA) decomposes effects before and after averaging. Beyond convexity-guaranteed raw energy attenuation, we test normalized order-dependent retention and exact-surface choices. A temporal gap need not cause a choice error, or vice versa.

The evaluation lesson is to measure local structure, held-configuration prediction and actual choices separately. *High additive fit certifies neither token-local independence nor valid choice; preservation comparisons also require reporting or controlling support construction.* These findings are complementary, not a causal chain.

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.

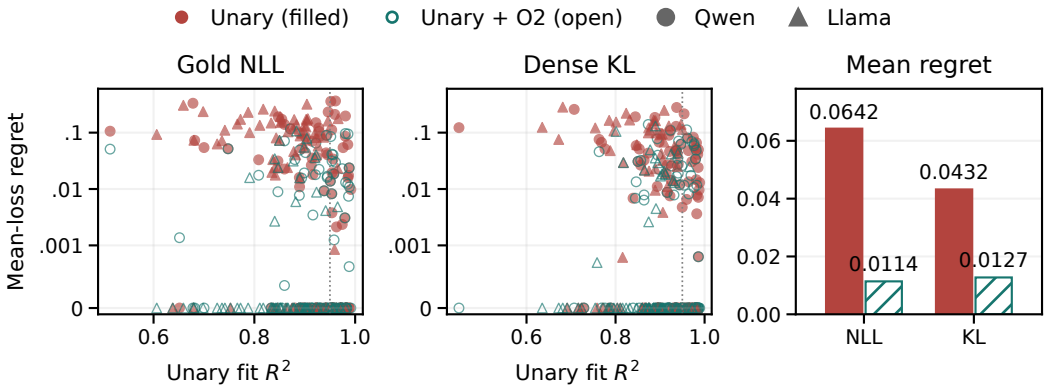

Figure 1: Fit and choice answer different questions. **Left/center:** All 320 comparisons, nested in 40 sources, with actual-loss regret for each projection (filled: unary; open: unary plus O2). Symmetric-log regret axes retain zeros; $R^2 = 0.95$ is an opened descriptive reference, not a validated rule. **Right:** Endpoint-separated mean regrets. The 77.5% reduction is a ratio of pooled means, not a pure-NLL or mean-percentage improvement.

- We introduce a complete token-state route-surface audit that separates the Jensen-guaranteed effect of averaging from empirical order-differential persistence and tests whether an additive aggregate preserves exact-surface choices.
- Pairwise reconstruction improves hindsight choice over unary on complete natural-document surfaces. A separate 160-document native study confirms average regret gains with matched fitting and choice sets; further studies extend domain, checkpoint and horizon coverage.
- On 80 new documents, we confirm a prespecified direction of support-construction influence on the unary–pairwise preservation gap, even with per-layer K and mean overlap held fixed.
- Two separate 48-document studies confirm calibrated prediction over unseen eight-band settings: first on the mean surface, then on document-first error against unary and a frozen shared-pair prior. Better prediction still need not yield a better choice.

Our claims concern endpoint surfaces, retrospective choices and separately calibrated predictions, not hidden-state mechanisms or a deployable router.

2 Related Work

Sparse-attention systems. Quest selects query-relevant KV pages without training (Tang et al., 2024); trained/adapted systems learn block, chunk or token selection (Yuan et al., 2025; Lu et al., 2025; Gao et al., 2025; Hu et al., 2026b). Controlled content and route-replay interventions already distinguish filtering from information integration (Ren et al., 2026). We instead measure order-wise temporal preservation and aggregate choice validity on complete joint surfaces under a common token history.

Cross-layer index reuse (Deshmukh et al., 2025; Bai et al., 2026; Sun et al., 2026; Zan et al., 2026) motivates the placement axis; reuse itself is systems context, not our contribution.

Feature interactions and attribution. Joint prompt attribution and scalable interaction explainers already model nonadditive effects in language models (Chang et al., 2025; Kang et al., 2025). ProxySPEX also studies within- and cross-layer attention-head interactions for pruning (Butler et al., 2025). We instead audit order-wise temporal preservation and route-choice regret, using fANOVA as a measurement tool, not a new decomposition or search algorithm.

Conditional importance and compensation. Downstream components can compensate for ablations via LayerNorm rescaling and anti-erasure (Rushing & Nanda, 2024). Conditional Co-Ablation

(CoAx) uses conditional effect growth and pairwise logit synergy, averaging energies across inputs/tokens (Gong et al., 2026). Our endpoint audit does not identify such a hidden-state compensation mechanism.

Aggregation and cancellation. Kramár et al. (2024, Eq. 5 and Appendix B.2) address cancellation across prompt pairs by averaging effect magnitudes rather than taking the magnitude of the average effect, separately from direct–indirect gradient cancellation. We therefore do not claim signed-effect cancellation under averaging as a new observation. Our empirical distinction is order-differential preservation on complete sparse-route lattices, together with a separate test of whether aggregate unary projections preserve exact-surface choices.

Temporal importance and sparse decoding. Temporal importance and accumulated long-decode error (Han et al., 2026; Zhang et al., 2026; Hu et al., 2026a) are deployment concerns, as are selector/kernel costs (Nawrot et al., 2026); they motivate but differ from our audit axes.

3 Problem and Estimand

Let $a = (a_1, \dots, a_4) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_4$ denote a fixed intervention rule over four depth bands and the target window. Native rules may recompute physical supports at each token. At position t on a common teacher-forced token history, $F_t(a)$ is the measured loss of intervention a. Token inputs are aligned; internal activations may depend on the route. The sequence-level surface is

$$\bar{F}(a) = \frac{1}{T} \sum_{t=1}^{T} F_t(a). \tag{1}$$

Cartesian audits freeze a product measure μ over their finite action lattice (Sobol, 1993; Hooker, 2007). Let P_S be the corresponding fANOVA projection for nonempty factor subset $S \subseteq \{1, 2, 3, 4\}$, write $f_{S,t} = P_S F_t$, and define $\|g\|_\mu^2 = \mathbb{E}_{a\sim\mu}[g(a)^2]$. Projection is linear, so

$$P_S \bar{F} = \frac{1}{T} \sum_{t=1}^{T} P_S F_t, \tag{2}$$

$$\left\| \frac{1}{T} \sum_t f_{S,t} \right\|_\mu^2 \le \frac{1}{T} \sum_t \|f_{S,t}\|_\mu^2. \tag{3}$$

Equation 3 follows from Jensen's inequality, not an empirical result. We therefore distinguish token energy $E_S^{\mathrm{tok}} = T^{-1} \sum_t \|f_{S,t}\|_\mu^2$, aggregate energy $E_S^{\mathrm{agg}} = \|T^{-1} \sum_t f_{S,t}\|_\mu^2$, and normalized temporal coherence $\rho_S = E_S^{\mathrm{agg}} / E_S^{\mathrm{tok}}$ when the denominator is positive. For order k, we sum energies before taking the ratio:

$$\rho_k(H) = \frac{\sum_{|S|=k} E_S^{\mathrm{agg}}(H)}{\sum_{|S|=k} E_S^{\mathrm{tok}}(H)}, \tag{4}$$

where (H) restricts both quantities to the first H target positions. Thus ρ_k is not an unweighted average of component coherences. O1 is unary, O2 pairwise, and O2+/O3+ denote degrees at least two/three. With $e_{k,t} = \sum_{|S|=k} \|f_{S,t}\|_\mu^2$ and $e_{+,t} = \sum_{k=1}^{4} e_{k,t}$, closure energy shares are $q_k = \sum_t e_{k,t} / \sum_t e_{+,t}$, $q_{2+} = \sum_{k=2}^{4} q_k$, and $s_2 = q_2 / q_{2+}$. Raw attenuation is guaranteed; order-wise retention is empirical.

Permuting measured surfaces within a completed interval preserves its coherence, but changes shorter prefix intervals and possibly their fitted decay slope. Neither statistic is an adjacent-token correlation time or a refresh interval.

The primary measure is uniform over the predefined Cartesian lattice. It is a reference measure that gives every audited schedule equal weight, not an estimate of a deployed router's visitation distribution.

Axis A measures temporal preservation. Axis B tests *choice validity* on $\mathcal{A}_{\mathrm{feas}} \subseteq \mathcal{A}$: the full lattice in Cartesian audits, fixed-Hamming-weight subsets in native audits. Let $P_{\le d} = P_\varnothing + \sum_{1 \le |S| \le d} P_S$,

including the mean, and define

$$\hat{a}_d = \arg\min_{a \in \mathcal{A}_{\text{feas}}} (P_{\leq d} \bar{F})(a), \qquad r_d = \bar{F}(\hat{a}_d) - \min_{a \in \mathcal{A}_{\text{feas}}} \bar{F}(a), \quad d \in \{1, 2\}. \tag{5}$$

Positive regret demonstrates a decision-relevant ordering error. Zero regret certifies an optimal choice, not the ordering of all remaining candidates. Both axes concern the specified surface; neither is a mediator for the other. Product-measure components need not remain orthogonal on a constrained subset; full-lattice projection and feasible-set refitting test different approximations.

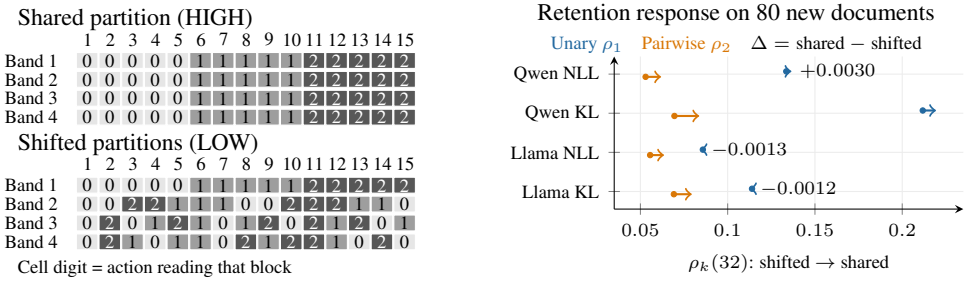

Figure 2: The support-construction condition test. **Left:** A canonical block ordering illustrates the two recipes; actual layouts randomize block identity. Each action selects five blocks; local and causal keys stay visible. Both recipes have mean overlap $5/3$, not different mean reading budgets. **Right:** Descriptive mean retentions move from shifted to shared. Annotations are unary-retention changes $\Delta\rho_1$, not gap changes Γ; paired gap intervals are in Table 3.

4 Complete-Surface Route Audit

We enumerate a frozen lattice, window and endpoint, checking dense replay before and after. Order/closure studies use 4,096-token prefixes and 32 teacher-forced predictions. Band choices are fixed; the 256-token local suffix and causal keys remain visible. Q/K/V are route-specific, without value patches. Appendix B defines these libraries; Figure 2 depicts the separate support intervention.

Exact tokenwise and mean-surface fANOVA requires no coalition sampling. The prespecified decay exponent is $\alpha_k = -\text{slope}(\log H, \log \rho_k(H))$ from ordinary least squares over $H \in \{4, 8, 16, 32\}$. The order study tests $(\alpha_2 + \alpha_3 + \alpha_4)/3 - \alpha_1$; the 40-source closure tests $\alpha_2 - \alpha_1$. These are finite-window summaries, not asymptotic laws.

Controlled tokenwise fractions $T^{-1} \sum_t e_{k,t}/e_{+,t}$ differ from the closure's ratios of summed energies q_k before source averaging (Appendix B.1).

Source identity is the statistical unit: controlled/order studies bootstrap 16 shared queries by model; the 40-source study averages models within source before family-balanced bootstrap and requires each model's prespecified direction. Retrospective projections use all observations; Section 6.7's restricted-calibration regression is not exact fANOVA. Follow-ups designed after inspecting the data are labeled *opened* and kept separate from prespecified confirmation.

5 EXPERIMENTS

Table 1: Primary studies. Source counts are distinct identities shared across models. Action spaces are per library; order persistence and pairwise consequence use two libraries. The last row reuses the preceding 64 sources; prediction uses two separate 48-document cohorts and samples its lattice.

Question	Models	Source	Space	Sources	Status
Cancellation	Qwen 4B, OLMo 7B	GovReport	4^4	16	prospective
Order persistence	Qwen 4B, Llama 8B	PG-19/GovReport	3^4	16	fresh
Pairwise consequence	Qwen 4B, Llama 8B	PG-19/GovReport	3^4	40	fresh
Native Quest transfer	Quest-Llama 8B	PG-19/GovReport	2^4	40	registered
Native feasible set	Quest-Llama 8B	PG-19/GovReport	70	160	fresh
Learned-selector scope	Seer-Qwen 4B	AIME	2^4	12	development
Support intervention	Qwen 4B, Llama 8B	PG-19/GovReport	3^4	80	fresh
Held-route prediction	Qwen 4B, Llama 8B	PG-19/GovReport	3^8	$48 + 48$	fresh
Horizon dependence	Qwen 4B, Llama 8B	PG-19/GovReport	3^4	80	fresh
Domain scope	Qwen 4B, Llama 8B	Papers/Python	3^4	64	fresh
Checkpoint scope	Qwen 32B	Same papers/Python	3^4	64	matched

Controlled and natural-text studies. Controlled GovReport queries (Huang et al., 2021) use Qwen3-4B-Instruct-2507 and OLMo-3-1025-7B, four-of-16-block schedules and 64 gold tokens. The natural-text order and 40-source studies balance PG-19 (Rae et al., 2020) and GovReport, using Qwen and Llama-3.1-8B-Instruct with two five-of-15-block libraries. Order-study primaries are Qwen and Llama; OLMo-3-7B-Instruct is secondary. The 40-source study tests gold-NLL nonadditive energy and pairwise share, plus equally weighted NLL/KL decay and regret gains. Coupled energy summaries are not independent confirmations (Appendix C).

Native and learned-selector neighborhoods. Quest uses Llama-3.1-8B-Instruct and 40 sources disjoint from closure. Native boundary swaps preserve K and layer-specific Q/K/V, with endpoint and feasible sets

$$F_t(a) = D_{\mathrm{KL}}(p_{\mathrm{native},t} \,\|\, p_{a,t}), \qquad \mathcal{A}_h = \left\{ a \in \{0,1\}^4 : \sum_i a_i = h \right\}, \quad h \in \{1,2,3\}.$$

Native 0000 has zero self-KL and is excluded. Full-lattice projections and opened slice-unary fits are scored within $\mathcal{A}_h$ (Appendix A.4), not as improvements over native Quest. A separate 160-document study fits both degrees directly on 70 weight-four eight-band candidates under fixed H100 execution (Appendix P). SeerAttention-R on its pinned Qwen3-4B backbone instead uses four Full/Shared index-computation factors on 12 AIME queries. Its temporal object is the O2+ share decrease, not pure-pair coherence (Appendix A.5); choice again uses fixed-weight sets. Neither comparison benchmarks kernel speed; equal K need not imply equal latency.

Intervention, prediction and scope. Later studies test support construction, calibrated prediction, and QASPER/RepoBench scope (Dasigi et al., 2021; Liu et al., 2024) with a matched-source 32B extension. These use 4K prefixes/32 targets, not QA or completion scores. A separate 80-document study keeps 4K prefixes but uses nested 32/64/128 targets (Appendix O).

6 RESULTS

6.1 SEQUENCE AGGREGATION SUPPRESSES TOKEN-LOCAL INTERACTION

The controlled query-mean tokenwise degree-three-plus fraction is 0.08170 versus aggregate 0.01335 in Qwen, and 0.04834 versus 0.01165 in OLMo (Figure 3). All 16 source gaps are positive in both models. Centered low-order coherence also exceeds degree-three-plus coherence, with positive paired-bootstrap lower limits in both models. Unlike raw attenuation, this normalized order difference is not guaranteed by Equation 3.

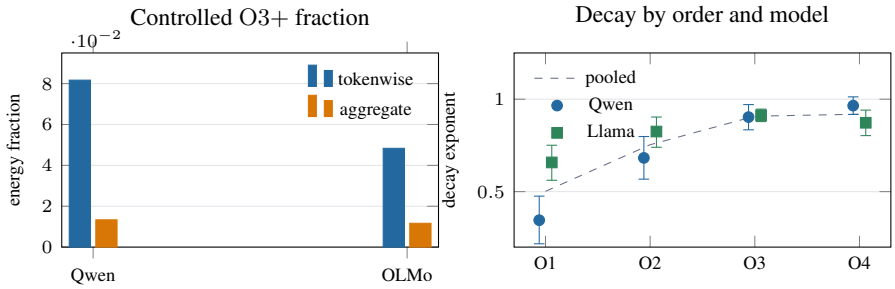

Figure 3: Temporal cancellation across two populations. **Left:** Controlled degree-three-plus fractions: mean of per-token fractions versus fraction of the 64-token mean surface. **Right:** Model means and opened pointwise 95% intervals (16 source units, corpus-stratified); the dashed line retains the pooled exponents. These are not simultaneous intervals or a strict per-query order ladder.

6.2 Aggregate retention decreases with interaction order

On the untouched natural-text population, the higher-minus-unary decay gap is 0.5047 with a 95% query-bootstrap interval $[0.3668, 0.6293]$ for Qwen and 0.2123 $[0.1202, 0.3127]$ for Llama. All four model-by-source-family means are positive. Pooled order-specific exponents increase from 0.5017 at unary order to 0.7537, 0.9080, and 0.9188 at orders two through four (Figure 3, right).

The contrast is positive in 28/32 primary model-query units; three Llama and one Qwen PG-19 queries have the opposite sign. Llama's mean O4 exponent (0.8726) is below O3 (0.9130). The confirmed mean contrast is not a strict ladder for every model or query.

6.3 Pairwise structure is material on both audit axes

In the 40-source study, gold-NLL O2+ accounts for 0.1466 of token-averaged nonconstant energy, with family-balanced 95% source-bootstrap interval $[0.1345, 0.1600]$, above the registered 0.10 operational floor. Within that residual, the O2 share is 0.8672 $[0.8578, 0.8761]$. Both models have the prespecified direction; Appendix A gives model-specific values.

On Axis A, averaging the two fixed NLL/KL endpoints gives a pairwise-minus-unary decay-exponent gap of 0.2106 $[0.1410, 0.2792]$; the model-specific gaps are 0.3316 and 0.0895. On Axis B, degree-two reconstruction reduces the equally weighted mean of separately evaluated NLL/KL 32-token hindsight regrets from 0.0537 to 0.0121, a gain of 0.0416 $[0.0304, 0.0537]$ and a 77.5% relative reduction (Figure 1). The registered decision gain is positive for $37/40$ source units. Exact-optimum matches rise from $120/320$ to $207/320$: pairwise terms correct 118 unary misses and create 31 new errors. These 320 descriptive comparisons are nested within 40 source units. Projection follows full-surface observation, so the result establishes hindsight decision relevance on this population, not an inference-time algorithm.

Robustness to measure, route quality, and audit budget. Opened unary fit has median $R^2 = 0.9134$; among 91 of 320 comparisons with $R^2 \geq 0.95$, unary misses 46 optima. Their regret falls from 0.03067 to 0.01222 with O2. This threshold is descriptive; a residual–margin diagnostic does not yield a useful unary safety gate (Appendix E). Zero-forward checks retain all four conclusions under every moderate $(.6, .2, .2)$ product-measure skew, and decision gain on each cell's 21 most dense-faithful schedules (0.0388 $[0.0280, 0.0503]$). On the recorded 33-cell design, pairwise beats unary but not best-measured search; this gain is design-sensitive. Strong $(.8, .1, .1)$ concentration exposes a genuine measure boundary (Appendix A.3).

Prospective longer-window choice. After eight-document development (Appendix N), a frozen study on 80 new documents tests $\Delta_{128} - \Delta_{32}$, where Δ_H is the equally weighted NLL/KL regret gain above. One 128-target tensor supplies nested horizons under the same 4K prefix. The primary source-paired change is -0.01029 $[-0.02036, -0.00021]$ (two-sided equal-corpus Welch 95% interval): the anticipated absolute decrease is supported, with the upper bound near zero. Yet H128 mean regret

falls from 0.03920 to 0.00575 with pairwise terms, an 85.3% reduction versus 80.8% at H32. All eight H128 model/library/loss gain intervals are positive, but nominal secondary, not simultaneous. Absolute gains decrease, but relative value persists; causal mediation is untested (Appendix O).

An endpoint-null decision control. On 16 opened sources (64 surfaces), additive-logit-null O2 with actual unary fixed recovers $-1.9\%/21.6\%$ of Qwen's NLL/KL choice gain and $-4.6\%/7.5\%$ for Llama. Direct selection on the entire null instead recovers 55.7%/99.6% and 12.3%/16.2%, respectively. These distinct controls have model/endpoint-dependent reach (Appendix C.2.1).

6.4 NATIVE QUEST: REGISTERED AUDIT AND ALIGNED FEASIBLE-SET CHOICE

Table 2: Separate native-KL hindsight studies, not a before/after comparison. Intervals use 40/160 document units; (a) averages three choices, (b) uses fixed H100 execution. Fit labels exclude execution checks. Allocation exception: Appendix G.

	(a) Registered: 40 docs	**(b) Fresh: 160 docs**
Fit set (labels U/P)	Full 2^4 cube (16/16)	$\mathcal{A}_4^{(8)}$ (70/70)
Choice set	$h = 1, 2, 3$ slices	Same $\mathcal{A}_4^{(8)}$
$r_U \to r_P$ (10^{-4})	$1.200 \to 0.211$	$3.479 \to 1.396$
Gain [95% CI] (10^{-4})	0.989 $[0.411, 1.747]$	2.082 $[0.908, 3.844]$
Mean regret reduction	82.4%	59.9%
Exact optimum hits (U/P)	$77 \to 103$ of 120	$32 \to 35$ of 160
Unary–pair coherence [95% CI]	0.0433 $[0.0245, 0.0628]$	Not tested

Both comparisons audit perturbation fidelity, not improvement over unperturbed Quest. All original 40 sources pass native replay/same-K checks and both audit criteria pass: coherence and regret gains are positive in $33/40$ and $24/40$ sources. Regret gain is 5.54% of mean within-weight action spread (Appendix A.4).

This four-band comparison narrows under feasible-set refitting: slice-unary matches $106/120$ optima versus cube-pairwise's $103/120$, arithmetically closing 87.58% of the regret gap; weights one and three are saturated. Weight-two gain $[-1.50, 11.31] \times 10^{-5}$ and pooled gain $[-0.87, 3.43] \times 10^{-5}$ intervals include zero, with six versus sixteen labels at weight two. These opened controls do not establish incremental superiority (Appendix A.4).

Fresh confirmation against feasible-set unary. After eight-document development, a frozen Llama/H100 study collects 160 new documents. Both projections fit uniformly to the same 70 eight-band weight-four actions, with ranks 8/28; neither is saturated. Mean regret falls 59.9% (Table 2b); the primary interval and prespecified Welch sensitivity are positive. Gains improve/worsen/tie on 93/41/26 documents, despite only three additional exact hits. This confirms *average regret*, not input-wise optimality. All labels precede choice; the result is conditional hindsight evidence, not hardware invariance (Appendices F,P).

6.5 SEERATTENTION-R: CHOICE RELEVANCE WITH UNCONFIRMED SHARE ATTENUATION

SeerAttention-R development tests Full/Shared index-computation placement. Pairwise terms reduce regret from 3.08×10^{-5} to 1.12×10^{-5} KL ($9/12$ positive) and raise exact matches from $18/36$ to $26/36$. Its nonadditive-share decrease is 0.03337 $[-0.01117, 0.08709]$, so transport remains unconfirmed rather than reversed. Binary geometry explains much of the smaller effect scale but not all query heterogeneity (Appendix C). This complements Quest without establishing a cross-system ranking.

6.6 SUPPORT CONSTRUCTION PROSPECTIVELY CHANGES THE TEMPORAL ORDER GAP

At fixed per-layer K, *Shared partition* (HIGH) reuses one three-way partition across four bands; *Shifted partitions* (LOW) use balanced ternary shifts (Figure 2). Both have five-block actions and

mean overlap $5/3$. Mean cumulative remote coverage also matches exactly ($325/27$ blocks); action-wise coverage distributions, grouping and depth use differ. HIGH/LOW remain the frozen record IDs (Appendix H).

For each complete surface, compute ρ_1 and ρ_2 at horizon 32, then

$$\Gamma = (\rho_1 - \rho_2)_{\mathrm{LOW}} - (\rho_1 - \rho_2)_{\mathrm{HIGH}}.$$

After opened development, we freeze the prediction of positive mean Γ on both Qwen NLL and dense-KL before new outcomes. Eighty new documents (40 per corpus) receive four layouts; contrasts are averaged within document. Both individual 95% lower limits must exceed zero. Llama is prespecified secondary.

Table 3: Fresh support intervention. Qwen NLL/KL are co-primary; Llama is secondary. Individual equal-corpus Welch-t intervals use 80 document units, not 320 layouts, and are not simultaneous. Positive Γ means gap narrowing.

Model	Endpoint	LOW gap	HIGH gap	Γ [95% CI]
Qwen	NLL	0.08058	0.07466	0.00592 [0.00138, 0.01047]
Qwen	KL	0.14225	0.13565	0.00661 [0.00065, 0.01256]
Llama	NLL	0.03020	0.02087	0.00933 [0.00368, 0.01497]
Llama	KL	0.04478	0.03330	0.01148 [0.00404, 0.01893]

Both primary criteria pass (Table 3). Pairwise retention increases more than unary (Figure 2). Qwen's mean gap falls by 7.35%/4.64% for NLL/KL; Llama's secondary decreases are 30.88%/25.64%. These are ratios of mean gaps, not explained causal fractions. All mean gaps remain positive; Qwen has positive source means in only 48/80 and 50/80 cases. The result concerns an average under the frozen layout procedure, not an input-wise law. Direct model-difference intervals include zero.

An opened eight-document Shared-repertoire control retains the Qwen contrast (Appendix I); rotation and endpoint controls only partly constrain its explanation (Appendix H.4). HIGH increases Qwen's all-action mean KL by 0.01039 $[0.00070, 0.02009]$: narrowing the preservation gap is not evidence of better fidelity or a better selected policy.

6.7 Calibrated pairwise models predict unseen configurations

Splitting each band in two yields $3^8 = 6{,}561$ routes, retaining tied four-band routes. After eight-document development, frozen unary/pairwise ridge models are tested on 48 new documents, balanced across GovReport/PG-19. Each supplies 129 fit and 32 secondary-tuning observations (161 for the full comparison); 128 disjoint held settings never enter fitting/tuning. Original $U, P, P_{1/2}$ use only the 129 fit labels. Calibration is *per new document*.

Table 4: NLL RMSE reductions relative to unary (%). C/I: contiguous/interleaved. C1: mean-surface and opened document-first gains on 48 documents. C2: document-first gains on 48 further documents for original P and the prespecified $P_{1/2} = (U + P)/2$. C2 brackets are individual crossed 95% intervals; half-blend intervals are opened nominal reporting. Bold marks the primary condition for original P only.

Model/library	C1 surface (%)	C1 document (%)	C2 original P (%) [95% CI]	C2 fixed half (%) [95% CI]
Qwen/C	36.87	−4.56	0.01 [-7.45, 6.90]	18.82 [15.37, 22.11]
Qwen/I	39.19	6.28	6.56 [-0.52, 13.15]	18.30 [15.09, 21.40]
Llama/C	49.53	23.54	23.03 [17.97, 27.86]	26.41 [23.97, 28.82]
Llama/I	60.23	23.53	28.69 [22.26, 34.72]	28.69 [25.96, 31.36]

The first cohort's primary mean-surface gain, 60.23% $[28.20, 64.29]$, passes. Fixed-design ridge commutes with source averaging, but squared error does not: mean-surface evaluation removes cross-document error variation. Relative RMSE gains need not have a fixed ordering. Qwen-C's mean-surface regret worsens from zero to 0.04399 despite 36.87% lower RMSE (Appendix K).

The separately frozen second cohort tests *mean document RMSE* with the same budget and 128 new held settings. Its shared-pair comparator fixes pair coefficients from C1 training labels and refits unary per document. The Llama-interleaved primary reduces error by 28.69% $[22.26, 34.72]$ versus unary and 28.06% $[22.10, 33.72]$ versus this prior, passing all criteria and improving 45/48 and 46/48 documents, respectively. Original P's secondary Qwen NLL intervals include zero.

Predictor mixing and estimator controls. Fixed $P_{1/2}$ improves mean document RMSE in all four C2 cells: 18.82%/18.30% in Qwen C/I, with 48/48 improving documents in each. It mixes both lower-order and pair coefficients; no held label chooses its weight. Opened controls retain the gain with half the fitted pair field on fixed U, not lower-order changes alone. Qwen-C half-blend lowers mean held-set regret from 0.08672 to 0.05911, but has 12/8/28 better/worse/tied choices despite 48/48 RMSE improvements. Rankings are within 128 candidates, not global optima. The 130-query eight-band pilot favors direct search over these frozen 129-fit procedures (Appendices L,D.4). An opened 161-label block-ridge control further improves prediction but worsens Llama choice regret relative to fixed half-blend (Appendix L.2).

6.8 Both audit objects extend to new domains and a larger checkpoint

Complete $3^4 \times 32$ audits on 32 new NLP papers and 32 Python repository clusters pass all eight NLL criteria (two models/domains/objects). A frozen Qwen3-32B extension on the *same opened documents* passes its four-component conjunction (Table 5).

Table 5: Domain/checkpoint scope at $H = 32$ (NLL). Individual intervals resample 32 source units per domain, averaging libraries within source. The 32B block adds a checkpoint, not independent documents.

Model	Text domain	$\rho_1 - \rho_2$ [95% CI]	$r_U - r_P$ [95% CI]
Qwen 4B	NLP papers	0.1036 [0.0768, 0.1321]	0.0810 [0.0398, 0.1271]
Qwen 4B	Python	0.0686 [0.0464, 0.0930]	0.0543 [0.0252, 0.0895]
Llama 8B	NLP papers	0.0252 [0.0100, 0.0416]	0.0566 [0.0310, 0.0865]
Llama 8B	Python	0.0472 [0.0258, 0.0707]	0.0689 [0.0442, 0.0960]
Qwen 32B	NLP papers	0.2029 [0.1654, 0.2420]	0.0668 [0.0273, 0.1112]
Qwen 32B	Python	0.1505 [0.1088, 0.1924]	0.0748 [0.0554, 0.0945]

Training, checkpoint and band width (9 versus 16 Qwen layers) also change, so 32B extends coverage rather than identifying a scale effect (Appendix M).

7 Discussion

Consequences for evaluation. Validate choices separately from fit and prediction: small mean errors can cross decision margins (Appendix E). Match fitting measure and feasible set: the new native confirmation aligns them, unlike the older cube-to-slice comparison.

Acquisition. Direct search wins the tested budgets (Appendix D), but conditions on joint choices and is not a unary model. Better prediction does not establish cheaper acquisition. A separate 80-document test finds 5.02% anchor-relative KL reduction for a non-surrogate policy, but not direct-search superiority (Appendix Q).

Scope and explanatory boundaries. Support-recipe effects do not identify an internal mechanism (Appendices C.2,J). Fixed-history studies, including H128, establish neither generated quality nor arbitrary-band or model-family generality.

8 Conclusion

Order-dependent retention coexists with predictive and choice-relevant pairwise structure. Validate choices directly, reporting support construction, endpoint, fitting measure and feasible set.

Reproducibility Statement

Hash-bound manifests and appendix protocols specify model revisions, source-level units, interventions, endpoints, interval procedures, mechanics and stop rules. The anonymous supplementary package contains code, tests, saved measurements, frozen selectors, decision traces, analysis protocols and a SHA-256 inventory. These cover the complete-surface and endpoint-null studies, support and OA controls, both prediction cohorts, domain/32B and horizon extensions, native feasible-set confirmation, calibration/regularization diagnostics, and the prospective calibration-policy test. To meet the 100 MB upload limit, four large token-array files are omitted from the review edition; the README distinguishes endpoint-array replay from inference recomputed using included document-level records. Export code and hashes identify the omitted inputs. Full token-level reanalysis of those studies requires those inputs; the documented review-edition commands are self-contained. CPU replay does not regenerate model outputs or LM-head replay. A separate optional local-model example executes a small mask audit with repeat/no-op/restoration checks (Appendix B.2); it is not a rerun of all historical studies or a new empirical confirmation.

Ethics Statement

This work evaluates pretrained language models using public or licensed text and mathematical problem sources. It involves no human subjects, user profiling, or deployment decisions.

AI Use Statement

Generative AI tools assisted with manuscript drafting and revision; literature retrieval, related-work discovery and reference checking; research ideation, experimental design, implementation, numerical analysis and interpretation; and developing and checking mathematical derivations and proof arguments. Assistance also covered figure and artifact preparation. References were checked against primary sources. Code and numerical analyses were checked using the tests and independent recomputations described in the reproducibility materials; these checks do not constitute formal proof verification. The author takes responsibility for the final text, mathematical claims, code, experimental decisions, reported results and citations.

References

Yushi Bai, Qian Dong, Ting Jiang, Xin Lv, Zhengxiao Du, Aohan Zeng, Jie Tang, and Juanzi Li. IndexCache: Accelerating sparse attention via cross-layer index reuse, 2026. URL https://arxiv.org/abs/2603.12201.

Landon Butler, Abhineet Agarwal, Justin Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. ProxySPEX: Inference-efficient interpretability via sparse feature interactions in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 72306–72340, 2025. doi: 10.52202/085713-2426. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/68dced1638d71d4d21598f79ae91262e-Abstract-Conference.html.

Yurui Chang, Bochuan Cao, Yujia Wang, Jinghui Chen, and Lu Lin. JoPA: Explaining large language model's generation via joint prompt attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 22106–22122, 2025. doi: 10.18653/v1/2025.acl-long.1074. URL https://aclanthology.org/2025.acl-long.1074/.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of NAACL-HLT*, pp. 4599–4610, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Dhruv Deshmukh, Saurabh Goyal, Nipun Kwatra, and Ramachandran Ramjee. Kascade: A practical sparse attention method for long-context LLM inference, 2025. URL https://arxiv.org/abs/2512.16391.

Yizhao Gao, Shuming Guo, Shijie Cao, Yuqing Xia, Yu Cheng, Lei Wang, Lingxiao Ma, Yutao Sun, Tianzhu Ye, Li Dong, Hayden Kwok-Hay So, Yu Hua, Ting Cao, Fan Yang, and Mao Yang. SeerAttention-R: Sparse attention adaptation for long reasoning, 2025. URL https://arxiv.org/abs/2506.08889.

Zhiren Gong, Zihao Zeng, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits, 2026. URL https://arxiv.org/abs/2607.01940v1.

Ruizi Han, Miao Zhang, Ziyue Qiao, and Liqiang Nie. Evolving sparsity: Leveraging token importance dynamics for efficient LLM decoding with sparse attention. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 11554–11566, 2026. doi: 10.18653/v1/2026.acl-long.530. URL https://aclanthology.org/2026.acl-long.530/.

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. doi: 10.1198/106186007X237892.

Junhao Hu, Fangze Li, Mingtao Xu, Feifan Meng, Shiju Zhao, Tiancheng Hu, Ting Peng, Anmin Liu, Wenrui Huang, Chenxu Liu, Ziyue Hua, and Tao Xie. Lil: Less is less when applying post-training sparse-attention algorithms in long-decode stage. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1898–1912, 2026a. doi: 10.18653/v1/2026.findings-acl.91. URL https://aclanthology.org/2026.findings-acl.91/.

Xiang Hu, Xinyu Wei, Hao Gu, Minshen Zhang, Tian Liang, Huayang Li, Lei Zhu, Yan Wang, Sirui Han, Yushi Bai, Kewei Tu, Haitao Mi, and Leo Liang. Hierarchical sparse attention done right: Toward infinite context modeling, 2026b. URL https://arxiv.org/abs/2607.02980.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1419–1436. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.112. URL https://aclanthology.org/2021.naacl-main.112/.

Justin Singh Kang, Landon Butler, Abhineet Agarwal, Yigit Efe Erginbas, Ramtin Pedarsani, Bin Yu, and Kannan Ramchandran. SPEX: Scaling feature interaction explanations for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28878–28903. PMLR, 2025. URL https://proceedings.mlr.press/v267/kang25a.html.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745v1.

Tianyang Liu, Canwen Xu, and Julian McAuley. RepoBench: Benchmarking repository-level code auto-completion systems. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pPjZIOuQuF.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, et al. MoBA: Mixture of block attention for long-context LLMs, 2025. URL https://arxiv.org/abs/2502.13189.

Piotr Nawrot, Jianing Li, Renjie Huang, Sebastian Ruder, Kelly Marchisio, and Edoardo Ponti. The sparse frontier: Sparse attention trade-offs in transformer LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38667–38701, 2026. doi: 10.18653/v1/2026.findings-acl.1926. URL https://aclanthology.org/2026.findings-acl.1926/.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Xingyu Ren, Youran Sun, Chugang Yi, and Haizhao Yang. Understanding sparse attention selectivity in long-context foundation models via counterfactual evaluation. *arXiv preprint arXiv:2608.01676*, 2026. URL https://arxiv.org/abs/2608.01676.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855. PMLR, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Ilya M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.

Yutao Sun, Yanqi Zhang, Li Dong, Jianyong Wang, and Furu Wei. You only index once: Cross-layer sparse attention with shared routing, 2026. URL https://arxiv.org/abs/2606.06467.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-aware sparsity for efficient long-context LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47901–47911. PMLR, 2024. URL https://proceedings.mlr.press/v235/tang24l.html.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Y. X. Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention, 2025. URL https://arxiv.org/abs/2502.11089.

Wen Zan, Jiaqi Zhang, Jianchao Tan, Hong Liu, Cunguang Wang, Xiang Li, Duyue Ma, Guanyu Wu, Yifan Lu, Fengcun Li, Yerui Sun, Peng Pei, Yuchen Xie, and Xunliang Cai. LongCat sparse attention: Taming the lightning via streaming-aware hierarchical cross-layer indexing, 2026. URL https://arxiv.org/abs/2608.01662.

Haoyue Zhang, Hualei Zhang, Xiaosong Ma, Jie Zhang, and Song Guo. LazyEviction: Lagged KV eviction with attention pattern observation for efficient long reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 36335–36352, 2026. doi: 10.18653/v1/2026.acl-long.1683. URL https://aclanthology.org/2026.acl-long.1683/.

APPENDIX READER'S GUIDE

The appendices distinguish temporal preservation, predictive accuracy and choice validity. *Opened* analyses use already-inspected data and are not additional independent confirmations. *Confirmation* evaluates a question and decision rule fixed before the relevant new outcomes; its scope remains specific to the stated population and protocol. *Closure* names the separate 40-source, two-model pairwise-consequence study. Evidence status and local provenance are detailed in Appendix C; historical allocation exceptions are in Appendix G.

Reader's question	Where to read
How are surfaces and statistics defined?	Measurement contracts and aggregation: Appendix B.
Does good fit preserve route choices?	Complete-surface results, fit diagnostics and endpoint controls: Appendices A, E and C.2.1.
What qualifies the native comparisons?	Registered Quest, execution sensitivity and aligned feasible-set confirmation: Appendices A.4, F and P.
What does support construction change?	Recipe intervention, repertoire balance and pair-incidence controls: Appendices H, I and J.
What predicts unmeasured routes, and at what acquisition cost?	Separate prediction cohorts and acquisition comparisons: Appendices K, L and D.
How far do domain and horizon results extend?	Domain/checkpoint scope, opened horizon study and prospective horizon test: Appendices M, N and O.
Does prompt-only calibration improve held-out fidelity?	A separate prospective policy test and measured stage costs: Appendix Q; not a pairwise optimizer claim.

A ADDITIONAL RESULTS

A.1 CONTROLLED AND NATURAL-TEXT ESTIMATES

Table 6 collects the earlier controlled, fresh-order, and opened-audit quantities. The separately prespecified fresh closure follows in Table 7; opened reconstructions are never relabeled as confirmation.

Table 6: Main numerical results and evidence roles. Fractions use the study-specific nonconstant-energy aggregation in Appendix B.1.

Evidence	Quantity	Qwen	Other family
Prospective	token O3+ fraction	0.08170	0.04834 (OLMo)
Prospective	aggregate O3+ fraction	0.01335	0.01165 (OLMo)
Fresh	decay gap	0.5047	0.2123 (Llama)
Fresh	95% CI	[0.3668,0.6293]	[0.1202,0.3127]
Opened	token pairwise fraction	0.21255	0.19410 (OLMo)
Opened	aggregate pairwise fraction	0.09805	0.08542 (OLMo)
Opened	unary–pairwise coherence	0.05975	0.07091 (OLMo)

The natural-text exact-sign diagnostic is positive in 28 of 32 primary model-query units. Its four exceptions are three Llama PG-19 queries and one Qwen PG-19 query, with gaps from -0.0055 to -0.1706. This diagnostic was report-only; the registered population confidence intervals and all four model-by-source-family means pass. OLMo-3-7B-Instruct is positive on all 16 secondary query units, with a gap of 0.4518 and interval $[0.3711, 0.5425]$, but its lower dense continuation quality prevents it from rescuing or vetoing the two primary models.

A.2 FRESH PAIRWISE CLOSURE

The separate closure study uses 40 source identities as the independent units. Table 7 reports the four prespecified components. The first two rows are coupled summaries of the same gold-NLL energy

Table 7: Fresh pairwise closure. Pooled intervals come from the registered family-balanced source bootstrap. Model columns are fixed scope-guard point estimates, not separately powered confidence intervals.

Quantity	Pooled estimate [95% CI]	Qwen	Llama
O2+ / nonconstant energy	0.1466 [0.1345, 0.1600]	0.1432	0.1501
O2 / O2+ energy	0.8672 [0.8578, 0.8761]	0.8551	0.8794
Pairwise-minus-unary decay	0.2106 [0.1410, 0.2792]	0.3316	0.0895
Unary-minus-degree-two regret	0.0416 [0.0304, 0.0537]	0.0299	0.0534

decomposition and are not independent confirmations. The latter two average the frozen gold-NLL and dense-KL endpoints within source.

All four pooled lower bounds exceed their frozen floors $(0.10, 0.50, 0, 0)$, and all eight model-specific directions pass. The source-level directions are $33/40$ above the materiality floor, $40/40$ above the pairwise-majority floor, $33/40$ with positive temporal gaps, and $37/40$ with positive regret gains. Every model-by-family mean and both route-library means have the registered direction. The pairwise composition sensitivity based on a ratio of mean normalized energies is 0.8567, compared with 0.8672 for the primary mean of cellwise ratios.

Across the 320 model–source–library–endpoint choice comparisons, exact optimum matches increase from 120 to 207. Degree-two terms correct 118 unary misses and create 31 new errors. These rows are descriptive and nested within the 40 resampled source identities; they are not 320 independent observations.

Endpoint-separated decision results (opened breakdown). Table 8 disaggregates the existing choices without refitting or changing their tie rules. NLL and KL each select their own minimizer. The pooled relative reduction divides the pooled regret gain by pooled unary regret; it is not a mean of the two percentage reductions. The 160 comparisons per endpoint remain nested within the same 40 sources.

Table 8: Existing 32-token choices separated by endpoint. No new confidence test replaces the registered pooled result. Regrets are in raw loss units.

Endpoint	Unary regret	Pairwise regret	Reduction	Optimum matches
Gold NLL	0.06420	0.01139	82.3%	56/160 → 103/160
Dense KL	0.04318	0.01274	70.5%	64/160 → 104/160
Equal endpoint mean	0.0537	0.0121	77.5%	120/320 → 207/320

A.3 Opened robustness sensitivities

Table 9 reports the complete P294-X0 sensitivity grid computed from the immutable P293 surfaces with zero model forwards. The product-measure analysis enumerates every assignment of the favored action; its reported lower bound bootstraps the minimum across assignments rather than selecting a favorable mode. Dense-fidelity restrictions rank schedules by 32-token mean dense KL. The budgeted analysis uses one nested D-optimal design constructed from schedule coordinates without endpoint access. Reproduction uses its recorded action IDs, not a new QR pivot selection.

Observation-design sensitivity and reproduction. Pivoted QR has nonunique choices for this schedule matrix. Regenerating the design on an x86 numerical stack changes the 33-cell pairwise-over-unary gain to 0.00392 $[-0.01976, 0.02347]$, versus the recorded-design row above. These are different measured action sets, not rounding of one fitted result. The gain is therefore conditional on the realized design, not invariant over valid QR designs. The package now includes the historical action lists, extracted after verifying the original full-result hash; its default replay checks raw-surface reconstruction against the original source rows and intervals without regenerating the design. Explicit QR regeneration remains an opened sensitivity, not a substitute result or a new confirmation.

Table 9: Opened artifact-only robustness checks. Every row was fixed before the new metric was computed and cannot replace the fresh P293 primaries.

Sensitivity	Result	Interpretation
All 81 favored-action modes, $(.6, .2, .2)$	worst temporal LB 0.1052; worst decision LB 0.0172	all four original floors survive
All 81 modes, $(.8, .1, .1)$	composition and temporal floors survive; materiality and worst decision interval do not	strong concentration remains a measure boundary
Best dense-KL quarter (21/81 routes)	regret gain 0.0388 [0.0280, 0.0503], $36/40$ positive	weakest routes are not required
Same 33/81 measured cells	pairwise-over-unary gain 0.0197 [0.0072, 0.0330]	helps on the recorded design
Pairwise versus best measured cell, 33/81	-0.0017 [-0.0127, 0.0090]	no advantage over direct measured search
Held-out RMSE, 41/81 cells	gain 0.0435 [0.0372, 0.0503]	better route-cell reconstruction

The moderate-grid result shows that uniform weighting is not necessary within this outcome-independent family. The strong-grid failure prevents a stronger measure-invariance claim, and none of these weights estimates a deployed router's visitation distribution. Likewise, the budgeted result reduces the cost of an offline audit relative to full enumeration but still evaluates request-specific route cells and supplies no inference-time predictor.

A.4 Registered Quest native-neighborhood result

The 40 Quest sources are disjoint from the 40-source closure and from the two opened Quest mechanics rows. They exhaust the remaining rows in the pre-existing inventory, balanced across source family. A later coverage audit identified one prior allocation omitted from that inventory; the registered result is preserved alongside a 39-source sensitivity (Appendix G).

Table 10: Registered architecture-native Quest result. The page-swap action preserves K and each layer's own Q/K/V.

Quantity	Estimate [95% CI]	Positive	Classification
Unary-minus-pairwise coherence	0.0433 [0.0245, 0.0628]	$33/40$	confirmed
Unary-minus-pairwise regret	0.000099 [0.000041, 0.000175]	$24/40$	confirmed

All 40 surfaces pass all-zero native equivalence, complete repeat, native replay, same-state one-page swap, and fANOVA reconstruction. Median dense/native KL is 0.0535, median top-1 agreement is 0.9062, and no row has low unary or pairwise energy. Decision gains are positive/zero/negative on $24/11/5$ sources. Across the fixed-Hamming-weight sets $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ per source, matches to the constrained native-KL minimizer increase from $77/120$ to $103/120$.

Mean within-weight unary and pairwise regrets are 1.20×10^{-4} and 2.11×10^{-5}, respectively. Their 82.4% relative reduction corresponds to 5.54% of the mean within-weight action spread. Pairwise terms correct 30 unary misses, introduce 4 errors, and have corrected-cell oracle-margin median 1.11×10^{-4}.

An opened source-level diagnostic did not support a positive monotonic association between temporal and ranking effects: Spearman correlations are -0.263 on the closure's source-level means and -0.481 on Quest. This is consistent with not treating Axis A as a mediator of Axis B; it does not establish causal independence or exclude nonlinear relations.

An independent implementation reads all 40 raw NPZ surfaces, reconstructs the binary fANOVA without the primary analyzer, and exactly reproduces both point estimates, both bootstrap intervals, and all choice counts. Quest's pages and Seer's Full/Shared placement remain different causal coordinates.

Feasible-set unary refit (opened). The primary projections use the full 16-cell product measure; their choice sets have 4, 6 and 4 cells at weights 1, 2 and 3. An additive model $\beta_0 + \sum_i \beta_i a_i$ has rank four on each set, so weights one and three are saturated. At weight two, its uniform least-squares projection is

$$(P_{\text{slice}}y)(a) = \text{mean}_{b\in\mathcal{A}_2}\, y(b) + \tfrac{1}{2}\{y(a) - y(\mathbf{1}-a)\}, \qquad y = \bar{F}.$$

We cross-check this closed form with an independent least-squares solve. For the new slice choices only, the constant-free complement differences are compared in exact arithmetic on the stored floating-point means, with lexicographic ties. All original full-cube choices remain unchanged.

Table 11: Quest opened fitting-scope audit. Hits are out of 40 at each weight. The last two columns are weight-two regrets in units of 10^{-4} native-reference KL. Cell counts describe observed outcomes, not equal measured runtime; slice cells are 4/6/4 for weights 1/2/3.

Projection	Cells	$h=1$	$h=2$	$h=3$	Mean r	Max r
Full-cube unary	16	16	27	34	1.927	35.893
Full-cube pairwise	16	32	31	40	0.526	5.982
Slice unary	4/6/4	40	26	40	1.000	11.502
Anchored unary	5	40	33	31	3.662	98.775

Refitting at weight two reduces unary regret, but the full-cube pairwise projection still has 47.4% lower mean regret and 31 rather than 26 optimum matches. It corrects eight slice-unary misses and introduces three. Its regret is lower/higher on 9/8 sources and equal on the other 23, so it does not dominate every source. Across all weights, slice-unary has $106/120$ matches versus $103/120$ for full-cube pairwise; the saturated weights therefore matter to the headline comparison.

An additional opened paired-source bootstrap places the improvement beyond slice-unary in context (Table 12). Average the three within-weight differences inside each source for the pooled row, then resample 20 sources per corpus, 20,000 times (seed 202609183221). The pooled and weight-two intervals both include zero. These are approximate individual intervals conditional on the observed action sets, not new confirmation or simultaneous coverage. The original cube-unary primary remains unchanged. Across all weights the mean regrets are 0.0001199651 (cube-unary), 0.0000333480 (slice-unary) and 0.0000210663 (cube-pairwise). The arithmetic path through slice-unary closes 87.58% of the original mean gap, but this is not a causal share attributable to measure mismatch.

Table 12: Opened slice-unary minus cube-pairwise regret. Values and intervals are in 10^{-5} native-reference KL units. Signs count sources; zero differences are retained. There are 40 independent sources, not 120.

Scope	Gain [10^{-5}; 95% CI]	+	−
Overall	1.23 [-0.87, 3.43]	8	12
$h=1$	−1.06 [-2.05, -0.26]	0	8
$h=2$	4.74 [-1.50, 11.31]	9	8
$h=3$	0.00 [0.00, 0.00]	0	0

Refitting pairwise on the six weight-two choices gives rank six and thus interpolates every outcome. Such a comparison would be algebraically saturated, not a new validation. The main 40-source ternary audit instead fits and selects on the same full product domain, so this Quest fitting-set issue does not explain away that result.

Large-error cases also prevent an accuracy-only reading. The largest slice-unary weight-two regret is 1.1502×10^{-3}; pairwise reduces it to 5.9823×10^{-4} but still misses the optimum. On another source slice-unary is exact while pairwise incurs 5.3063×10^{-4} regret. All source-level choices and failures are retained in the artifact. The anchored baseline uses only native and four singleton losses; its 33/40 weight-two hits coexist with a much larger worst regret.

This tests a specified least-squares representation, not impossibility of every unary scoring rule. The slice fit already observes every feasible outcome, so direct selection would have zero regret. Full-cube pairwise uses sixteen outcomes rather than six; no query-efficiency claim follows.

Near-optimal selection sensitivity (opened). We keep each method's original selected action $\hat{a}_{x,h}^{(m)}$ fixed and evaluate $\mathbf{1}[r_{x,h}^{(m)} \le \epsilon]$, where $r_{x,h}^{(m)} = \bar{F}_x(\hat{a}_{x,h}^{(m)}) - \min_{a \in \mathcal{A}_h} \bar{F}_x(a)$. Figure 4 retains every empirical regret breakpoint; neither projections nor actions are refitted as tolerance changes. On these opened sources, the pairwise empirical rate is never below the unary rate across the full curve. Both reach 100% once tolerance exceeds all observed regrets. This describes constrained fidelity selection, not generated-quality or systems improvement, and introduces no new promotion threshold.

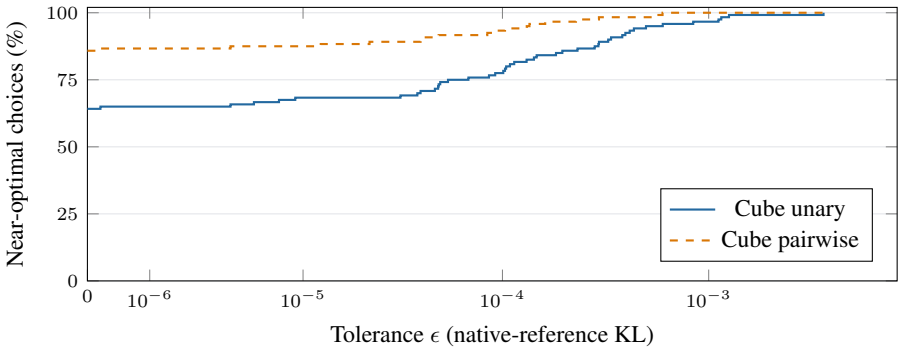

Figure 4: Opened near-optimal sensitivity of the original full-cube Quest choices. Rates average the three Hamming sets within each source, then the 40 sources; the 120 choices are not independent samples. No inferential band is reported. Horizontal position uses $\log(1 + \epsilon/10^{-6})$ to include zero; tick labels show ϵ. At zero, counts are $77/120$ and $103/120$.

A.5 Released-selector scope

Seer's registered temporal statistic uses token-averaged energies:

$$\Delta_{\text{share}} = \frac{E_{2+}^{\text{tok}}}{E_1^{\text{tok}} + E_{2+}^{\text{tok}}} - \frac{E_{2+}^{\text{agg}}}{E_1^{\text{agg}} + E_{2+}^{\text{agg}}}, \qquad E_{2+} = E_2 + E_3 + E_4.$$

It is distinct from Quest's pure-pair coherence contrast. Its sign compares retention of unary and all nonadditive energy, while its magnitude also depends on their initial shares. Choice regret averages the three fixed-weight sets $\mathcal{A}_h$, $h = 1, 2, 3$, within each query.

Table 13: SeerAttention-R development scope. Mechanics, fidelity, and hindsight choice relevance pass; nonadditive-share attenuation is unconfirmed.

Check	Object	Result
Execution	native replay	pass
Fidelity	local reference agreement	pass
Headroom	identifiable route variation	pass
Temporal	nonadditive-share attenuation	unconfirmed
Choice	pairwise hindsight relevance	pass

Δ_{share} has mean 0.03337, median 0.02595, and $8/12$ positive query units. Every leave-one-query-out mean remains positive, but the registered query-bootstrap interval includes zero. Pairwise choices have mean raw regret improvement 0.00001961 $[0.00000402, 0.00004059]$, with $9/12$ positive query units and 56.8% mean per-query relative reduction. Mean unary and pairwise regrets are 3.08×10^{-5} and 1.12×10^{-5} KL. Every decision-gain leave-one-query-out mean is positive, ranging from 0.00001106 to 0.00002202 KL.

The opened zero-forward choice reconstruction finds 18/36 exact oracle matches under the unary projection and 26/36 under the pairwise projection. Pairwise terms correct 10 unary misses and create 2 new errors. These 36 rows are nested within 12 query units and are descriptive, not an enlarged sample.

The 56.8% averages $(r_q^U - r_q^P)/r_q^U$ over all 12 queries after averaging each query's regrets over Hamming weights. Under the original convention, a query with $r_q^U = 0$ contributes zero (one query, also with $r_q^P = 0$); negative gains remain included. In contrast, $1 - \bar{r}^P/\bar{r}^U$ gives 63.7% reduction in mean regret, using unrounded values. An opened scale calibration instead divides the summed improvement by the summed within-Hamming-weight action spread and yields 1.44%. This denominator was not registered and is reported only to prevent the relative-regret statistic from being mistaken for a system-level quality gain. This concerns Full/Shared placement on the measured local surface; it does not show that Seer's native block scorer omits a pairwise term.

B Implementation and Analysis Details

Table 14 collects the complete execution contract for the headline closure study in one place; all layer ranges have inclusive endpoints. The paragraphs below provide the corresponding details for the full study set.

Table 14: Compact reproduction contract for the headline 40-source study. The released manifests bind every source identity and file hash.

Item	Exact contract
Models and bands	Qwen3-4B-Instruct-2507: layers 0:8, 9:17, 18:26, 27:35; Llama-3.1-8B-Instruct: 0:7, 8:15, 16:23, 24:31. Exact revisions are listed below.
Source selection	First 20 untouched confirmation rows per family in a pre-existing SHA-256 split order; all 40 are used without replacement. PG-19 and GovReport each contribute 20 distinct documents.
Token window	No chat template or special tokens. For each native tokenizer, take the 4,096 tokens immediately before the bound document character boundary and the following 32 tokens. Token sequences and document contents are hash checked at execution.
Intervention	Split the first 3,840 prompt tokens into 16 blocks of 240. In depth band b, action a_b retains one predefined five-of-15 remote set for all 32 prediction rows. Remote block 0 is excluded; the final 256 prompt tokens and causal target keys are always visible. Values are not directly patched; downstream activations and Q/K/V may differ across routes.
Surface and endpoints	Two fixed route libraries, $3^4 = 81$ schedules each, complete lexicographic enumeration. Store all 32 teacher-forced gold NLL, dense-to-sparse KL, and dense/sparse top-1 agreement values.
Inference	Source identity is the unit. Average Qwen and Llama within source, balance PG-19 and GovReport, and use 50,000 stratified percentile bootstrap replicates with shared indices (seed 2026083193).
Numerics	Float64 analysis; missing, duplicate, nonfinite, or incomplete cells invalidate the affected component. Schedule ties use tolerance 10^{-12} and lexicographic index order. Dense pre/post replays are bitwise exact.
Compute	6,720 physical forwards (13,120 sequence-forward equivalents) on one NVIDIA GB10; 5.81 total elapsed hours. Peak allocated CUDA memory is 8.20 GiB for Qwen and 15.92 GiB for Llama. No optimizer step.

B.1 Surface contracts

The controlled prospective surface uses four route factors with four actions per factor, 256 schedules per query, and 64 stored gold-token NLL values per schedule. Both natural-text studies use four factors with three actions per factor, 81 schedules per library/query cell, and 32 stored target states. The order study has 16 source identities; the separate pairwise closure has 40 new identities, two route libraries, and two fixed model systems. Quest uses four binary same-K page-swap bands and 16 schedules on 40 additional sources; Seer uses four binary Full/Shared placements and 16 schedules. These Cartesian schedules are complete lexicographic products; no outcome selects a subset of cells. The separate 70-action constrained study is defined in Appendix P.

The controlled study divides the 3,840 remote prompt tokens into 16 blocks of 240 tokens and uses the four fixed policies $\{0, 1, 2, 6\}$, $\{0, 3, 4, 7\}$, $\{1, 3, 5, 8\}$, and $\{2, 4, 5, 9\}$ at each depth band. The

natural-text studies use the same block size but deliberately exclude remote block 0 from direct reads at target-prediction rows: its first 240 prompt keys are masked there, not kept as an always-visible sink. Earlier prefix rows remain dense, so this does not erase all indirect influence of those tokens through other prefix representations. Their contiguous library partitions blocks 1–15 as $\{1{:}5\}$, $\{6{:}10\}$, and $\{11{:}15\}$; the interleaved library uses $\{1, 4, 7, 10, 13\}$, $\{2, 5, 8, 11, 14\}$, and $\{3, 6, 9, 12, 15\}$. Every action therefore retains the same number of remote keys. The final 256 prompt tokens and causally available target tokens remain readable in every cell.

These complete-surface audits require finite observations and exact dense replay around the intervention run. The controlled and natural-text experiments use teacher-forced token NLL as the primary utility endpoint. The natural-text study also stores dense-to-sparse token KL and top-1 agreement as fidelity diagnostics. Route actions apply to frozen target-prediction rows while target suffix keys remain available according to each registered operator.

Model and numerical bindings.

Study	Model binding(s)	Additional binding
Controlled	Qwen3-4B-Instruct-2507 `cdbee75f17c0`	OLMo-3-1025-7B `a81bae42db39` (co-primary)
Fresh order/closure	Qwen (same); Llama-3.1-8B-Instruct `0e9e39f249a1`	OLMo-3-7B-Instruct (order only; secondary)
Quest native swap	Llama-3.1-8B-Instruct `0e9e39f249a1`	Quest `01c1623bf939`
Seer scope	Qwen3-4B `1cfa9a720891`	SeerAttention-R gate `d29752cb96b6`

Full repository IDs and revision hashes are frozen in the manifests. Controlled and fresh runs use frozen weights, SDPA, and bfloat16 execution; OLMo-Instruct cannot rescue or veto the fresh-order primaries and is not part of the closure study. The 40 closure identities share a document and boundary character across tokenizers, but each model uses its native 32-token continuation; identical cross-tokenizer target spans are not claimed. The Seer primary endpoint uses an FP32 LM-head readout, with native bfloat16 secondary. Its correctness-only PyTorch reference carries no official-kernel or latency-equivalence claim.

Seer's four binary factors are anchor–follower clusters, not a partition of all layers: $4 \to \{5, 6\}$, $13 \to \{14, 15\}$, $22 \to \{23, 24\}$, and $31 \to \{32, 33\}$. Full (zero) computes each follower's native mask; Shared (one) reuses its anchor's mask at the same token and KV head. Anchor masks are recomputed for each token, rather than carried over from a previous token; other layers retain native selection. Only the mask is borrowed: each follower uses its own Q/K/V, which may change through earlier interventions. The pinned Seer development manifest gives the complete operator and trajectory contract.

The Quest reference follows the released Python evaluation path: GQA K/V are expanded to query heads, signed page-max scores select top-K over all pages, and no current page is forced outside top-K. Action one exchanges the lowest-scoring selected page with the highest-scoring unselected page in an enabled depth band. The native set is fixed before deterministic boundary-tie handling; K, Q/K/V ownership, page size 16, and token budget 2,048 remain unchanged. Its Python reference likewise supports no optimized-kernel claim.

Exact decomposition. For the finite product measure, component functions are reconstructed by iterated marginalization. Numerical tests verify that the sum of the constant and all nonempty subset components reproduces every cell and that summed component energies reproduce the centered surface energy to floating-point precision. Order fractions are computed from direct squared fANOVA energy; signed temporal coherence is never used as a substitute for raw magnitude.

Energy-fraction aggregation. Let $e_{k,t} = \sum_{|S|=k} \|f_{S,t}\|_\mu^2$ and $e_{+,t} = \sum_{k=1}^{4} e_{k,t}$. The controlled study's tokenwise degree-three-plus fraction is

$$\frac{1}{T} \sum_t \frac{e_{3,t} + e_{4,t}}{e_{+,t}},$$

averaged across queries within each model. The closure instead forms

$$q_k = \frac{\sum_t e_{k,t}}{\sum_t e_{+,t}}, \qquad q_{2+} = q_2 + q_3 + q_4, \qquad s_2 = \frac{q_2}{q_{2+}}$$

within each model/source/library cell, then averages those quantities within source and across sources. A mean of token fractions and a fraction of token-averaged energies generally differ; both registered conventions are retained. Coherence always uses the ratio of summed component energies in Equation 4, not a mean of component-wise ratios. The controlled code assigns zero to a zero-energy token fraction; the natural-text protocols invalidate nonpositive required retention ratios rather than dropping affected cells.

The finite-horizon decay statistic is the negative ordinary-least-squares slope of log retained-energy ratio on log horizon at horizons $4, 8, 16, 32$. Any common logarithm base gives the same slope. The order study's contrast is $(\alpha_2 + \alpha_3 + \alpha_4)/3 - \alpha_1$; the pairwise closure's is $\alpha_2 - \alpha_1$. The fresh-order confirmation averages across the two route libraries and two endpoints within query before family-balanced query bootstrap resampling, with 10,000 frozen replicates. The pairwise closure treats source identity as the independent unit, averages the two fixed models within source, stratifies resampling by PG-19 and GovReport, and uses 50,000 frozen replicates with shared resample indices across all four components. Magnitude and composition use gold NLL; temporal and decision components equally weight gold NLL and dense KL.

Decision analysis. At each registered horizon, the unary and degree-two projections select their own minimum projected horizon-mean loss. Their actions are evaluated using the unprojected exact surface, and regret is measured against the exact optimum. All coefficient refreshes are retrospective and uncharged only because this is a measurement analysis. Any future scheduler must charge feature collection, prediction, route construction, and sparse execution separately. The closure's 320 model–source–library–endpoint choice comparisons are descriptive rows nested within 40 source units, not independent observations.

B.2 A PORTABLE LOCAL-MODEL EXECUTION EXAMPLE

The optional `examples/portable_audit` package lets a reader supply local Hugging Face safetensors weights and UTF-8 text, execute a small complete four-band mask lattice, and obtain token NLL, dense-reference KL, exact finite-product projections and hindsight choice statistics. It uses the existing query/key-mask adapter bodies; model equations and attention are implemented by the pinned Transformers runtime, not reimplemented by the audit. The separate package does not import private research paths, manifests or licensed corpus text, and does not download model weights.

The qualified example uses Torch 2.12.1 and Transformers 5.12.1, native Llama/Qwen3 full-attention classes, eager attention, one sequence and no cache. Defaults are 128 prefix tokens, 8 teacher-forced targets, 16 remote keys plus 16 recent prefix keys, and two choices in each of four equal depth bands. Prefix rows remain dense; target-row key access changes while positions and gold continuation are fixed. The complete 16-setting lattice plus six control forwards checks repeats, all-keys bypass and materialized no-ops, mask budgets and restoration. Eager kernels still compute dense attention matrices: this is not sparse-kernel acceleration or a replication of the paper's 4K protocol.

From the example directory, after installing its optional model dependencies:

```
python -m portable_audit --model SNAPSHOT \
  --text-file example_text.txt --output RUN \
  --device cuda --dtype bfloat16
```

The package passes 36 CPU tests, including isolated-directory execution with network access blocked and actual tiny native HF models. A separate GB10 smoke runs the pinned Qwen3-4B-Instruct-2507 and Llama-3.1-8B-Instruct checkpoints on the same supplied example text: 22 forwards each, all checks pass with zero observed checked-logit discrepancy. This is engineering qualification, not two independent scientific documents or a new performance claim. Output records raw numerical endpoints, token alignment, selected supports, checks, code/config provenance and measured execution counts. Token IDs may reveal input text; users should not publish private inputs through the output record.

C Evidence Status and Scope

Experiments specified before outcomes are distinguished from opened reconstructions and sensitivities on existing surfaces. The latter are not additional independent replications. An earlier direct-energy correction placed pure order four at 0.175–0.290%, versus 14.78–18.22% for all non-additive energy. The current claims use that corrected decomposition. Original contracts, technical amendments and failure records are retained in the artifact; these local records are not third-party preregistration timestamps.

C.1 Numerical and intervention boundaries

Degree-three and degree-four components are small and precision-sensitive, so the correction rests on directly reconstructed pairwise energy and its separately confirmed consequences. Fixed product measures and weak routes make the energy fractions properties of the tested geometry, not deployed route frequency or a quality–cost frontier; matched action counts do not establish matched latency. Binary geometry explains part, but not all, of Seer's smaller and more heterogeneous temporal contrast. Any new router study must therefore fix its population and coordinate contract before observing outcomes.

C.2 Chronology and endpoint-composition controls

On earlier natural-document surfaces, permuting stored token surfaces retained an interaction-order contrast. A separate sixteen-source study confirmed a normalized signed-composition contrast but not its specified chronology effects. These are not new shuffled curves for either 40-source study, and they do not identify a serial-order mechanism. The endpoint-null comparison below separately examines whether nonlinear loss applied to additive logits accounts for pairwise field structure and decision gain.

C.2.1 Opened pairwise endpoint-null decision comparison

Population and readout map. We use sixteen earlier natural-document sources (eight PG-19 and eight GovReport), Qwen/Llama and both ternary libraries: 64 complete $3^4 \times 32$ surfaces. These are sixteen source identities, not 64 independent samples or a new P293/Quest confirmation. Stored BF16 hidden tensors are replayed through the original BF16 LM heads with the original two-route, 33-row batch geometry; the first 32 prediction rows are evaluated. The resulting measured logits are projected onto the full-lattice constant-plus-unary space in FP64. Exact NLL and dense-reference KL are then applied, without re-quantizing the null logits. Original endpoints and O2 energy/inner-product summaries reproduce. No new backbone inference is used.

Shared-unary comparison. Let F be the actual endpoint, z the measured logit field, and $F_{\mathrm{null}} = \ell(P_{\le 1} z)$ the nonlinear endpoint of additive logits. Set $U = P_{\le 1} F$, $A_2 = P_2 F$, and $N_2 = P_2 F_{\mathrm{null}}$. We compare U, $U + A_2$, $U + N_2$, and the residual-only diagnostic $U + (A_2 - N_2)$. Each minimizes its mean32 prediction over all 81 actions; all choices are evaluated on the original actual mean loss. The same 10^{-12} lexicographic tie tolerance is used for all arms. Every arm is retrospective and uses complete-surface information; direct raw selection would know the optimum.

Field and decision metrics. For each source, concatenate its two libraries' component fields, giving them equal weight. Define

$$R_0^2 = 1 - \frac{\|A_2 - N_2\|^2}{\|A_2\|^2}, \qquad \cos = \frac{\langle A_2, N_2 \rangle}{\|A_2\| \|N_2\|}, \qquad \mathrm{NRMSE} = \frac{\|A_2 - N_2\|}{\|A_2\|}.$$

Compute these before and after mean32 aggregation, then average across the sixteen sources. These are field-fit descriptors, not learned-regressor validation scores. Negative R_0^2 is retained. Decision recovery is $(r_U - r_{U+N_2})/(r_U - r_{U+A_2})$ using mean raw regrets, not a mean of per-source ratios. No new interval, gate or favorable-source restriction is introduced in this opened analysis.

Actual versus null pairwise substitution gives lower/higher/tied source-mean regret on 11/5/0 (Qwen NLL), 10/6/0 (Qwen KL), 9/2/5 (Llama NLL), and 11/1/4 (Llama KL) sources, with tolerance 10^{-12}.

Model	Loss	Token R_0^2	Mean R_0^2	Mean cosine	Mean NRMSE	Gain recovery
Qwen	NLL	0.230	0.351	0.645	0.772	-1.9%
Qwen	KL	0.406	0.497	0.719	0.668	21.6%
Llama	NLL	0.042	-0.214	0.419	1.071	-4.6%
Llama	KL	0.069	-0.060	0.452	1.010	7.5%

Table 15: Partial field reproduction versus decision recovery on the sixteen opened sources. "Mean" columns apply to the mean32 component field; entries average source-level ratios after within-source library pooling. Gain recovery uses mean regrets from Table 16. Neither R_0^2 nor recovery is a causal explained fraction.

Table 16: Six opened choice controls, scored by actual mean loss. Entries are mean regret (optimum hits out of 32: sixteen sources times two libraries, not independent trials). The first four arms share actual unary terms; the last two optimize the null directly. Means are descriptive.

	Qwen		Llama	
Choice surface	NLL	KL	NLL	KL
U	0.08985 (11)	0.09702 (9)	0.03805 (16)	0.06523 (14)
$U + A_2$	0.02038 (18)	0.02345 (14)	0.00731 (22)	0.00695 (24)
$U + N_2$	0.09115 (8)	0.08114 (8)	0.03945 (13)	0.06085 (12)
$U + (A_2 - N_2)$	0.02614 (20)	0.03730 (15)	0.04475 (18)	0.04458 (17)
F_{null}	0.05112 (11)	0.02374 (16)	0.03428 (12)	0.05581 (14)
$P_{\leq 2}F_{\text{null}}$	0.05902 (11)	0.03834 (13)	0.03459 (13)	0.05523 (15)

For Qwen, the null partially reproduces both tokenwise and mean32 O2 fields, but recovers only $-1.9\%/21.6\%$ of the NLL/KL decision improvement. The remaining field error is substantial: this is not near-exact field recovery followed by a large decision failure. For Llama, field reproduction is weak and mean32 R_0^2 is negative, so its poor decision recovery is not the same field-versus-choice contrast. Across models, this shared-actual-unary hybrid does not reproduce most mean gain; it is not uniformly inferior on every source. All family/source breakdowns are retained in the artifact.

Residual-only choices retain much of Qwen's improvement but perform poorly for Llama; Llama NLL regret even exceeds unary. Thus the useful contribution cannot simply be assigned to an isolated residual. Although $A_2 = N_2 + (A_2 - N_2)$, these fields are not orthogonal and argmin-based regret benefits do not add. The shared-unary comparison limits this pair-field substitution, not direct optimization of the entire null or all curvature explanations; it identifies neither the originating component nor a mechanism for temporal cancellation.

The null also obeys a useful consistency identity. For the same reference distribution p_d and gold token y,

$$\text{NLL}(a) - \text{NLL}(d) - D_{\text{KL}}(p_d \| p_a) = (p_d - e_y)^\top (z(a) - z_d).$$

The right-hand side is linear in logits, so its O2 vanishes for additive logits: $P_2\text{NLL}_{\text{null}} = P_2\text{KL}_{\text{null}}$. This is an exact endpoint identity, not a new theorem or evidence that actual KL is an endpoint-only null.

Direct-null choice and cross-order orthogonality. An additional opened CPU control selects the minimizers of F_{null} and $P_{\leq 2}F_{\text{null}}$, then scores both on actual F. All 16 sources, two libraries and both endpoints remain included; libraries average within source. No coefficient is tuned and no primary criterion changes.

Direct-null mean regret nearly matches actual pairwise for Qwen KL (0.02374 versus 0.02345), recovering 99.60% of the mean gain over unary, with 7/7/2 better/worse/tied source means. This is not an equivalence test. Recovery is 55.75% for Qwen NLL and $12.28\%/16.15\%$ for Llama NLL/KL. Degree-two null recovery is $44.38\%/79.76\%$ and $11.27\%/17.16\%$, respectively. These are descriptive ratios, not causal shares. The null's decision adequacy cannot be judged from its pair field alone; neither a universal curvature explanation nor its universal failure follows. All unfavorable source/corpus cases are released.

For the same product measure, $\langle U, N_2 \rangle = \langle F, P_{\leq 1} P_2 F_{\text{null}} \rangle = 0$, even though the parent functions differ. This identity holds tokenwise and after averaging. In contrast, N_2 and $A_2 - N_2$ occupy the same degree-two space and need not be orthogonal. Nonadditive argmin gains alone prove neither nonorthogonality nor an internal mechanism. Independent conditional-marginal and tensor-basis implementations reproduce all 768 old/new choices; this uses saved endpoints, not new model calls.

The standalone package includes the 64 saved actual/null endpoint arrays. `artifact.endpoint_bridge` regenerates the four shared-unary arms (512 arm/endpoint choices); `artifact.pat_endpoint` adds the two direct-null arms, giving 768 choices across all six arms. A separate tensor-basis check reproduces the original marginal-projection results. The package does not regenerate hidden tensors or the head replay; older chronology/O4 result summaries remain provenance, not rerun model experiments.

D LOCATION REUSE AND RESTRICTED-OBSERVATION FOLLOW-UPS

The original 40-source study serves as development data here. A further 180 document identities (90 PG-19, 90 GovReport), selected using the historical exclusion registry, were evaluated with the same checkpoints, four bands, two ternary libraries, 4096-token prefix and 32-token teacher-forced endpoint. The location test was frozen before these new outcomes. Geometry and acquisition analyses were proposed after they were opened: they are not additional fresh confirmations or explanations of temporal preservation. The registry omitted four older document allocations; Appendix G reports this defect and the 176-source exclusion sensitivity without changing the original result or its unmet joint criterion.

D.1 A PRESPECIFIED FIXED-LOCATION RETENTION TEST

Development fixed bands 2–3 for Qwen and 1–2 for Llama (one-based). All 81 actions are observed on each new source to compute its unary U_x and pair fields $\Psi_{x,ij}$. We evaluate actual-loss regret of $\arg\min_a[U_x(a) + \Psi_{x,e_m}(a)]$: only pair identity e_m transfers. Controls average separately selected actions' outcomes over all six pairs or the three adjacent pairs, not averaged fields. Development-best-adjacent matches the fixed rule; development-energy selection also matches for Llama.

The registered criterion requires positive lower endpoints for all four source-paired 95% intervals and positive model/endpoint point contrasts. Libraries and NLL/KL are averaged within source; 50,000 family-stratified bootstrap draws retain all 180 sources in their original order. These are marginal intervals, not a simultaneous confidence set.

Table 17: Registered location retention. Gain is control-minus-fixed regret; $+/-/=$ counts source directions. Qwen's adjacent interval contains zero, leaving the joint criterion unmet without proving its effect zero.

Model	Control	Gain	Marginal 95% CI	$+/-/=$
Qwen	Random pair	0.00602	$[0.00239, 0.00975]$	113/64/3
Qwen	Adjacent pair	0.00274	$[-0.00030, 0.00587]$	101/72/7
Llama	Random pair	0.02200	$[0.01596, 0.02869]$	124/47/9
Llama	Adjacent pair	0.01749	$[0.01198, 0.02360]$	113/51/16

Pooled retention of 52.9% of full-pairwise versus unary regret improvement is descriptive, not a replacement for the unmet criterion. This is selective retention after full observation, not cheaper measurement, coefficient transfer or architectural causation.

D.2 OPENED TRANSFER OF A LOW-DIMENSIONAL REPRESENTATION

Pure O2 has 24 dimensions: six centered 3×3 pair fields, four degrees of freedom each. Separately by model/library/endpoint, an origin-through rank-four SVD basis fits unit-normalized development O2 vectors. The basis is frozen on the 180 opened sources, but target projection coordinates and the common unary use each complete target surface.

Let $M \in \{0,1\}^{3\times 16}$ encode the three disjoint five-block masks and $C = I_3 - \mathbf{1}\mathbf{1}^\top/3$. Their centered support Gram is $G = CMM^\top C = 5C$, proportional to $\mathbf{1}[a_i = a_j] - 1/3$ on each pair. The support control fits four axes within the span of these six pair directions. Other rank-four controls use the fixed-pair space or block-diagonal development covariance. We also average 128 random-space and 128 refitted source-relabelled outcomes. Applying the same action permutation to all bands within each training source preserves pair energy, singular values, within-source alignment and G.

Table 18: Opened representation transfer. Capture is mean target O2 energy retained; regret uses actual mean loss. Rank-four methods have equal O2 dimension; unary and full O2 are references.

Representation	Qwen		Llama	
	Capture	Regret	Capture	Regret
Learned rank-4	48.8%	0.03248	54.3%	0.03198
Fixed-pair rank-4	29.3%	0.03318	58.2%	0.03127
Support-Gram rank-4	42.1%	0.03091	34.0%	0.03162
Relabelled training mean	46.3%	0.03223	54.3%	0.03048
Block-covariance rank-4	50.1%	0.03111	53.8%	0.02623
Random rank-4 mean	16.6%	0.04095	16.8%	0.05303
Unary only	0.0%	0.04545	0.0%	0.06594
All actual O2 (rank-24)	100.0%	0.01760	100.0%	0.00511

Learned rank four improves over random spaces and unary, but its regret intervals against fixed-pair, support-Gram and relabelled controls all include zero. Fixed-pair comparisons also have opposite NLL/KL point signs within each model. This neither establishes distinctive learned geometry nor proves its absence. Other predefined ranks and normalizations remain in the artifact.

D.3 Opened selection with restricted observations

Selectors see queried scalar mean losses, not target exact unary/O2. Nine constant/unary and four frozen O2 columns give 13 parameters. Pivoted QR followed by leverage selection fixes an observation design. At budget $b \in \{17, 25\}$, $b - 1$ losses fit coefficients by least squares. Evaluating the recommendation consumes the final query; if already observed, the next unused design point is queried. Return the best observed action.

Comparators use fixed-pair, support-Gram or unary designs, plus direct best-measured on each b-point design. Thus unary comparisons concern complete acquisition procedures, not identical-label regression alone. Cyclic search uses either fixed 0123 order or an order chosen among 24 on development only. Frozen random completion fills the budget after convergence; repeated queries are cached. No order is chosen from evaluation outcomes.

Table 19: Opened query-budget comparison on the same 180 sources. Each row charges all scalar observations, including the final recommendation. Regret is not generated-answer error or serving latency.

Queries	Selector	Qwen regret	Llama regret
17	Learned rank-4	0.02040	0.02777
17	Fixed-pair rank-4	0.01958	0.02928
17	Unary fit	0.02482	0.03906
17	Shared design, best measured	0.04241	0.08633
17	Cyclic (0123) + completion	0.00541	0.01078
17	Cyclic (dev. order) + completion	0.00469	0.01356
25	Learned rank-4	0.01503	0.01789
25	Fixed-pair rank-4	0.01319	0.01512
25	Unary fit	0.02097	0.02793
25	Shared design, best measured	0.03057	0.05638
25	Cyclic (0123) + completion	0.00271	0.00628
25	Cyclic (dev. order) + completion	0.00324	0.00823

At 17 queries, learned-rank-4 gains over development-order cyclic are -0.01572 $[-0.02067, -0.01169]$ (Qwen) and -0.01420 $[-0.01968, -0.00869]$ (Llama). Both endpoints favor direct search; fixed-order cyclic also wins, and the direction persists at 25. Surrogate recommendations improve over static probing and unary acquisition, not over the stronger controls. Intervals use 20,000 family-stratified source draws and are opened marginal 95% summaries. In this lattice, fitting a reusable surrogate can aid proposals without outperforming direct conditional search.

Table 20: Opened query-budget comparison: actual mean NLL at each method's selected observed action (lower is better), using the same four sources within each row. U/P/H denote unary, pairwise and fixed half; Rev./Fwd. denote reverse/forward search. Each method receives at most 130 distinct observations, including cross-method cache returns. All source-level outcomes and supplemental dense KL at the NLL-selected action are retained in the replay. No confidence intervals, confirmation, global-regret or speedup claim.

Model	Library	U	P	H	Rev.	Fwd.	Random
Qwen	Contiguous	2.785716	2.785551	2.783689	2.751026	2.751026	2.854990
Qwen	Interleaved	2.911223	2.913021	2.907402	2.856723	2.834906	2.924027
Llama	Contiguous	5.201825	5.149137	5.154819	5.123496	5.123496	5.254862
Llama	Interleaved	5.225529	5.194028	5.180911	5.173995	5.172447	5.265882

D.4 Opened eight-band live query-budget pilot

The preceding comparison uses four-band archived complete surfaces. A separate pilot instead executes real queries in the $3^8 = 6561$-configuration domain on four already-opened C2 documents (two per corpus, selected by identity-hash order), with both checkpoints and both libraries. It tests per-document selection, not transfer of one static configuration between documents. Methods, sources and budgets were fixed before inspecting the calibration- expansion results in Appendix L.1.

Each method receives at most 130 distinct observations. Original U, P and fixed half use the same 129-action training design and frozen penalties, score all 6561 configurations, and validate one nomination if unobserved. If already measured, no substitute query is supplied. Forward/reverse cyclic searches use fixed coordinate orders and seeded restarts; random search uses a fixed permutation. Each spends 130 distinct observations. All methods return their best measured action, with the same 10^{-12} choice-tie convention. Methods have private information ledgers: sharing a physical execution cache does not waive another method's observation charge. No unqueried true loss or exact fANOVA coefficient is available to a method.

Both direct-search orders have lower average selected NLL than each surrogate procedure in all four cells. Pairwise and half improve over unary in several cells, but do not establish an equal-budget advantage over direct search. These outcomes do not test the new block-ridge rule or 257-observation fits, and do not rule out all structural acquisition methods.

There are 12,479 charged method observations and 7,566 distinct physical science queries, with 4,913 cross-method cache hits still charged. Duplicate- batch execution equality is checked on every physical query. Including qualification/repeat/restoration checks, 8,038 physical forward calls are recorded. Two H100 workers finish collection and backup in about 50 minutes; this shared-cache, check-heavy runtime is not a per-method production speed benchmark. Saved query values and all six decision traces are independently replayed. Dense-reference KL at the NLL-selected action is supplemental. Without a complete 6561-action surface, global regret is unknown. This is opened cost evidence, not a new confirmation, online router or generation test.

E AGGREGATE FIT VERSUS CHOICE VALIDITY

This opened reanalysis uses all 320 model–library–endpoint comparisons nested in the same 40 sources; it adds neither model execution nor fresh confirmation. For each continuation-mean surface $\bar{F}$ and its uniform full-cube unary projection U, define

$$R^2_U = 1 - \frac{\sum_a (\bar{F}(a) - U(a))^2}{\sum_a (\bar{F}(a) - \mathbb{E}_a \bar{F})^2}.$$

The median is 0.9134. A post-hoc descriptive threshold of 0.95 selects 91 comparisons, with 46 unary and 34 pairwise optimum misses. Within that subset, mean actual-loss regret is 0.03067 for unary and 0.01222 for unary plus O2. Only one comparison has $R^2_U \geq 0.99$: these data do not establish frequent choice failures at arbitrarily near-perfect fit. Threshold counts are descriptive and are not independent source trials or a newly registered definition of adequate fit.

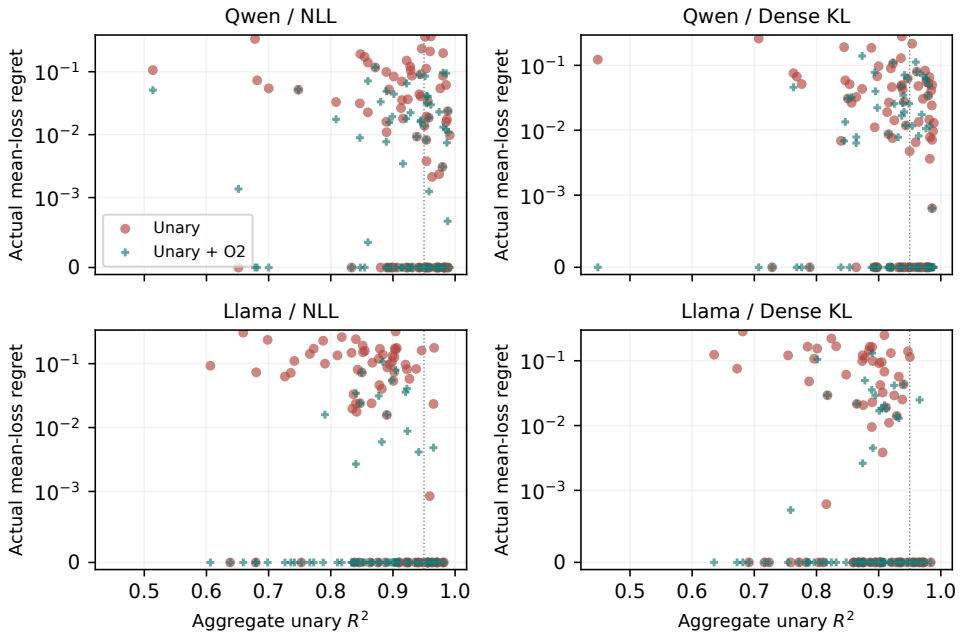

Figure 5: Every P293 comparison, separated by model and endpoint (80 per panel, sharing 40 sources). Both choices are evaluated on the original actual mean-loss surface. Regret uses a symmetric-log axis with a linear region below 10^{-3}, retaining exact zeros. The dotted $R^2 = 0.95$ line is a descriptive reference, not a validated selection rule. Global fit and argmin validity answer different questions; this plot does not establish a deployed workflow's error or a query-efficient method.

The package recomputes every projection, fit score, selected action and regret from the saved token arrays. An independent conditional-marginal calculation agrees with the tensor-basis projections to floating-point precision. The 0.95 subgroup is not used to alter any original action, fit, hypothesis test or source inclusion rule.

An illustrative Qwen/PG-19 contiguous-library KL case has $R^2_U = 0.98461$: unary selects $(0, 0, 1, 1)$ with regret 0.04123, whereas O2 selects the exact optimum $(0, 2, 1, 1)$. This example is chosen nearest the median unary regret among high-fit errors improved by O2. It is outcome-conditioned, not a randomly selected or representative population estimate.

What a margin can and cannot explain. Let $\bar{F} = U + R$, and let a_U minimize U. A competitor a beats it exactly when

$$R(a_U) - R(a) > U(a) - U(a_U).$$

Thus a small average squared residual can exceed a particular competing margin in the relevant direction. This identity is not a new theorem or an identified neural mechanism. With $m_U =$

$\min_{a \neq a_U}[U(a) - U(a_U)]$, the sufficient condition $\mathrm{osc}(R) < m_U$, where $\mathrm{osc}(R) = \max_a R(a) - \min_a R(a)$, guarantees the same optimum; more generally $r_U \leq \max(0, \mathrm{osc}(R) - m_U)$.

The full-residual diagnostic is conservative here: it certifies none of 320 unary choices, despite 120 being correct, and only 12 degree-two choices. Within the 91 high-fit comparisons the corresponding counts are 0 and 4, and the median unary residual-oscillation/margin ratio is 13.84. The numerical certificate requires a strict 10^{-12} margin above residual oscillation; all formulas are independently checked on conditional-marginal projections. These descriptive counts are nested within the original 40 sources and use the full measured surface. Failure to certify is not failure of the choice. An exact directional crossing diagnoses an observed error but is not a pre-action predictor; a maximum measured at a few probes is not a bound on every unobserved candidate without additional assumptions.

Opened candidate-localized certificate. Using the same complete surfaces, let $w = \mathrm{osc}(R)$ and $C = \{a : U(a) - U(a_U) \leq w\}$. Outside C, the known global bound excludes a better actual loss. Inside C, replace the residual range by $\mathrm{osc}_C(R)$ and compare it with the minimum surrogate gap over $C \setminus \{a_U\}$; a singleton is vacuously certified. Apply the same construction to the degree-two approximation, without changing any choice. At the existing 10^{-12} strict margin, certified unary choices rise from 0 to 5 of 320; degree-two certificates rise from 12 to 79. All are correct. Median candidate counts are 22 and 6, but 115 correct unary and 128 correct degree-two choices remain uncertified. Localization helps, yet does not show that irrelevant outliers uniquely cause the global certificate's conservatism. It still uses the full measured residual: it is not a cheaper observable gate. An independent directional-crossing check agrees with actual regret on all 640 approximation/surface rows, with no incorrect certificates.

F QUEST ARITHMETIC SENSITIVITY

We perform an opened sensitivity audit on eight existing Quest sources, selecting four per corpus by the SHA256 ordering of source identifiers, without consulting their effects. All sixteen same-K boundary-swap actions and the original token windows are retained. This is not a new forty-source confirmation or a test of generated quality.

Two changes are separated. First, we evaluate FP32 and FP64 LM-head and KL arithmetic on *identical* hidden states and supports. Second, we cast the same stored checkpoint weights to FP32 and rerun the backbone, with TF32 disabled. This does not recover unquantized training weights. Each backbone uses its *own* native Quest distribution as the KL reference; the higher-precision backbone may change queries, keys and selected pages. It is not a same-QKV or fixed-support numerical comparison.

Backbone	Readout	$\Delta\rho$	r_U (10^{-4})	$r_{\leq 2}$ (10^{-4})	Matches U→O2
BF16	FP32	0.04114	0.8222	0.2625	$16 \rightarrow 18$
BF16	FP64	0.04114	0.8222	0.2625	$16 \rightarrow 18$
FP32	FP32	0.03621	2.0781	0.1554	$12 \rightarrow 20$
FP32	FP64	0.03621	2.0781	0.1554	$12 \rightarrow 20$

Table 21: Opened eight-source arithmetic sensitivity. Values are source means, with regret also averaging the three Hamming sets within source. Optimum matches are out of 24 nested choices, not independent trials. $\Delta\rho$ is unary-minus-pairwise temporal coherence. FP64 here denotes readout arithmetic, not a full-FP64 model.

The original BF16 endpoint surfaces, surface-hidden digests and native support digests match in all eight sources. Changing only readout arithmetic changes none of the 24 oracle, unary or pairwise choices within either backbone. Rerunning the backbone in FP32 changes 9 oracle, 8 unary and 10 pairwise choices; none of the 128 complete action-support trace digests is byte-identical across backbones. This does not mean every individual page changes. Both mean audit directions persist in this subset, but exact route identity is execution-dependent. Positive coherence contrasts occur in 8/8 BF16 and 6/8 FP32 sources; positive regret gains occur in 3/8 and 5/8, respectively. These descriptive counts do not establish a population-wide precision-invariance result.

Every completed execution checks native/action-zero equivalence, native and joint-action replay from an unchanged base cache, and exact same-K boundary swaps. A separate CPU log-softmax calculation checks each readout, and an independent Walsh implementation checks the resulting fANOVA and regrets. The artifact reproduces the table from the saved endpoint arrays. It does not regenerate the backbone, KV caches or hidden-state generation.

F.1 OPENED EIGHT-BAND FEASIBLE-SET AND EXECUTION CONTROLS

A separate development pilot uses eight existing documents and eight binary depth bands. Fit both unary and degree-two projections uniformly on the SAME 70 weight-four configurations: their ranks are 8 and 28, so neither is saturated. All labels are observed before retrospective choice. The original GB10 pilot does not establish an incremental mean choice gain over feasible-set unary.

The same eight documents are subsequently replayed under a qualified deterministic H100 execution. Both use the same pinned Llama-3.1-8B-Instruct weights, BF16/SDPA backbone and FP32 head/KL readout, unlike the preceding backbone-precision intervention. An initial strict bridge failed; those failures and original GB10 outcomes remain recorded. The deterministic profile passes repeats, no-op, cache, own-KV/support, coarse/fine and restoration checks, but does not match GB10 surfaces. Table 22 reports both results, not a replacement of the uncertain pilot by the favorable execution.

Table 22: Opened SAME70 native-reference KL choice. Eight document units, not 70 independent samples per source. Intervals are nominal paired source intervals; the H100-minus-GB10 contrast uses the same documents.

	SAME70: values $\times 10^{-4}$			
Execution	r_U	r_P	Gain	Nominal 95% CI
Original GB10	5.17786	4.33785	+0.84000	[-1.04336, +3.18894]
Deterministic H100	3.16710	0.92958	+2.23752	[+0.08589, +5.06151]
Paired H100–GB10	–	–	+1.39752	[-2.34919, +5.65918]

Across executions, configuration labels change in 7/8 unary choices, 6/8 pairwise choices and 5/8 actual optima. Each execution has its own native reference and query-dependent physical supports, so identical labels need not mean identical supports. The positive opened H100 interval is not fresh confirmation or evidence of a positive population hardware difference: the direct paired gain-difference interval includes zero. These results qualify native choice claims, not the separately measured fixed-support prediction or preservation objects. Appendix P separately reports the subsequent 160-document confirmation under the frozen H100 profile; it does not upgrade these opened results or establish cross-execution invariance.

A subsequent opened Welch sensitivity uses the same paired source gains and equal-corpus weights, not a new success rule. Its GB10, H100 and paired H100-minus-GB10 intervals are respectively $[-3.1003, 4.7803]$, $[-1.4346, 5.9096]$, and $[-5.1095, 7.9045]$, all $\times 10^{-4}$. All include zero, with approximate degrees of freedom 3.16, 5.63, and 4.22. Both interval procedures rely on only four sources per corpus; neither guarantees population coverage. The original bootstrap result remains reported, but the H100 pilot is inferentially sensitive, not robust new confirmation.

The original 40-action, equal-changed-layer sensitivity and cross-scoring of configuration labels are also retained in the package. Neither is used to replace the 70-action comparison. The software verifies saved numerical surfaces and paired statistics; it does not rerun either backbone or identify an internal cause of the execution differences.

G HISTORICAL ALLOCATION COVERAGE AND EXCLUSION SENSITIVITY

A subsequent provenance audit found that the historical exclusion registry covered JSONL files but omitted an older JSON source manifest. Four PG-19 documents in the 180-source location/representation population had prior allocations: two had been evaluated and two were protected reserves with no recorded outcome forwards in that earlier study. One PG-19 document in the 40-source Quest population had also been evaluated earlier. Document hashes, not just identifiers, agree. This establishes prior document exposure or allocation, not identical token windows or outcome-dependent source selection. The 40-source two-model closure population does not intersect these allocations; its numerical results are unchanged.

We preserve the original protocols, arrays, selections and registered verdicts, but do not describe the affected populations as entirely unallocated. The following opened sensitivity removes all four follow-up documents and the one Quest document, without refitting a basis, changing a pair or rerunning selection. Estimates retain equal corpus weights with 86/90 PG-19/GovReport sources for the follow-ups and 19/20 for Quest. Intervals use 20,000 within-corpus source resamples; they are retrospective nominal intervals, not replacement confirmatory tests or simultaneous bands.

Table 23: Allocation-exclusion sensitivity. Follow-up gains are adjacent-minus-fixed regret; Quest contrasts compare the registered cube-unary and cube-pairwise projections. Nested choices are not independent sources.

Population	Quantity	Estimate [95% interval]
Qwen, 176 sources	Fixed-location gain	$0.00280\ [-0.00025, 0.00593]$
Llama, 176 sources	Fixed-location gain	$0.01765\ [0.01208, 0.02381]$
Quest, 39 sources	Unary-minus-pairwise coherence	$0.04096\ [0.02229, 0.06060]$
Quest, 39 sources	Unary-minus-pairwise regret	$0.0001010\ [0.0000411, 0.0001755]$

The fixed-location joint criterion remains unmet. Both random-pair comparisons retain positive interval bounds, while Qwen's adjacent comparison includes zero. Quest retains positive mean temporal and decision contrasts; cube-unary, cube-pairwise and slice-unary exact matches are respectively 75, 101 and 104 of 117 nested choices. These do not alter the choice-set or information-budget limitations of the original comparisons.

On the same 176 follow-up sources, the 17-query shared-basis surrogate has mean regrets 0.02053/0.02757 for Qwen/Llama, versus 0.00546/0.01079 for fixed-order cyclic search. At 25 queries the corresponding values are 0.01515/0.01762 versus 0.00273/0.00643. The restricted-observation competitive conclusion is therefore unchanged. The artifact supplies the exclusion identities, unchanged per-source statistics and a standalone sensitivity command.

H PROSPECTIVE SUPPORT-CONSTRUCTION INTERVENTION

This study tests a condition of temporal preservation, not a new route selector. Its primary target is the change in the normalized unary–pairwise gap at horizon 32, not interaction magnitude, generated quality, or a causal explanation of the earlier ranking results. The development, opened controls and fresh confirmation populations remain separate.

H.1 PHYSICAL CONSTRUCTION AND MATCHED QUANTITIES

The two checkpoints and revisions are Qwen3-4B-Instruct-2507 (`cdbee75f17c0`) and Llama-3.1-8B-Instruct (`0e9e39f249a1`); full revisions are in the artifact. Four bands partition their 36 and 32 layers into groups of nine and eight. Each source has a native-tokenizer 4,096-token prompt and 32 teacher-forced targets. Dense prefix formation and hard score cuts on target prediction rows follow the controlled operator. All conditions retain the same local suffix, causal keys, five 240-token eligible remote blocks per action, and per-layer K.

For each source/layout seed, permute the 15 eligible block IDs into a 5×3 array B. A band's action $a \in \{0, 1, 2\}$ reads

$$S_b(a) = \{B_{g,(a+c_{bg}) \bmod 3} : g = 0, \ldots, 4\}.$$

HIGH uses $c_b = (0, 0, 0, 0, 0)$ in every band. LOW uses $(00000), (00112), (01021), (01202)$, whose pairwise modular differences have counts $\{1, 2, 2\}$ over the five groups. These are the lexicographically first valid code triple appended to the zero code; every realized partition/seed is included in the package. Each band thus partitions the same 15 blocks into three disjoint five-block actions.

For bands i, j, let $G_{ij}(a, b) = |S_i(a) \cap S_j(b)|$ and $C = I - \mathbf{1}\mathbf{1}^\top/3$. Both conditions have mean overlap $\frac{1}{9}\sum_{a,b} G_{ij}(a, b) = 5/3$ blocks for every band pair. However, $\|CG_{ij}C\|_F^2$ equals 50 for HIGH and 2 for LOW. HIGH/LOW therefore denote pattern concentration, not larger/smaller mean overlap. The intervention changes the whole support recipe, including content grouping and depth-dependent use; it does not isolate this scalar Gram norm or hold marginal loss fixed.

Cumulative coverage is matched in mean, not routewise. Let $C(a) = |\bigcup_{b=1}^{4} S_b(a_b)|$. Every eligible block occurs in one of three actions in each band. Independent uniform actions therefore give $\mathbb{E}_\mu C = 15[1 - (2/3)^4] = 325/27$ for *both* recipes, at every physical layout. Prefix-depth means likewise equal $15[1 - (2/3)^j]$ for $j = 1, \ldots, 4$. Thus the intervention does not change average direct remote coverage. It does change its distribution: Shared has counts $\{5 : 3, 10 : 42, 15 : 36\}$ over 81 actions; Shifted has $\{10 : 3, 11 : 12, 12 : 48, 13 : 15, 14 : 3\}$. Only three Shared schedules read five distinct blocks; shared partitions do not force all bands to choose one action. Variance and content/reuse patterns remain coupled aspects of the recipe, not isolated causes. Direct support union is also not the total causal receptive field, since prefix states are formed densely.

An exact geometry-only replay checks all 81 actions and frozen layouts without using loss outcomes. Selecting equal-coverage actions post hoc would change the product measure and its decomposition, not repair the original experiment.

Earlier overlap-concentration manipulations in the research record already varied support geometry at matched budgets. Their stored mean-NLL surfaces cannot recover temporal retention. The new question here is the prospective direction of the *order-differential temporal response*, not invention of an overlap manipulation.

H.2 DEVELOPMENT, ALLOCATION AND FROZEN INFERENCE

Eight previously opened documents, four per corpus, each with four seeded layouts supplied development evidence. Their Qwen mean Γ estimates were $0.01291/0.01762$ (NLL/KL). These informed the positive direction and Qwen co-primary designation before any confirmation output. They are not additional confirmation units. Llama was secondary from the outset.

The confirmation has 80 new document identities, 40 each from GovReport and PG-19, and four independently seeded layouts per document, paired across checkpoints. An expanded JSON/JSONL allocation inventory excludes previously evaluated, allocated and protected identities, raw/normalized document hashes, and public PG-19 split identities. The earlier coverage defect in Appendix G is not

inherited. Fresh means disjoint experimental allocation, not absence from pretraining or new-domain transfer. No document, layout, endpoint, or sample size was changed after outcomes.

For each model, document, layout and endpoint, compute $\rho_k = E_k^{\mathrm{agg}}/E_k^{\mathrm{tok}}$ from the complete 81×32 surface. Form $\Delta\rho_k = \rho_{k,\mathrm{HIGH}} - \rho_{k,\mathrm{LOW}}$ and $\Gamma = \Delta\rho_2 - \Delta\rho_1$, then average the four contrasts within document. Ratios precede layout averaging; neither pooled-energy ratios nor layout-level pseudoreplication are substituted.

Let $\bar{\Gamma}_f, s_f^2, n_f$ denote the sample mean, variance and size of corpus f. The equally weighted mean and its variance estimate are

$$\widehat{\Gamma} = \tfrac{1}{2}\sum_f \bar{\Gamma}_f, \quad v_f = \frac{s_f^2}{4n_f}, \quad V = \sum_f v_f, \quad \nu = \frac{V^2}{\sum_f v_f^2/(n_f-1)}.$$

The prespecified interval is $\widehat{\Gamma} \pm t_{.975,\nu}\sqrt{V}$. The factor four in v_f comes from the squared one-half corpus weight, not from the four within-document layouts. Qwen's NLL and dense-KL lower bounds must *both* exceed zero, a conjunctive decision, not selection of a favorable endpoint. These approximate individual intervals are not a simultaneous confidence set; the two endpoints share logits. A prespecified 20,000-draw corpus-stratified document bootstrap (seed 202609143141) gives the same primary decisions and remains sensitivity, not a replacement test. There was no interim scientific inspection or optional stopping. These are local hash-bound pre-execution records, not an external preregistration service.

H.3 Effect sizes and heterogeneity

Table 24: Fresh retention means and source directions. Positive and mixed counts each have denominator 80; mixed means positive and negative layout contrasts coexist within that source. All four layouts remain included.

Model	Endpoint	LOW ρ_1/ρ_2	HIGH ρ_1/ρ_2	$\Gamma > 0$	Mixed layouts
Qwen	NLL	$0.13357/0.05299$	$0.13655/0.06189$	48/80	73/80
Qwen	KL	$0.21176/0.06951$	$0.21939/0.08374$	50/80	70/80
Llama	NLL	$0.08578/0.05558$	$0.08444/0.06357$	47/80	64/80
Llama	KL	$0.11395/0.06917$	$0.11278/0.07948$	51/80	63/80

The relative mean-gap decreases in the main text divide mean Γ by the mean LOW gap. They are not means of document-relative percentages or fractions of mechanism explained. Qwen's fresh effects are smaller than its development estimates, and all four remaining mean gaps are positive. Llama's larger relative decreases partly reflect its smaller LOW gaps: direct paired Qwen-minus-Llama Γ is -0.00340 $[-0.01094, 0.00413]$ for NLL and -0.00488 $[-0.01361, 0.00386]$ for KL. These comparisons do not establish model-response differences or architecture causality. Qwen's KL corpus means are 0.00160 (GovReport) and 0.01162 (PG-19); these descriptive strata do not establish a corpus moderator.

O2 retention alone also increases: Qwen $0.00890/0.01423$ and Llama $0.00799/0.01031$, each with a positive individual interval. Qwen unary retention changes by $0.00298/0.00763$, and Llama by $-0.00134/-0.00117$. Thus $\Delta\rho_2$ must not be substituted for the primary gap change. Source-level values, interval variants and layout contrasts are retained as profiles in the review edition.

HIGH-minus-LOW all-action mean loss is 0.00786 $[-0.00167, 0.01740]$ for Qwen NLL and 0.01039 $[0.00070, 0.02009]$ for Qwen KL; Llama's changes are 0.00449 $[-0.00100, 0.00997]$ and 0.00341 $[-0.00207, 0.00889]$. These are uniform-lattice mean losses, not selected-policy quality. They neither establish a quality gain nor prove every HIGH route is worse.

H.4 Opened explanatory controls and fresh linear readout

Table 25: Opened controls on the same eight development documents, layout zero only. Intervals are nominal source-bootstrap 95% intervals; these are not fresh confirmations or independent eight-source replications.

Model	Endpoint	Rotation-balanced Γ	Actual-minus-null Γ
Qwen	NLL	$0.02158\ [0.00031, 0.04039]$	$0.01067\ [-0.00802, 0.03730]$
Qwen	KL	$0.02447\ [0.00265, 0.04410]$	$0.02078\ [0.01126, 0.03303]$
Llama	NLL	$-0.02569\ [-0.07532, 0.00865]$	$-0.01809\ [-0.07454, 0.03051]$
Llama	KL	$-0.00281\ [-0.07360, 0.05043]$	$-0.00509\ [-0.07975, 0.06018]$

Finite depth-placement sensitivity. Cyclically rotate LOW's four partition assignments over depth, holding HIGH fixed. Compute retention per execution before averaging the four rotations. Qwen's positive mean balanced Γ remains, but falls from its original layout-zero value by 0.01758 $[0.00119, 0.04114]$ for NLL and 0.01735 $[0.00666, 0.02968]$ for KL. Llama's balanced intervals include zero. All LOW rotations preserve K, mean overlap and Gram norms. Nevertheless, HIGH uses only one partition, and cyclic four is not all 24 permutations. This finite placement contrast neither fully matches content repertoires nor decomposes the effect into independent causal contributions. The fresh trial tests the original recipe, not the rotation ensemble.

Specified endpoint null. Passively captured post-normalization hidden states reproduce native BF16 head outputs. Project the full measured logit lattice onto constant plus unary components in FP64, then apply the same nonlinear NLL/KL readout to actual and projected logits without null requantization. The Qwen KL actual-minus-null Γ is positive, whereas NLL is uncertain (Table 25). This limits the specified additive-logit explanation, not endpoint nonlinearity in general. The review edition supplies actual/null source profiles, not endpoint tensors; it does not rerun hidden-state generation or head replay.

Fresh linear-readout diagnostic. Subtract actual NLL and dense-reference KL *surfaces before* fANOVA. Their nonconstant difference is one logit-linear readout, not all logit interaction. Its fresh Γ is $-0.00118\ [-0.00401, 0.00166]$ in Qwen and $0.00013\ [-0.00253, 0.00279]$ in Llama. The original positive response is not established in this diagnostic. Retention ratios are nonlinear, so $\Gamma(F_{\mathrm{NLL}} - F_{\mathrm{KL}})$ is not $\Gamma(F_{\mathrm{NLL}}) - \Gamma(F_{\mathrm{KL}})$. This neither demonstrates absent internal interaction nor attributes the complete effect to loss curvature.

H.5 Opened horizon and energy description

After the H32 confirmation, we additionally recompute the fixed prefix set $H \in \{4, 8, 16, 32\}$ from the same endpoints, retaining all 80 documents and four layouts. Table 26 is opened description, not four new confirmations or a new test of decay-exponent differences. The H32 rows reproduce the original contrasts and intervals; the primary is unchanged.

Table 26: Mean $\Gamma(H)$ and nominal pointwise 95% Welch-t intervals at every inspected prefix horizon, using 80 document units. Shorter horizons are opened analyses; H32 repeats the original result. These are not simultaneous intervals or additional confirmation criteria.

Model	H	NLL Γ [95% CI]	KL Γ [95% CI]
Qwen	4	$0.01469\ [-0.00131, 0.03069]$	$0.01369\ [-0.00287, 0.03026]$
Qwen	8	$0.00268\ [-0.00999, 0.01535]$	$0.01469\ [0.00294, 0.02644]$
Qwen	16	$0.00727\ [-0.00007, 0.01461]$	$0.00965\ [0.00010, 0.01920]$
Qwen	32	$0.00592\ [0.00138, 0.01047]$	$0.00661\ [0.00065, 0.01256]$
Llama	4	$0.01119\ [-0.00909, 0.03146]$	$0.01367\ [-0.00723, 0.03457]$
Llama	8	$0.00997\ [-0.00242, 0.02236]$	$0.01354\ [-0.00278, 0.02986]$
Llama	16	$0.01041\ [0.00214, 0.01868]$	$0.01732\ [0.00739, 0.02726]$
Llama	32	$0.00933\ [0.00368, 0.01497]$	$0.01148\ [0.00404, 0.01893]$

All H4 intervals include zero, as do Qwen NLL's H8/H16 and Llama NLL/KL's H8 intervals. The point estimates are nonmonotone and establish neither a constant shift nor a universal horizon trend. The package also provides every order/condition's token and aggregate energy means. At H32, Qwen NLL O2 token energy rises from 0.21701 to 0.23415, while aggregate energy rises from 0.01424 to 0.01888. Both change; these averages do not apportion causal contributions to the retention response. In particular, mean retention is not the ratio of these mean energies. No shorter-horizon result replaces the prespecified H32 criterion.

H.6 Execution and reproduction boundary

All 1,280 fresh surfaces completed on eight RTX 5090 replicas using frozen PyTorch 2.11.0/CUDA 12.8, Transformers 5.12.1, BF16 backbone and FP64 readout. Pre-outcome qualification checked cold processes, cards, singleton/two-route batches and GB10 (maximum endpoint difference 1.78×10^{-14}); checked fingerprints and repeats matched exactly. Independent saved-surface verification reproduced the statistics and verdict to 1.78×10^{-15}. The review edition's `artifact.submission_replay` recomputes tables, controls, horizons and Figure 2's means from source profiles, not omitted token arrays. The root README maps replay/export code. This is not model or allocation-registry replay; parallelism adds no sample units or serving gains.

I OPENED SHARED-REPERTOIRE BALANCE

This finite-design control extends the eight-document, layout-zero development analysis in Appendix H.4, not the 80-document confirmation. It asks whether the contrast persists when the Shared reference averages the same partition repertoire as the cyclic Mixed ensemble. No scientific significance gate or automatic next experiment was specified.

I.1 DESIGN, REUSE AND ESTIMANDS

Let $P_0, \ldots, P_3$ be LOW's four partitions. Shared$_j$ uses P_j in every band; Mixed$_r$ places $P_{(b+r) \bmod 4}$ in band b. Both ensembles give each band every partition with frequency $1/4$. This matches partition-identity distributions, not measured singleton losses or internal activations. The four cyclic assignments are not the full 24-permutation ensemble. We reuse Shared$_0$ and Mixed$_{0:3}$ and generate Shared$_{1:3}$, adding $8 \times 2 \times 3 = 48$ surfaces to 80 reused surfaces. The independent units remain eight opened documents, four per corpus, not 128 surfaces. Model revisions, physical layout, tokens, K, BF16/SDPA execution and FP64 endpoints are unchanged; no fresh document is introduced.

For each individual surface, compute $g = \rho_1(32) - \rho_2(32)$ before configuration averaging. For document d, define

$$M_d = \tfrac{1}{4} \sum_r g(\text{Mixed}_r), \qquad S_d = \tfrac{1}{4} \sum_j g(\text{Shared}_j), \qquad S_{0d} = g(\text{Shared}_0).$$

Then $\Gamma_{0,d} = M_d - S_{0d}$ and $\Gamma_{\text{bal},d} = M_d - S_d$, with

$$\Gamma_{\text{bal},d} - \Gamma_{0,d} = S_{0d} - S_d.$$

Here Γ_0 is the rotation-balanced column of Table 25, not the fresh Table 3 contrast or a comparison with only Mixed$_0$. Intervals use the original 20,000-draw paired-document bootstrap stratified by corpus (seed 20260913318). They are opened, nominal and individual.

Table 27: Opened Shared-repertoire control, eight document units. All shift intervals include zero. No joint success criterion was imposed, and NLL/KL are not independent replications.

Model	Endpoint	Γ_{bal} [95% CI]	$\Gamma_{\text{bal}} - \Gamma_0$ [95% CI]
Qwen	NLL	$0.01237\ [0.00940, 0.01547]$	$-0.00921\ [-0.02839, 0.01342]$
Qwen	KL	$0.01833\ [0.01139, 0.02515]$	$-0.00614\ [-0.02701, 0.02024]$
Llama	NLL	$0.00411\ [-0.00409, 0.01407]$	$0.02980\ [-0.00597, 0.07659]$
Llama	KL	$0.00924\ [0.00011, 0.01980]$	$0.01205\ [-0.04217, 0.07765]$

The Qwen balanced contrast is positive on both endpoints, with positive document values in 8/8 NLL and 7/8 KL cases. Its balanced $\Delta\rho_1/\Delta\rho_2$ means are $-0.00112/0.01124$ (NLL) and $-0.00223/0.01610$ (KL). Llama NLL remains uncertain; Llama KL has a positive nominal interval with a small lower bound. The Qwen point estimates fall from $0.02158/0.02447$ to $0.01237/0.01833$, but the shifts themselves are not established as nonzero. The balanced reference still includes Shared$_0$ with weight $1/4$: persistence does not prove that its contribution is absent.

Under uniform random physical layouts, every Shared$_j$ has the same marginal distribution as Shared$_0$: each arises by a fixed within-row permutation of the randomized block array. Hence the expectation of $S_{0d} - S_d$ over that layout distribution is zero. The realized shift on these eight layout-zero instances is not evidence that the earlier random-layout confirmation was systematically biased. The control improves the finite-design interpretation, not identification of a unique hidden pathway.

I.2 SIGNED GAPS AND CONCENTRATED NEGATIVE CONTRIBUTIONS

Positive Γ means a decrease in signed g, not necessarily in $|g|$. Llama's mean Mixed/Shared gaps are $-0.00555/-0.00966$ for NLL and $0.00212/-0.00712$ for KL. Despite the sign crossing between the reported KL sample means, no document changes the sign of its Mixed versus Shared ensemble-mean gap in the balanced comparison (0/8 for each endpoint). Individual configurations within a

document may still disagree. Neither a systematic input-wise reversal nor architecture causality follows.

An opened arithmetic audit locates Llama NLL's old negative Γ_0 mean: four documents are positive and four negative, but two account for 92.4% of the negative-magnitude sum. Table 28 shows the largest negative contributor. It is retained, not excluded or called noise.

Table 28: Outcome-selected descriptive example: GovReport R40616, Llama NLL, the most negative old Γ_0. This is not a representative sample or a predictor of which new documents behave similarly.

M_d	S_{0d}	S_d	$\Gamma_{0,d}$	$\Gamma_{\mathrm{bal},d}$
-0.06167	0.11996	-0.06177	-0.18163	0.00010

On that document, the large discrepancy is specific to the Shared_0 comparison and is nearly absent after Shared averaging. This identifies an arithmetic source of the result, not its internal cause. Magnitude summaries also depend on where absolute values enter: $|\mathbb{E}_d M_d| - |\mathbb{E}_d S_d|$, $\mathbb{E}_d(|M_d| - |S_d|)$, and averaging individual configuration magnitudes answer different questions and can disagree. All are retained in the artifact; none replaces the signed estimand or supplies a new inferential claim. Changing three or four bands together does not identify a single band's causal contribution. No further partition/band attribution is claimed.

I.3 Execution and reproducibility

All 48 new surfaces completed on GB10. After an interrupted run, recorded hash checks and cold qualification preceded collection of unfinished surfaces; failure and resume receipts are retained in the artifact. Each source replayed checked old Shared/Mixed batches and verified mask budgets, first/last repeats, dense restoration and teardown. Independent saved-surface verification agreed to 1.78×10^{-15}.

The review edition retains 48 new endpoint surfaces but omits the 80 reused token surfaces. `artifact.submission_replay` recomputes contrasts and intervals from supplied source profiles, retaining the descriptive case. The root README maps replay/export code; plans and sanitized failure records remain. Full tensor reconstruction needs omitted inputs; no model, hidden-state or source-registry replay is claimed.

J Opened Pair-Incidence-Matched Follow-ups

This separate development study tests two constructions with identical full pairwise support-incidence tables, not only identical mean overlap. It uses eight already-opened documents (four per corpus), two physical layouts and four cyclic depth placements, on the same checkpoints. The resulting $8 \times 2 \times 4 \times 2 \times 2 = 256$ full surfaces are not independent samples: inference averages layouts and placements within each document first. There was no directional or significance-based success gate and no fresh confirmation followed. This is not a retest of the 80-document recipe contrast in Section 6.6.

J.1 Construction and preservation contrast

Index 27 physical blocks by $(x, y, z) \in \{0, 1, 2\}^3$ and assign four band labels, modulo three, using

$$\text{Linear} = (x, y, z, x + y + z), \qquad \text{Quadratic} = (x, y, z, x + y + z^2).$$

Action a selects the blocks with label a in its band. Each action reads nine 128-token blocks ($K = 1152$, unlike the earlier $K = 1200$); every pair's 3×3 intersection table has all entries equal to three. Both constructions have 27 distinct membership signatures. Linear has uniform triple tables; writing $w = x + y + z^2 \pmod 3$, Quadratic changes the (x, y, w) triple to counts 0, 1 and 2, while the other three remain uniform. This matches pair incidence but does not isolate a single internal mechanism or all higher-order properties. Both use the same 4096-token prompt, local 256-token window and 32 targets.

For each surface let $g = \rho_1(32) - \rho_2(32)$ and compute $D_g = g_{\text{quad}} - g_{\text{lin}}$ before configuration averaging. Positive D_g means a larger *signed* gap, not better quality. Intervals use the original 20,000-draw corpus-stratified document bootstrap (seed 202609153150); they are individual, nominal and opened, not simultaneous or confirmatory.

Table 29: Pair-incidence-matched development comparison, eight document units. All four mean-gap intervals include zero; this neither establishes equivalence nor supplies a new directional confirmation.

Model	Endpoint	g_{lin}	g_{quad}	D_g	Nominal 95% CI
Qwen	NLL	0.03722	0.03654	-0.00068	$[-0.00524, +0.00311]$
Qwen	KL	0.08121	0.08320	+0.00199	$[-0.00414, +0.00869]$
Llama	NLL	0.00831	0.01093	+0.00262	$[-0.00390, +0.01026]$
Llama	KL	0.01914	0.02460	+0.00546	$[-0.00161, +0.01335]$

With only four sources per corpus, these exploratory intervals do not rule out small effects or quantify their detection power. The original bootstrap is not retrospectively replaced by a different success criterion. These uncertain contrasts do not establish equivalence or an isolated triple-overlap mechanism. The following diagnostics concern a different scalar readout, not a replacement for the preservation-gap estimand.

J.2 SCALAR READOUT AND REFLECTION CALIBRATION

For a fixed checkpoint, let q_{dt} be the original dense distribution and y_{dt} the target at document d, token t. With saved post-normalization hidden state h and LM-head matrix W, define

$$\ell_{dt}^{64}(h) = \left(W^\top (q_{dt} - e_{y_{dt}})\right)^\top h, \qquad F_X(d,P,r) = \frac{1}{81 \cdot 32} \sum_{a,t} \ell_{dt}^{64}(h_{X,d,P,r,a,t}) \,.$$

Here $e_{y_{dt}}$ is the target one-hot vector, X is Linear or Quadratic, P a physical block assignment and r a cyclic placement. Both heads have no bias. We evaluate the projection in FP64 while keeping the original BF16 hidden states and weights fixed; this is not full-precision backbone execution. The native rounded-logit readout is $q_{dt}^\top z - z_{y_{dt}} = \mathrm{NLL} - \mathrm{KL} - H(q_{dt})$; the entropy cancels in paired contrasts. The head64 readout instead removes final-head rounding, not backbone quantization or all numerical effects.

Reflection R exchanges the physical assignments at (x, y, z) and $(x, y, -z)$ modulo three. We acquire only Linear at RP, adding 128 surfaces, and reuse the original two base layouts per document. For fixed d, r,

$$D_d(P,r) = F_{\mathrm{quad}}(d,P,r) - F_{\mathrm{lin}}(d,P,r),$$
$$M_d(P,r) = F_{\mathrm{quad}}(d,P,r) - \frac{F_{\mathrm{lin}}(d,P,r) + F_{\mathrm{lin}}(d,RP,r)}{2}.$$

D is plain and M partially symmetrizes the Linear arm. Since R preserves the uniform layout measure, $\mathbb{E}_P M = \mathbb{E}_P D$; no observation of Quadratic(RP) or bitwise reflection invariance is assumed. Each estimator averages four placements and two base layouts within document, then eight documents equally (four per corpus). Reflections are not new independent layouts or sources, and F is not the nonlinear retention statistic g.

Table 30: Opened head64 scalar contrasts, eight document units. Both interval procedures are shown for both estimators. Bootstrap percentiles enumerate all 65,536 within-corpus empirical resamples; Welch intervals use the equal-corpus formula in Appendix H with $n_f = 4$. These individual nominal intervals are not fresh confirmation or exact population-coverage guarantees.

Model	Estimator	Mean	Bootstrap 95% CI	Welch 95% CI
Qwen	Plain D	-0.00358	$[-0.00673, -0.00050]$	$[-0.00826, +0.00110]$
Qwen	Reflected M	-0.00099	$[-0.00456, +0.00285]$	$[-0.00667, +0.00469]$
Llama	Plain D	-0.00563	$[-0.01171, +0.00033]$	$[-0.01590, +0.00464]$
Llama	Reflected M	-0.00353	$[-0.00903, +0.00078]$	$[-0.01211, +0.00504]$

Planning criterion, not an effect-sign test. The frozen rule would prefer reflection for future acquisition only if Qwen's cost-adjusted ratio $\mathcal{R} = 1.5\widehat{V}_{\mathrm{one}}(M)/\widehat{V}_{\mathrm{one}}(D) < 0.8$ and every equal-corpus leave-one-source-out ratio was below one. The factor 1.5 charges three rather than two surfaces. After averaging placements, $\widehat{V}_{\mathrm{one}}$ averages the within-corpus variances of two-layout source means, plus half the corpus-average within-source layout variance. Qwen gives $\mathcal{R} = 2.155$ and leave-one-out values 1.850–2.738: the rule is not met. Llama's descriptive ratio is 0.948; it does not replace the prespecified Qwen rule. These small-sample variance proxies neither guarantee power nor measure GPU speedups. We do not choose the plain estimator because its mean is more negative.

Table 30 exposes sensitivity to both coupling and inference: Qwen's plain bootstrap interval excludes zero, but its Welch interval and both reflected intervals include zero. Neither Llama estimator resolves a mean direction. No robust new mechanism, preservation-gap explanation or quality/cost improvement is claimed, and no fresh confirmation was launched. The review edition's `artifact.submission_scope` recomputes both tables from per-document statistics; full token arrays are omitted. The root README maps replay/export code and retained diagnostic records. Neither archived searches nor model/head execution are rerun. This opened branch neither replaces nor invalidates the fresh support-recipe result.

K PROSPECTIVE CALIBRATED PREDICTION AT EIGHT-BAND RESOLUTION

The first cohort below tests mean-surface prediction. Its opened document-level diagnostics motivate, but are not replaced by, the separate document-first confirmation in Appendix L.

Development and frozen question. An opened eight-document study (four per original corpus) nominated the Llama-interleaved predictor. Its mean-surface NLL RMSE reduction relative to unary was 25.76%; the other cells were -10.29%, 3.79% and 9.99% (Qwen contiguous, Qwen interleaved, Llama contiguous). This motivated a *new-document prediction* test, not a claim about an optimized selector. The primary condition, regularization, calibration roles and success criterion were frozen before the new outcomes. Secondary cells cannot rescue a failed primary. This study does not reuse the new sources to select a new method.

Sources and coordinates. The confirmation allocates 24 GovReport CRS train documents and 24 PG-19 training-shard documents. The GovReport split changes because earlier validation allocations depleted the eligible pool: this is independent-document validation, not a literal repeat of the old validation-split population. Outcome-blind ordering excludes known used and reserved identities, raw and normalized document hashes. One character anchor per document is shared across tokenizers; native 4,096-token prefixes and 32-token targets may differ. Length eligibility is checked without a dense-quality filter. The fixed 48-document allocation is not a claimed power guarantee and is not enlarged after inspecting outcomes.

Both original checkpoints, libraries, remote K$=1{,}200$ and local 256-token suffix remain unchanged. Splitting each Qwen nine-layer band into four/five layers, and each Llama eight-layer band into four/four, yields eight factors. Assigning equal actions to both halves exactly recovers the original four-band coordinates. The full space has $3^8 = 6{,}561$ actions; its complete loss surface is not measured. The inherited full-rank QR design contains 129 fitting actions and 32 disjoint tuning actions. A seeded uniform sample of 128 held actions excludes all 225 development actions (129 fit, 32 tune, 64 test) and the tied four-band diagonal. These same action roles apply to every source, model and library.

Estimator and information access. Write $Y_s(a) = 32^{-1}\sum_t F_{s,t}(a)$. The degree-one and degree-two feature maps contain 17 and 129 columns, respectively, including an intercept and orthonormal ternary contrasts. For $d \in \{1, 2\}$, we fit each source separately:

$$\widehat{\beta}_{s,d} = \arg\min_{\beta} \sum_{a \in A_{\text{fit}}} (Y_s(a) - \phi_d(a)^\top \beta)^2 + \lambda_d \|\beta_{-0}\|_2^2, \qquad \widehat{Y}_{s,d}(a) = \phi_d(a)^\top \widehat{\beta}_{s,d}.$$

The intercept is unpenalized; zero regularization uses least squares. The development grid was $\{0, 1, 10, 100\}$, and the chosen lambdas are frozen per model/library/endpoint (Table 33). This restricted fit is not the exact eight-factor fANOVA of an unobserved full surface. There are $n = p = 129$ nominal fitting rows and pairwise columns, but ridge is not unregularized interpolation: its effective fitted degrees of freedom are $\mathrm{tr}[X(X^\top X + \Lambda)^{-1}X^\top]$, with an unpenalized intercept and a generalized inverse where required. Equality of counts does not establish zero residual degrees of freedom or an identified variance cause. The opened budget and block-penalty controls in Appendix L test estimator dependence, not that cause. The design is full rank with condition number 13.98. Frozen pairwise ridge has fitted degrees of freedom 125.90 at $\lambda = 1$ (Qwen contiguous) and 110.25 at $\lambda = 10$ (the other cells), independently checked by centered SVD. Model responses differ even for the same design and penalty; nominal $n = p$ alone cannot account for those differences. Fit outputs and held predictions are committed before test endpoints enter analysis. The fitting interface reads only the fit/tune roles.

Let U and P denote the original fitted predictors. Constant, U, P, $(U + P)/2$ and $U + \alpha(P - U)$ are all reported in the package. The secondary blend estimates one clipped $\alpha \in [0, 1]$ from the *mean* tuning surface of the 48 documents, separately in each model/library/endpoint; it does not estimate 48 different alphas. It cannot replace the original primary P. Although U and P need only the 129 fitting labels after lambda freeze, the complete comparison acquires and charges all $129 + 32 = 161$ calibration labels per document. The 128 test labels are evaluation-only. This design does not test coefficient, interaction-direction or shared-geometry transfer without calibration.

Estimand and uncertainty. Define $\bar{Y} = 48^{-1}\sum_s Y_s$ and $\overline{\widehat{Y}}_d = 48^{-1}\sum_s \widehat{Y}_{s,d}$. The primary estimand is

$$G = 1 - \frac{\sqrt{128^{-1}\sum_{a\in A_{\text{test}}}(\bar{Y}(a) - \overline{\widehat{Y}}_2(a))^2}}{\sqrt{128^{-1}\sum_{a\in A_{\text{test}}}(\bar{Y}(a) - \overline{\widehat{Y}}_1(a))^2}}.$$

The sole primary is Llama-interleaved NLL: point $G \geq 0.10$ and the lower limit of its two-sided crossed 95% bootstrap interval must exceed zero. We use 20,000 replicates (seed 202609163193), resampling 24 documents within each corpus and 128 test-action indices. The paired draws are shared across methods/models; the secondary blend's alpha is recomputed from resampled tuning predictions. Training actions and lambdas stay fixed. Source-only intervals conditional on the held set are also released. These approximate individual intervals are not simultaneous coverage guarantees. Documents, test actions and token/forward counts are distinct units.

The primary result passes with $G = 60.23\%$ and crossed interval $[28.20, 64.29]\%$. Its pairwise $R^2 = 0.9323$ and Spearman correlation 0.9579 are descriptive, not extra confirmation criteria. The old development checks ($G \geq 0.10$, $R^2 \geq 0.75$, rank ≥ 0.8, and no regret increase above 0.001 nat) are retained in the package but do not redefine the primary test. The point gain's increase over development is not evidence of a universal individual effect: document population, held settings and averaging over 48 rather than eight sources all change.

Table 31: Full C1 NLL reporting retained alongside the expanded main table. Surface: original mean-surface gain and crossed 95% interval. Document: opened ratio-of-mean-document-RMSE gain. r_U, r_P are mean-surface choice regrets among the 128 held candidates, not global optima.

Model/library	Surface (%) [95% CI]	Document (%)	r_U	r_P
Qwen/C	36.87 [4.18, 46.98]	−4.56	0.00000	0.04399
Qwen/I	39.19 [12.32, 49.63]	6.28	0.04861	0.00000
Llama/C	49.53 [19.41, 58.66]	23.54	0.05647	0.00000
Llama/I	60.23 [28.20, 64.29]	23.53	0.00000	0.00000

Why the mean surface, and what averaging changes. The frozen primary evaluates the cohort-average route-response function outside its fitting configurations. This can describe expected loss across the sampled population, but is not an evaluation of a fixed policy transferred to new documents without calibration. For a fixed feature design and ridge penalty, the prediction operator A_d is linear in training responses, so

$$\frac{1}{n}\sum_s A_d Y_{s,\text{fit}} = A_d\left(\frac{1}{n}\sum_s Y_{s,\text{fit}}\right).$$

Accordingly, fitting each document separately does not change the meaning of the mean-surface primary. An alternate-basis augmented solve verifies this identity numerically. Individual-document information is a separate question.

Let $e_{s,d} = Y_s - \widehat{Y}_{s,d}$ on the fixed held set. We additionally report

$$G_{\text{doc}} = 1 - \frac{\text{mean}_s\sqrt{\text{mean}_a\, e_{s,2}(a)^2}}{\text{mean}_s\sqrt{\text{mean}_a\, e_{s,1}(a)^2}}.$$

This ratio of mean document RMSEs is neither the original mean-surface gain nor the mean of document-relative gains. Table 32 puts the document-first estimand and uncertainty alongside the primary's reported source directions. The added intervals use 20,000 within-corpus source draws (24 per corpus; seed 202609183220), conditional on the 128 held actions. These are opened, nominal summaries, not crossed intervals or a replacement for the prespecified mean-surface primary.

Table 32: Opened document-first prediction summaries. Gain is G_{doc}; source-only intervals condition on the held configurations. + counts documents with smaller pairwise RMSE. Original predictors and all sources are unchanged.

Model/library	Endpoint	RMSE_U	RMSE_P	Gain (%) [95% CI]	+
Qwen/C	NLL	0.13962	0.14598	−4.56 [-11.30, 2.15]	23/48
Qwen/C	KL	0.12710	0.12627	0.65 [-6.80, 7.72]	31/48
Qwen/I	NLL	0.12564	0.11776	6.28 [1.09, 11.12]	30/48
Qwen/I	KL	0.11506	0.10318	10.33 [4.36, 15.70]	36/48
Llama/C	NLL	0.12820	0.09803	23.54 [17.54, 28.59]	42/48
Llama/C	KL	0.11522	0.08756	24.01 [18.51, 28.65]	42/48
Llama/I	NLL	0.11561	0.08840	23.53 [16.88, 29.39]	41/48
Llama/I	KL	0.10524	0.07969	24.28 [17.31, 30.44]	40/48

The contrast can also be understood through the exact error identity

$$\mathbb{E}_{s,a}e_s(a)^2 = \mathbb{E}_a(\mathbb{E}_s e_s(a))^2 + \mathbb{E}_{s,a}(e_s(a) - \mathbb{E}_s e_s(a))^2.$$

For Qwen contiguous NLL, pairwise improves the first term but increases the second, worsening mean document MSE. Both terms improve for Llama. This describes error heterogeneity of fitted estimators, not an internal mechanism.

Table 33: Prediction diagnostics. R_P^2 and rank concern the mean surface. RMSE + counts documents with lower pairwise error on their own held surface; α is the secondary tuning-only blend weight. These are not 48 independent confirmations.

Model/library	Endpoint	λ_U/λ_P	R_P^2	Spearman ρ	RMSE +	α
Qwen 4B / C	NLL	1/1	0.967	0.982	23/48	0.936
Qwen 4B / C	KL	1/1	0.972	0.985	31/48	0.950
Qwen 4B / I	NLL	1/10	0.914	0.950	30/48	1.000
Qwen 4B / I	KL	10/10	0.916	0.955	36/48	1.000
Llama 8B / C	NLL	0/10	0.950	0.977	42/48	1.000
Llama 8B / C	KL	1/10	0.956	0.983	42/48	1.000
Llama 8B / I	NLL	0/10	0.932	0.958	41/48	1.000
Llama 8B / I	KL	0/10	0.931	0.956	40/48	1.000

Prediction is not choice superiority. Mean-surface improvement can coexist with heterogeneous per-source errors: NLL improves for only 23/48 Qwen-contiguous documents, versus 41/48 in the primary cell. Choice minimizes each predicted *mean* surface over the 128 held actions and is scored on their actual mean loss. Ties within 10^{-12} use the smallest lexicographic action ID. Qwen-contiguous's unary choice has zero held regret but its pairwise choice has 0.04399; both primary-cell methods have zero regret. Neither is known to attain the 6,561-action global optimum. These facts prevent promoting prediction confirmation into a selector, cost or zero-query transfer claim. The package also reports each corpus and source, the half/blended controls and paired checkpoint contrasts; secondary comparisons do not identify architecture causes.

Table 34 instead lets each document choose its own held candidate, then averages actual regret. The opened source-paired gain intervals are positive in the two Llama NLL cells and include zero in the two Qwen NLL cells. These differ from the first cohort's mean-surface regrets: in particular, a cohort-optimal choice need not be optimal for each source. The methods retain their original expensive calibration and face no matched-cost direct-search comparator here.

Table 34: Opened document-specific choices among 128 held candidates. $\bar{r}_U, \bar{r}_P$ average document regrets; gain is $\bar{r}_U - \bar{r}_P$. Intervals use the same conditional paired-source draws as Table 32, not new primary tests.

Model/library	Endpoint	Mean r_U	Mean r_P	Gain [95% CI]
Qwen/C	NLL	0.08486	0.08354	0.00132 [-0.03663, 0.04937]
Qwen/C	KL	0.06877	0.06354	0.00522 [-0.03735, 0.05135]
Qwen/I	NLL	0.06918	0.05470	0.01447 [-0.01269, 0.04244]
Qwen/I	KL	0.04380	0.04002	0.00378 [-0.02451, 0.02859]
Llama/C	NLL	0.12461	0.05340	0.07121 [0.03363, 0.11331]
Llama/C	KL	0.08373	0.04293	0.04081 [0.00500, 0.08145]
Llama/I	NLL	0.08688	0.04572	0.04116 [0.00337, 0.08460]
Llama/I	KL	0.08980	0.04751	0.04229 [0.01070, 0.07568]

Individual fitted pair terms versus a shared field (opened). To distinguish individual information from a cohort field, hold each original pairwise fit's intercept and unary coefficients fixed, then replace its fitted pair coefficients by zero or by the equal-corpus mean from the other 47 documents. No held labels choose this replacement or a mixing weight; the diagnostic itself was proposed after outcomes were known. Table 35 reports all cells. Llama's original terms predict better than both replacements; Qwen contiguous NLL does not. This supports individualized fitted pair information in some cells, not exact full-lattice O2, an internal cause, or an efficient way to acquire coefficients. The lower-order coefficients still come from the original joint fit; these controls are not independently optimized unary models. LOO donors overlap, so no naive independent-source interval or fresh confirmation is attached.

Table 35: Opened fitted-component replacement. Gains are reductions in mean document RMSE of the original predictor versus each replacement. Shared pair fields exclude the evaluated document and weight the corpora equally. Each + column counts documents favoring the original predictor over that control.

Model/library	Endpoint	vs. zero (%)	+	vs. LOO (%)	+
Qwen/C	NLL	−4.74	23/48	−4.45	28/48
Qwen/C	KL	0.54	29/48	−1.21	29/48
Qwen/I	NLL	10.83	37/48	10.99	37/48
Qwen/I	KL	12.66	40/48	11.60	37/48
Llama/C	NLL	25.42	45/48	25.36	44/48
Llama/C	KL	26.01	47/48	25.74	44/48
Llama/I	NLL	25.48	44/48	24.71	45/48
Llama/I	KL	25.94	41/48	24.98	45/48

Table 36: Secondary KL prediction and held-set choice, with the same crossed interval procedure as Table 4. All four are secondary, not independent replications of the NLL primary.

Model / library	RMSE reduction [95% CI]	r_U	r_P
Qwen 4B / C	45.55% [15.84, 55.10]	0.00390	0.00390
Qwen 4B / I	41.26% [18.26, 51.04]	0.04787	0.07303
Llama 8B / C	50.97% [22.35, 59.43]	0.04133	0.00000
Llama 8B / I	61.39% [29.89, 65.52]	0.04493	0.00000

Execution and reproduction. All $48 \times 2 \times 2 \times 289 = 55{,}488$ scientific settings completed on GB10 with the fixed BF16/SDPA backbone and FP64 endpoint readout. Recorded interruptions and resumes preserve sources, settings, estimators and criteria; only hash-verified complete batches were reused. The review edition's `artifact.submission_replay` refits from exact per-document train/tune/test mean losses and recomputes predictions, choices and intervals. C1 mean arrays are retained exactly; omitted token arrays concern domain/32B below. The root README maps replay/export code. CPU replay neither regenerates model outputs nor independently certifies source allocation.

L INDEPENDENT DOCUMENT-FIRST PREDICTION CONFIRMATION

Question and new population. Mean-surface accuracy need not imply accuracy on individual documents. After the first cohort and its component-replacement diagnostics were opened, we froze a distinct test of document-specific prediction, retaining Llama-interleaved NLL as the sole primary condition. We allocate 48 further documents, 24 each from GovReport CRS train and PG-19 train, without changing the first cohort's estimand or success criterion. Outcome-blind selection excludes known used/reserved identities and raw/normalized text hashes in a registry spanning 536 historical files, including the first cohort. A separate source verifier reconstructs inclusion and native input windows. This verifies known recorded overlap, not pretraining exposure or all semantic duplicates.

The two checkpoints, fixed support libraries, eight-band coordinates, 4,096-token prefixes, 32-token targets and K remain unchanged. The 129 fit and 32 tuning configurations are inherited. The 128 new held configurations exclude all development and first-cohort actions and the tied four-band diagonal; they are shared across documents, models and libraries. No source or action is replaced based on outcomes. Coefficients still require calibration on each new document: this is not zero-query source transfer.

Frozen shared-pair comparator. Besides the same unary U and pairwise P ridge estimators and fixed lambdas, we test a shared-pair prior S. Fit the degree-two model to the first cohort's equal-corpus mean *129 training responses only*, retaining its fitted pair coefficients $b^{\mathrm{old}}_{\mathrm{pair}}$ and setting its intercept and unary coefficients to zero. Define the fixed field

$$h(a) = \phi_{\mathrm{pair}}(a)^\top b^{\mathrm{old}}_{\mathrm{pair}}, \qquad \widehat{Y}_{s,S}(a) = h(a) + \phi_1(a)^\top \widehat{b}_s,$$

where $\widehat{b}_s$ is a fresh unary ridge fit to $Y_s(a) - h(a)$ on the new document's 129 fitting actions, with the frozen unary penalty. Neither cohort's held labels construct the prior. Unlike the opened component swap, this comparator refits its lower-order terms after fixing the pair field. Beating it tests added value of document-specific calibration against this particular prior, not every shared representation or exact O2. In particular, it tests a frozen mean loss-pair field, not the KV/index-sharing operators used by cross-layer inference systems. Its weak advantage over unary does not establish that such systems cannot amortize computation.

The primary uses the original P, not a tuned replacement. Half blends, zero-pair ablations and $U + \alpha_s(P - U)$ are secondary. Here each document's $\alpha_s \in [0, 1]$ is fitted using only its 32 tuning responses, unlike the first study's cohort-level blend. No test labels enter fitting or tuning; predictions are hash-committed before test evaluation. The complete comparison charges 161 new-document calibration observations. The prior additionally uses 48×129 earlier observations per model/library, so it is not a cost-free comparator. The 128 held labels are evaluation-only. The constant half weight and tuning-only rule were specified before C2 model outcomes; neither was introduced after inspecting the Qwen test results.

Document-first estimand and prespecified decision. For each control $C \in \{U, S\}$, compute

$$R_{s,m} = \sqrt{\frac{1}{128} \sum_{a \in A_{\mathrm{test}}} (Y_s(a) - \widehat{Y}_{s,m}(a))^2}, \qquad G_C = 1 - \frac{48^{-1} \sum_s R_{s,P}}{48^{-1} \sum_s R_{s,C}}.$$

This is a ratio of mean document RMSEs: square roots precede source averaging. It is neither RMSE on the mean surface nor a mean of document-relative gains. Confirmation requires all three frozen conditions in Llama-interleaved NLL: $G_U \geq 0.10$, its crossed 95% lower limit above zero, and the corresponding lower limit for G_S above zero. The two contrasts share the same pairwise predictions and are not independent replications. No secondary result can rescue a failed primary.

Twenty thousand crossed bootstrap replicates (seed 202609183233) resample 24 sources within each corpus and 128 held-action indices. Each replicate applies a shared action draw to all sources/methods, recomputes each source's RMSE on that draw, then averages with the paired source weights. Fit designs, lambdas and the development prior remain fixed: uncertainty is conditional on these learned choices, not a bootstrap of prior training. The intervals are individual approximate intervals, not simultaneous guarantees. Document counts, action counts and GPU evaluations are not interchangeable sample sizes.

Table 37: New document-first confirmation. Gain is G_C in percent with an individual crossed 95% interval; + counts documents with smaller pairwise RMSE than the indicated control. Only Llama-interleaved NLL is primary; all other cells and the counts are secondary/descriptive.

Model/library	Endpoint	vs. U (%) [95% CI]	+	vs. prior (%) [95% CI]	+
Qwen/C	NLL	0.01 [-7.45, 6.90]	27/48	−0.19 [-7.27, 6.65]	30/48
Qwen/C	KL	4.47 [-3.30, 11.49]	32/48	3.33 [-4.06, 10.20]	28/48
Qwen/I	NLL	6.56 [-0.52, 13.15]	33/48	6.08 [-1.15, 12.78]	32/48
Qwen/I	KL	10.73 [3.95, 17.12]	38/48	9.52 [2.50, 16.25]	32/48
Llama/C	NLL	23.03 [17.97, 27.86]	43/48	22.07 [16.85, 26.92]	43/48
Llama/C	KL	24.33 [19.43, 29.14]	46/48	23.33 [18.36, 28.16]	47/48
Llama/I	NLL	28.69 [22.26, 34.72]	45/48	28.06 [22.10, 33.72]	46/48
Llama/I	KL	31.47 [24.87, 37.61]	45/48	30.35 [24.02, 36.29]	46/48

Result and scope. The primary passes with $G_U = 28.69\%$ $[22.26, 34.72]\%$ and $G_S = 28.06\%$ $[22.10, 33.72]\%$. Pairwise mean document NLL RMSE is 0.07785, versus 0.10917 for unary and 0.10822 for the shared prior. There are 45/48 and 46/48 improving documents, respectively. Llama-contiguous also has positive secondary intervals, whereas both Qwen NLL cells include zero. The result is not a universal model-family effect, an architecture explanation, or proof that every document benefits. Differences from the first cohort's effect sizes do not isolate a population or hardware cause.

Secondary document-specific choices minimize each fitted surface over the 128 held candidates, with the inherited 10^{-12} tie tolerance and lowest action-ID rule. In the primary cell, mean actual regret falls from 0.09645 to 0.04921 versus unary; the gain is 0.04724 with a source-only 95% interval $[0.00719, 0.08797]$, conditional on the held set. This is not the primary crossed prediction contrast, the global 6,561-route optimum, or a matched-cost comparison with direct search. Full per-source predictions, corpus summaries, controls and both endpoints are released. Query-efficient selection, generation quality and repayment of calibration cost remain untested.

Prespecified scalar controls, separately reported uncertainty. Table 38 reports the same fixed half-blend $P_{1/2} = (U + P)/2$ and tuning-only blend $P_{\alpha,s} = U + \alpha_s(P - U)$ across every model/library/endpoint, rather than selecting a different method for each cell. Original fits, observations and P-primary decisions remain unchanged. These methods and their point estimates were prespecified secondary outputs; the added interval reporting is opened. For either blend m, the statistic is $1 - \text{mean}_s R_{s,m} / \text{mean}_s R_{s,U}$, using the same 20,000 paired crossed source/action draws as the primary. Calibration fits and tuning-derived α_s remain fixed under held-action resampling. Intervals are nominal per comparison, not a newly registered joint test or simultaneous coverage for all cells and methods.

Table 38: Prespecified scalar controls on C2, with opened nominal crossed 95% intervals. Gains are reductions in mean document RMSE versus unary; + counts improving documents. All cells/methods are retained. The original P primary in Table 37 is unchanged.

		Fixed half (%)		Tune-only (%)	
Model/library	Endpoint	[95% CI]	+/48	[95% CI]	+/48
Qwen/C	NLL	18.82 [15.37, 22.11]	48/48	18.55 [14.75, 22.23]	48/48
Qwen/C	KL	20.86 [17.17, 24.34]	48/48	20.26 [15.64, 24.67]	46/48
Qwen/I	NLL	18.30 [15.09, 21.40]	48/48	18.51 [14.65, 22.37]	48/48
Qwen/I	KL	17.87 [14.65, 21.01]	48/48	18.65 [14.31, 23.09]	48/48
Llama/C	NLL	26.41 [23.97, 28.82]	48/48	28.87 [25.47, 32.38]	48/48
Llama/C	KL	26.79 [24.34, 29.22]	48/48	29.10 [25.41, 32.97]	48/48
Llama/I	NLL	28.69 [25.96, 31.36]	48/48	32.74 [27.83, 37.70]	47/48
Llama/I	KL	29.80 [27.01, 32.40]	48/48	35.46 [30.43, 40.45]	48/48

The half-blend improves all 48 observed documents in each NLL cell, while the tuning-only blend improves 48, 48, 48 and 47 (Qwen C/I, Llama C/I). These are finite-sample directions, not guaranteed benefits for every input. The same fixed-half rule narrows, but does not remove, the Llama–Qwen relative-gain difference: the paired opened contrasts are 7.59 percentage points $[3.53, 11.76]$ for con-

tiguous and $10.39\ [6.21, 14.71]$ for interleaved. This locates estimator sensitivity, not an architecture cause or a causal fraction explained by shrinkage. In particular, $P - U$ includes intercept/unary changes from joint refitting, not only an isolated exact O2 field. Error balancing remains a possible explanation. No extra model observations are introduced by this reporting. These C2 contrasts retain the original 161-observation full-comparison budget and do not themselves test matched-cost search. Appendix D.4 separately compares the original 129-fit predictors with direct search at 130 queries; it does not test the later block-ridge family.

Opened fitted-component attribution and choice. To test whether lower-order refitting alone reproduces the half-blend gain, write the original predictor as $P = L + Q$, with L its fitted intercept/unary part and Q its fitted degree-two field. Hold all original coefficients fixed and compare $(U + L)/2$ with $U + Q/2$. These are post-hoc component interventions, not separately optimized methods or exact O2 of the unobserved full loss lattice. No held label selects a scale or a replacement predictor.

Table 39: Opened C2 fitted-component controls: reduction in mean document NLL RMSE relative to unary (%). Original half is $(U + L + Q)/2$. All cells are kept.

Model/library	$(U + L + Q)/2$	$(U + L)/2$	$U + Q/2$
Qwen/C	18.82	−0.08	18.65
Qwen/I	18.30	−1.83	20.10
Llama/C	26.41	−0.12	26.20
Llama/I	28.69	−0.22	28.87

Lower-order mixing alone does not reproduce the mean predictive gain, whereas adding $Q/2$ to fixed U retains substantial improvement. This supports useful fitted pair-field information within these predictors, not a unique internal cause or a causal percentage of gain. With $e = Y - U$, $l = L - U$, and $q = Q$, the per-document half-blend MSE gain is exactly $\langle el - l^2/4\rangle + \langle eq - q^2/4\rangle - \langle lq\rangle/2$. The cross term remains, and this MSE identity is not an additive attribution of the ratio of mean RMSEs. All eight endpoint cells and nominal crossed intervals are released; original primary estimates are unchanged.

Table 40: Opened half-blend NLL choices among the 128 held candidates. Gain is $r_U - r_{1/2}$ with paired source-only nominal 95% intervals conditional on that set; $+/-/=$ counts better/worse/tied document regrets.

Model/library	r_U	$r_{1/2}$	Gain [95% CI]	$+/-/=$
Qwen/C	0.08672	0.05911	0.02761 [0.00182, 0.05644]	12/8/28
Qwen/I	0.06307	0.04515	0.01792 [0.00391, 0.03295]	13/4/31
Llama/C	0.08304	0.04239	0.04064 [0.02186, 0.06106]	16/1/31
Llama/I	0.09645	0.03551	0.06094 [0.02872, 0.09572]	16/5/27

Half-blend reduces average held-set regret in all four NLL cells in this opened comparison, including both Qwen cells. The 48/48 RMSE directions do not imply 48/48 improved choices; Table 40 retains adverse and tied cases. Nor should the newly constructed $U + Q/2$ replace half on the basis of better RMSE: in Qwen-interleaved, its regret gain is only 0.000095 with interval $[-0.02464, 0.02162]$, despite its larger RMSE improvement. These controls neither certify the full 6561-route optimum nor compare equal- cost search. Source-only choice intervals and crossed prediction intervals answer different questions; neither is a new confirmatory verdict.

C2 execution and reproduction. All $48 \times 2 \times 2 \times 289 = 55{,}488$ scientific settings complete on two qualified H100 replicas using the fixed BF16 backbone and FP64 endpoint readout. Source allocations, methods and criteria remain frozen; there is no outcome-based stopping or automatic follow-up. Repeat, nested-coordinate, restoration and patch-teardown checks pass. An independent augmented-solve implementation agrees with stored predictions within 6.13×10^{-14} and checked quantities within 2.24×10^{-13}. The anonymous CPU package refits from saved calibration/test

means, reconstructs the shared prior from the earlier training labels, and replays the contrasts and intervals. It does not rerun GPU kernels or independently recover licensed source text and historical allocations.

L.1 Opened calibration-budget intervention

To test whether original P's predictive behavior depends on its calibration regime, we select eight already-opened C2 documents (four per corpus) by a declared identity-hash order, not effect size. Each retains its original 129 fit actions, 32 tuning actions and 128 held actions. Two predeclared, disjoint sets of 128 additional fitting actions each exclude all original roles and the tied four-band diagonal. The models, libraries, token windows and endpoints remain unchanged. All eight documents pass exact H100 dense, first-fit/test, repeat, coarse/fine and restoration checks before new acquisition.

Each nested 257-action fit is compared with its original 129 baseline on the SAME eight documents. To preserve the penalty in mean-squared-loss units, $\lambda_{257} = \lambda_{129}(257/129)$, leaving the intercept unpenalized. A separate fixed-sum-λ sensitivity is retained, not selected as the main comparison. Original U, P, fixed half and tuning-only blends are reported at each budget; no held labels select a penalty, design or mixing rule.

Table 41 defines NLL gain (%) as $100(1-\overline{\mathrm{RMSE}}_P/\overline{\mathrm{RMSE}}_U)$ against same-budget unary. In general, normalized ridge uses $\lambda_n = \lambda_{129}n/129$. Both extension designs A and B are retained, not best-of. Choices, half/tune-only, KL and fixed-sum sensitivity are in the full replay JSON. This is a nested quantity-and-coverage comparison, not a neural-cause, fresh-confirmation or efficiency claim.

Table 41: Opened fitting-budget comparison on the same eight C2 documents (four per corpus), not 16 sources. NLL gains (%) are reductions in mean document RMSE relative to same-budget unary. Each 257-label design adds its own disjoint 128-action extension to the original 129; both A and B are reported. C/I denote contiguous/interleaved libraries.

Cell	P_{129} gain	P_{257} A gain	P_{257} B gain
Qwen C	-3.6	33.6	35.3
Qwen I	14.6	31.1	31.3
Llama C	20.0	40.7	43.1
Llama I	29.2	49.8	46.8

Under normalized ridge, P's mean document NLL error decreases by 19.8–37.5% relative to its own 129-action fit. Each of the eight documents improves in each NLL cell for both extensions. At 257 observations, original P also beats same-budget half-blend on mean RMSE in all four NLL cells of both designs. This is new evidence that the nested acquisition regime matters, not proof that all original differences were caused solely by observation count: quantity, coverage and conditioning change together, and the sources/held targets are already opened. Two extensions do not create 16 independent sources.

Choice remains distinct. For Qwen-interleaved, P's mean fixed 128-candidate regret changes from 0.06020 at 129 to 0.08024 and 0.03944 for extensions A and B. Prediction improves in both, but choice worsens in A. All source, KL, penalty, half/blend and alpha results are released. The extra 128 training observations are a cost, not a query-efficiency improvement or a new optimizer.

The study adds 8192 model settings. A parent backup inventory race triggered a technical stop; completed batches were verified before explicit same-contract resume and cold qualification. Sources, code, tolerances, designs, penalties and the original deadline were unchanged; failure records remain in provenance. Independent rotated/augmented fits and 1280 scalar metric checks agree within 1.45×10^{-13}. This is an opened development result, not a new fresh confirmation or an internal mechanism claim.

L.2 Opened objective-aligned regularization control

More fitting labels are not the only possible remedy. On all 48 already-opened C2 documents (24 per corpus), we additionally tune unary and pairwise ridge separately on the original grid $\{0, 1, 10, 100\}$,

using only each document's 32 tuning labels. Denote these per-document fits by U^* and uniform P^*. The block-ridge control, block P^*, selects (λ_U, λ_Q) from the same 4×4 grid under

$$\|y - X\beta\|_2^2 + \lambda_U \|\beta_U\|_2^2 + \lambda_Q \|\beta_Q\|_2^2,$$

with an unpenalized intercept, 16 unary and 112 pair coefficients. Fits use only the original 129 training labels, without refitting on tuning labels. Exact tuning-MSE ties use numeric lexicographic grid order. The full comparison costs 161 observed labels; the original fixed predictors use only 129. This estimator family was proposed after C2 was opened, not prespecified for its confirmation. Held labels do not select its penalties, but the family itself still requires independent validation for a new generalization claim.

Let $H = (U + P)/2$ denote the half-blend of the original frozen fits. Table 42 reports NLL gain (%) for predictor M as $100(1 - \overline{\mathrm{RMSE}}_M / \overline{\mathrm{RMSE}}_{U^*})$. All eight NLL/KL cells, original-U comparisons, selections and per-source metrics are in the replay JSON. This newly proposed posthoc control is not a C2 primary change, fresh confirmation or method promotion, and is not part of P326 (Appendix D.4).

Table 42: Opened regularization control on 48 C2 documents. NLL gains (%) are reductions in mean document RMSE relative to tuned unary U^*; H is the original half-blend and P^* uses tuning-selected penalties. Fits use 129 labels; selection uses 32 more (161 total), never held labels. r is mean actual regret among 128 held candidates, retaining adverse block choices. C/I denote contiguous/interleaved libraries.

Cell	H gain	Uniform P^* gain	Block P^* gain	r_H	r_{block}
Qwen C	19.44	6.83	25.59	0.05911	0.03757
Qwen I	18.53	10.98	26.25	0.04515	0.04370
Llama C	26.96	24.48	33.24	0.04239	0.05121
Llama I	29.03	31.34	38.97	0.03551	0.03917

Tuned unary is not stronger on these NLL point errors: its mean RMSE increases by 0.28–0.77% relative to original U. Half-blend retains an 18.53–29.03% reduction against that comparator. Block ridge reduces RMSE by 25.59–38.97% against tuned unary without new model evaluations. Thus the opened results do not identify more labels as the only or cheapest remedy, and unary tuning does not explain away half's benefit in this bounded menu. They support estimator-policy dependence, not a neural mechanism or an automatically promoted replacement method.

Choice again differs: block ridge's NLL regret exceeds half's by 0.00882 and 0.00366 for Llama C/I. All eight endpoint cells, candidate penalties, source choices and adverse outcomes are retained. Independent centered, rotated-basis solves agree on every selected penalty and held prediction within 5.6×10^{-14}. This control does not modify either C2's primary or the separately frozen query-budget pilot.

M External-Domain and Matched Larger-Checkpoint Scope

Domain population and inclusion. The fresh domain study allocates 32 NLP papers and 32 distinct Python repository clusters. QASPER v0.3 train (Dasigi et al., 2021) supplies paper titles, abstracts and section/paragraph text; QA and evidence labels are not used. RepoBench Python v1.1 (Liu et al., 2024), restricted to its `in_file` stratum, supplies `all_code`, not retrieved snippets, `next_line` or answer labels. Thus "Python" denotes repository-derived text, and evaluation remains gold-continuation NLL/dense-reference KL, not a RepoBench completion score or QASPER QA accuracy.

The data revisions are `fdc9d8214fbab5dd782958601db4d678e6934a54` (QASPER metadata) and `8a7cf0c8942cc1aa066bf261839650ac55a2ff79` (RepoBench Python v1.1). The QASPER train text is read from its public v0.3 archive. Outcome-blind hash ranking starts from 888 paper candidates and 1,707 repository clusters. Geometry-only eligibility skips 12 papers and 158 repositories before selecting 32 per domain, so the population is conditional on sufficient text for the fixed window, not every benchmark item. A 445-record historical/reservation registry excludes known source identities and raw/normalized hashes. Selected-window seven-word-shingle checks find no pair above 0.25 containment. This is not a guarantee against semantic duplicates, unrecorded forks or pretraining exposure.

Each source has one shared character anchor across tokenizers, a native 4,096-token prefix, 32 targets, remote K$=1{,}200$ and local suffix 256. We use the original contiguous/interleaved libraries, four depth quarters and complete 81-action surfaces. The first panel contains the fixed Qwen3-4B-Instruct-2507 and Llama-3.1-8B-Instruct checkpoints. Its $64 \times 2 \times 2 = 256$ surfaces give 20,736 scientific settings. Corpus, model, library and endpoint are not multiplied into independent-source counts.

Larger checkpoint: matched sources, not a new sample. After the smaller-model outcomes were opened, a separately frozen scope study runs Qwen/Qwen3-32B, revision `9216db5781bf21249d130ec9da846c4624c16137`, on *all the same 64 documents*. Four contiguous bands each contain 16 of its 64 full-attention layers. Original character anchors are unchanged; all native windows match the original Qwen inputs exactly. The 128 new complete surfaces add 10,368 settings, not 64 independent documents. Checkpoint, training and depth differ; four bands span 16 layers each rather than the smaller Qwen's nine. Even a matched-source contrast cannot identify the effect of parameter count alone. This panel tests neither a model-family law nor the eight-band predictor from Appendix K.

Two objects and prespecified conjunctions. For each complete surface at $H = 32$, compute $T = \rho_1 - \rho_2$ and $D = r_U - r_P$, with order energies summed before retention ratios and projected choices scored on original mean loss. Average the two libraries *within source*, then average 32 source clusters per domain. For the 4B/8B panel, all eight NLL lower limits (two models, two domains, two objects) must exceed zero. For the 32B panel, all four must exceed zero. Both frozen criteria pass (Table 5). The respective source bootstraps use 20,000 draws and seeds 202609173204/202609173211, sharing draws across objects/endpoints. These intersection–union decisions require all component successes; individual approximate 95% intervals are not a simultaneous confidence set. KL is secondary and cannot rescue NLL.

Table 43: Secondary dense-reference KL scope. Libraries are averaged within each of 32 source clusters per domain. The 32B block reuses the same sources.

Model	Text domain	$\rho_1 - \rho_2$ [95% CI]	$r_U - r_P$ [95% CI]
Qwen 4B	NLP papers	0.1789 [0.1487, 0.2106]	0.0404 [0.0131, 0.0726]
Qwen 4B	Python	0.0941 [0.0703, 0.1196]	0.0327 [0.0118, 0.0559]
Llama 8B	NLP papers	0.0345 [0.0147, 0.0555]	0.0743 [0.0488, 0.1018]
Llama 8B	Python	0.0563 [0.0337, 0.0808]	0.0611 [0.0373, 0.0878]
Qwen 32B	NLP papers	0.2841 [0.2422, 0.3252]	0.0364 [0.0137, 0.0610]
Qwen 32B	Python	0.1694 [0.1243, 0.2151]	0.0539 [0.0307, 0.0806]

The domain NLL temporal contrast is positive in 30/32 and 29/32 Qwen sources, and 21/32 and 23/32 Llama sources (papers/Python); the corresponding choice gain counts are 23/32, 26/32, 20/32 and 25/32. These are mean effects, not per-source laws. Fixed-H retention does not establish an order-dependent decay exponent in these domains. Retrospective choice uses all 81 labels and therefore supplies no calibration-cost advantage or online policy improvement.

Execution controls and released scope. Two H100 PCIe 80GB devices run independent replicas with source sharding, not tensor parallelism. The backbone remains BF16/SDPA, endpoints FP64; there is no quantization or outcome-guided batch change. Technical checks cover dense/no-op replay, patch restoration, support budgets, repeats and batch/singleton equivalence. The smaller-panel GB10/H100 bridge reproduces checked NLL exactly with maximum KL difference 4.44×10^{-15}; a cold cross-card repeat also agrees. A source-input verifier initially translated CRLF; byte-preserving decoding corrected the verifier before model outcomes, without changing source bytes or native inputs. The failure receipt is retained. The 32B panel is separately qualified on already-opened documents outside these 64 sources; original receipts and local backups bind all outputs.

Independent conditional-marginal and contrast-basis implementations reproduce the complete-surface energies, choices and source intervals. The review edition's `artifact.submission_scope` recomputes intervals from per-document statistics, not omitted token endpoints. The root README maps retained hashes, protocols and replay/export code. This neither reruns models nor independently certifies source history.

N OPENED MATCHED-HORIZON SCOPE

Question and evidence status. Does the original short-window result persist when more teacher-forced target positions are averaged? This opened study extends eight already-inspected closure documents, four GovReport and four PG-19, to 128 targets. It is not fresh confirmation, a new primary, a long-prompt benchmark, or free-running generation. All model/library/endpoint conditions remain in the analysis. The original 40-source result is unchanged; the smaller reused sample cannot establish a universal horizon law or the absence of an uncertain effect.

Within each corpus, sources are ordered by the SHA-256 of their existing identity and the first four jointly eligible for both tokenizers are retained. The original character anchor, 4,096-token prefix and first 32 target IDs are preserved exactly; targets extend naturally without padding or a new anchor. All examined science candidates are eligible. A separate opened document is used only for execution checks. The original pinned Qwen and Llama checkpoints, four depth bands, two five-of-15 support libraries and all 81 actions are unchanged. One 128-target endpoint tensor supplies nested $H = 32, 64, 128$ prefixes, giving $8 \times 2 \times 2 \times 81 = 2{,}592$ scientific settings, not three independent cohorts. The prompt length stays fixed.

Statistics and complete reporting. For each source/model/library/loss, recompute token and aggregate order energies, ρ_k, and actual unary/degree-two hindsight choices at each H. Intervals resample four sources within each corpus, sharing draws across conditions and horizons (20,000 draws, seed 202609221280). They are nominal, opened, descriptive intervals, not simultaneous bands. Undefined energy ratios are retained as undefined. The supplement includes every source, horizon, NLL/KL endpoint and the diagnostic top1-agreement surface.

An additional opened summary matches the closure's weighting: average the two models, two libraries and *separately evaluated* NLL/KL regrets inside each source, then average sources (bootstrap seed 202609222100). Table 44 is not pure NLL, a mean of percentage improvements, or the regret of a single route jointly selected for both endpoints. The H128–H32 change in pooled gain is -0.03143 with nominal interval $[-0.05461, -0.01083]$. The same sample therefore retains a large H32 point advantage; it does not fail to reconstruct the original short-window result. The pooling does not replace the model-specific results below.

Table 44: Opened matched-source horizon summary. Reductions are ratios of pooled mean regrets. Eight source units, not 64 independent conditions.

H	U regret	P regret	Gain [95% CI]	Reduction
32	0.05172	0.01512	$0.03660\ [0.01003, 0.06282]$	70.76%
64	0.03138	0.00957	$0.02182\ [0.00847, 0.03498]$	69.52%
128	0.01355	0.00838	$0.00517\ [-0.00358, 0.01467]$	38.18%

H128 preserves some objects, not every advantage. Qwen retains positive NLL/KL preservation-gap intervals in both libraries, but its choice-gain intervals include zero. Llama-contiguous retains positive gap intervals; its NLL choice interval touches zero and KL spans zero. Llama-interleaved instead has uncertain gaps but positive NLL/KL choice-gain intervals (Table 45). An uncertain gap is not evidence that the gap is absent. These patterns require separate reporting of preservation and choice, not a causal ordering of the two.

Table 45: All H128 NLL/KL conditions, source means and nominal 95% intervals. C/I denote contiguous/interleaved support libraries.

Model/Lib.	Loss	$\rho_1 - \rho_2$ [95% CI]	Gain [95% CI]
Qwen/C	NLL	$0.10433\ [0.08180, 0.12843]$	$0.00563\ [-0.02493, 0.02928]$
Qwen/C	KL	$0.17657\ [0.13933, 0.21812]$	$-0.00708\ [-0.03012, 0.01638]$
Qwen/I	NLL	$0.10254\ [0.06752, 0.13626]$	$0.00265\ [-0.01082, 0.02144]$
Qwen/I	KL	$0.16486\ [0.12351, 0.20609]$	$-0.00311\ [-0.02031, 0.01726]$
Llama/C	NLL	$0.03603\ [0.01804, 0.05560]$	$0.00633\ [0.00000, 0.01493]$
Llama/C	KL	$0.04840\ [0.02923, 0.06840]$	$0.00114\ [-0.00862, 0.01092]$
Llama/I	NLL	$0.00438\ [-0.00347, 0.01166]$	$0.02724\ [0.00654, 0.05472]$
Llama/I	KL	$0.00088\ [-0.00250, 0.00494]$	$0.00857\ [0.00208, 0.01714]$

Headroom and local error, not a capability ranking. Opened complete-surface diagnostics distinguish global fit, remaining unary regret and harmful reranking (Table 46). At H128, Qwen-interleaved unary already matches five of eight NLL optima (seven for KL). Its mean NLL regret is 0.00869, versus 0.78250 for a uniform random action: the surface is not flat enough that arbitrary choices work. Llama-interleaved has more unary misses that pairwise terms recover. Llama-contiguous unary already matches six of eight NLL optima, so the pattern is not a simple ordering by model size. Native-token losses do not establish cross-tokenizer language capability; all these projections used all 81 labels.

Table 46: Opened NLL fit and choice diagnostics. Normalized regrets average document-wise regret divided by the full action-loss range; they are not ratios of the displayed population means or practical quality thresholds.

Model/Lib.	U R^2_{32}	U R^2_{128}	Hits_{128} U/P	$\text{Regret/range}_{128}$ U/P (%)
Qwen/C	0.922	0.956	3/6	1.33/0.84
Qwen/I	0.899	0.956	5/5	0.54/0.54
Llama/C	0.929	0.917	6/7	1.23/0.47
Llama/I	0.900	0.873	2/6	5.03/0.64

Qwen-contiguous improves NLL optimum matches from three to six, yet four positive source contributions totaling 0.14036 are partly offset by one loss of 0.09530; no source is removed. For this post-hoc illustrative extremum, write $F = U + Q + R$, where U includes the mean and unary terms, Q is pure pairwise and R contains degrees three and four. With $a_U = \arg\min U$ and $a_P = \arg\min(U + Q)$,

$$F(a_U) - F(a_P) = [(U + Q)(a_U) - (U + Q)(a_P)] + [R(a_U) - R(a_P)].$$

The predicted improvement 0.14636 is overturned by residual contrast -0.24166, producing actual gain -0.09530. Degree-two global R^2 is 0.9883 in that example. This locates a truncation error near the minimum; it does not identify a hidden-state mechanism. Nested least-squares projection cannot worsen global squared fit, but need not preserve the argmin.

What averaging does not identify. Qwen-interleaved unary R^2 increases from 0.8987 to 0.9561, whereas Llama-interleaved decreases from 0.9003 to 0.8731. Llama-contiguous also loses global unary fit while its choice regret decreases. Thus pooled regret attenuation is not a universal increase in additivity. Qwen-interleaved degree-two NLL regret itself increases from 0.00195 to 0.00604, despite the decrease in the pooled degree-two mean.

Four disjoint 32-token block diagnostics are released as within-source descriptions, not additional independent observations. Qwen's unary block directions align more strongly than its pairwise directions, but mean pairwise cross terms are positive in every model/library/loss group. We do not infer uniform opposing-sign cancellation. The target set, mean loss surface, selected routes and oracle all change with H. Projection linearity does not linearize argmin, and co-occurring energy and choice changes do not identify attenuation as a causal mediator. Nor does longer teacher-forced averaging measure generated task quality or a refresh interval.

Execution and reproducibility. GB10 execution uses the original BF16/SDPA model settings and FP64 endpoint arithmetic, without package or weight changes. Dense/no-op, causal suffix, repeat, mask inventory, restoration and cold checks pass. An initial full-128-row readout differs from historical H32 by at most 2.7×10^{-15} despite exact prefix logits. Before scientific collection, we retain the historical 32-row FP64 readout shape for four chunks of each single 128-target forward. A further cross-batch comparison retains exact logits/top1 but permits NLL/KL roundoff only within $32\epsilon_{64}(1 + \max(|x|, |y|))$ elementwise. Accepted cross-batch differences are not called exact. Same-shape repeats and every scientific action's historical first32 endpoints remain exact. Both failed qualification records are preserved; neither changed scientific sources, endpoints or hypotheses.

All 16 source/model tensors are complete. Independent marginal and orthogonal contrast calculations check energies, choices and source intervals; all 248,832 first32 endpoint scalars match the original archive exactly. The anonymous CPU command `artifact.horizon_scope` recomputes these tables from saved endpoints, not from model execution. The released numerical scope does not certify a new population or create a cost-efficient selector.

O PROSPECTIVE MATCHED-HORIZON CHOICE

From development to a fixed new question. Appendix N's eight opened documents suggested that the absolute choice value of degree-two reconstruction could decrease with the averaging window. We therefore froze a new 80-document study before allocation and model outcomes. Its sole primary is the paired change in choice gain from $H = 32$ to 128, not whether the H128 gain alone is positive. Neither the original closure primary nor the opened pilot is redefined. Absolute gains at all three horizons and condition-specific results are prespecified secondary summaries.

Population and matching. The 40 GovReport CRS-train documents and 40 PG-19 train-shard0 books exclude known earlier/reserved identities, raw/normalized document hashes and public PG-19 validation/test overlap. Selection uses no outcomes. Independent code reconstructs source bytes, selection order and 160 model-specific windows. This certifies local allocation, not pretraining or semantic disjointness. The changed GovReport split makes this a prospective new-population study, not an identical-population replication of the closure pilot.

Both original checkpoints, four depth bands, two five-of-15 libraries, and all 81 actions are unchanged. Each document has a fixed 4,096-token prefix and 128 natural teacher-forced targets under each tokenizer. Shared character anchors do not make cross-tokenizer target strings identical. The same 128-token tensor supplies nested $H = 32, 64, 128$ prefixes. There are $80 \times 2 \times 2 \times 81 = 25{,}920$ scientific settings, but only 80 independent document units. The study changes the averaging target set, not prompt length or a free-running generation horizon.

Estimand and frozen inference. Let $r_{U,d,m,l,e}(H)$ and $r_{P,d,m,l,e}(H)$ be the actual all-action regrets of the constant-plus-unary and degree-two projections at horizon H. Each NLL/KL endpoint selects its own route. Define

$$\Delta_d(H) = \frac{1}{8} \sum_{m,l,e} \big[r_{U,d,m,l,e}(H) - r_{P,d,m,l,e}(H) \big], \qquad D_d = \Delta_d(128) - \Delta_d(32),$$

where the sum covers two fixed models, two libraries and two losses. The primary is $\theta = \frac{1}{2}\mathbb{E}_G[D_d] + \frac{1}{2}\mathbb{E}_{PG}[D_d]$. This composite of native losses is not pure NLL, a capability ranking, or the regret of one route jointly chosen for both losses. Within-document pairing includes every model/library/endpoint.

The authoritative interval is two-sided 95% equal-corpus Welch–Satterthwaite:

$$\widehat{\theta} = \tfrac{1}{2}(\overline{D}_G + \overline{D}_{PG}), \qquad v_f = \frac{s_f^2}{4n_f}, \quad \widehat{V} = v_G + v_{PG}, \quad \nu = \frac{\widehat{V}^2}{v_G^2/(n_G - 1) + v_{PG}^2/(n_{PG} - 1)},$$

with $n_G = n_{PG} = 40$ and limits $\widehat{\theta} \pm t_{.975,\nu}\sqrt{\widehat{V}}$. Before outcomes, a negative upper limit was specified as support for the anticipated decrease, a positive lower limit as the opposite direction, and otherwise an unresolved result, not equivalence. No practical-effect threshold or outcome-dependent sample extension was specified. The fixed 50,000-draw source-stratified percentile bootstrap (seed 20260923803) is sensitivity only; it cannot replace the primary interval. Shared resampling draws preserve all within-source conditions and horizons. Secondary intervals are nominal, not separately powered or simultaneous.

Precision and complete collection. Before outcomes, paired pilot variance and cost planning fixed $N = 80$. Normal half-widths 0.00814 at the unstable pilot SD and 0.01221 at 1.5 times that SD were sensitivities, not coverage or success guarantees. All sources/conditions completed without outcome-based stopping or replacement; the realized Welch half-width is 0.01008.

Primary result: absolute decrease, not disappearance. The estimate is -0.010286 with interval $[-0.020362, -0.000209]$ (SE $= 0.005050$, $\nu = 68.03$). The bootstrap interval is $[-0.019981, -0.000368]$. Both are negative, but the primary upper limit is close to zero. The point decrease is smaller than the opened pilot's -0.03143; neither a large minimum decrease nor its mechanism is established. Corpus-specific point changes are -0.01683 (GovReport) and -0.00374 (PG-19). The primary concerns their fixed equal-weight mixture, not separate confirmations in each corpus.

Table 47: Fresh matched-horizon mean regrets. Gain intervals are nominal secondary 95% Welch intervals over 80 source units. Reduction is the ratio of pooled mean gain to pooled mean unary regret, not the mean of document percentages, a pure-NLL reduction or a generation-quality improvement.

H	U regret	P regret	Gain [95% CI]	Reduction
32	0.05412	0.01038	$0.04374\ [0.03579, 0.05168]$	80.81%
64	0.04478	0.00728	$0.03749\ [0.02967, 0.04532]$	83.73%
128	0.03920	0.00575	$0.03345\ [0.02602, 0.04088]$	85.34%

Pairwise mean regret decreases from 0.01038 at H32 to 0.00575 at H128; unary decreases from 0.05412 to 0.03920. Thus the absolute additional gain falls from 0.04374 to 0.03345, while the ratio-of-means regret reduction rises from 80.8% to 85.3%. The H128 pooled gain interval is $[0.02602, 0.04088]$. The primary absolute decrease is therefore compatible with substantial longer-window pairwise choice value, not evidence that pairwise information vanishes or becomes proportionately less useful.

Table 48: All fresh NLL/KL conditions. C/I denote contiguous/interleaved libraries. H128 gains and their H128–H32 changes have nominal secondary 95% intervals; neither eight simultaneous confirmations nor a within-every- condition horizon law is claimed.

Model/Lib.	Loss	Gain_{32}	Gain_{128} [95% CI]	Δ_{128-32} [95% CI]
Qwen/C	NLL	0.03805	$0.03692\ [0.01760, 0.05624]$	$-0.00113\ [-0.02695, 0.02470]$
Qwen/C	KL	0.03331	$0.03031\ [0.01929, 0.04134]$	$-0.00300\ [-0.02474, 0.01874]$
Qwen/I	NLL	0.02395	$0.01358\ [0.00271, 0.02445]$	$-0.01037\ [-0.02525, 0.00451]$
Qwen/I	KL	0.02015	$0.01977\ [0.01111, 0.02843]$	$-0.00039\ [-0.01229, 0.01152]$
Llama/C	NLL	0.08088	$0.04449\ [0.02873, 0.06025]$	$-0.03639\ [-0.06488, -0.00789]$
Llama/C	KL	0.05805	$0.04532\ [0.03001, 0.06063]$	$-0.01273\ [-0.04133, 0.01587]$
Llama/I	NLL	0.05104	$0.04004\ [0.02547, 0.05460]$	$-0.01101\ [-0.03576, 0.01374]$
Llama/I	KL	0.04446	$0.03717\ [0.02367, 0.05068]$	$-0.00728\ [-0.02277, 0.00820]$

All eight H128 gain intervals are positive, including both Qwen libraries. All eight absolute-change point estimates are negative, but seven individual change intervals include zero. Model-averaged changes are $-0.00372\ [-0.01592, 0.00848]$ for Qwen and $-0.01685\ [-0.03323, -0.00048]$ for Llama. Different significance statuses do not establish a between-model difference, much less an effect of model size. The fresh results do not retrospectively make the earlier eight-document H128 gain significant, or show that sample size alone explains the difference between cohorts.

What this connects, and what it leaves open. The study directly tests horizon dependence of choice value, complementing the separate preservation measurements. It does not establish preservation as a causal mediator. Longer target windows average different content and change the actual mean surface, both selected routes and the oracle. Headroom and higher-order residuals can also affect regret. Nor do three nested horizons establish a universal or condition-wise monotone law. Complete surfaces are observed before projection, so these results establish hindsight choice relevance, not query-efficient optimization, native Quest policy improvement or generated task accuracy.

Execution and portable reproduction. The unchanged GB10 BF16/SDPA runtime and fixed 32-token FP64 readout chunks reproduce both models' opened qualification fingerprints before collection. Each fresh source has dense/no-op/restoration checks, mask audits on every batch and independent H32 runs for the first two and last actions in each library. Those representative sparse logits/endpoints exactly match H128 prefixes; this is not an independent H32 replay of all 81 fresh actions. The inherited cross-batch readout guard is unchanged (Appendix N).

All 160 tensors and 13,120 batch receipts complete; independent contrasts verify 2,880 cells. No interim effect governed continuation. Full `artifact.horizon_confirmation` requires token endpoints omitted from the review edition. Its `artifact.submission_replay` recomputes inference from document records; the root README maps replay/export code. This does not rerun token decomposition, models or historical exclusion checks.

P PROSPECTIVE NATIVE FEASIBLE-SET CHOICE

A stronger comparator, a separate confirmation. The registered four-band Quest audit fits full-cube projections and evaluates choices within fixed-weight sets. Its opened slice-unary control changes both the fitting measure and label budget, and some small slices are saturated (Appendix A.4). The separate eight-band pilot instead fits both degrees to the same 70 feasible actions, with ranks 8 and 28. Its opened GB10/H100 outcomes are retained in Appendix F.1. The following fresh study tests average incremental choice value against that aligned unary comparator; it neither replaces the earlier populations nor changes their verdicts.

Before any new model outcome, we freeze 160 documents, the complete action set, checkpoint, execution profile, exact choice arithmetic and sole primary. There is no alpha tuning, optimizer search, favorable-subgroup promotion, optional stopping, source replacement, sample extension or alternate-hardware rescue. H100 is selected as the available qualified collection environment after inspecting both development executions; that selection and the positive opened H100 result are disclosed, not treated as independent confirmation.

Population and execution. The independent units are 80 GovReport CRS-train documents and 80 PG-19 train-shard0 books, selected with document/anchor seeds 202609193241/202609193242. Known allocated, reserved and failed identities and raw/normalized document hashes, including public PG-19 validation/test overlap, are excluded. Separate code reconstructs the canonical corpus, selection and tokenizer windows before collection. This verifies disjointness from recorded allocations, not absence from pretraining or semantic similarity. GovReport's split differs from earlier validation-based pilots; this is a new fixed population, not a same-split replay.

The checkpoint is Llama-3.1-8B-Instruct, revision `0e9e39f249a1` (full revision in the artifact). Quest uses BF16/SDPA with actionwise FP32 head/KL readout, a 4,096-token prefix and 32 consumed-target states under a fixed teacher-forced history. Query-dependent supports and own-layer Q/K/V are recomputed; each execution's native distribution defines the KL reference. Layers 0/1 remain dense, page size is 16 and $K = 2{,}048$. Each KL row uses the state after consuming one of the 32 continuation tokens; the prefill's final prediction is not an additional measured row. These are state-aligned distribution comparisons, not 32 gold-NLL target scores. There is no generated-quality, gold-NLL or latency endpoint in this study.

Two H100 PCIe 80GB workers use independently loaded replicas and the frozen deterministic profile: TF32 disabled, highest FP32 matmul precision, cuBLAS workspace `:4096:8`, unchanged BF16 reduction setting and seed 2026083192. Both cards pass the same exact opened qualification before fresh collection. Static even/odd shards assign one corpus to each card; corpus summaries do not isolate corpus from card effects. Inference conditions on this fixed execution, with documents, not devices, as units. The released sanitized protocol records model/package revisions and arithmetic; device identifiers and host paths are not required for CPU replay.

Identical feasible set and measure. The eight layer bands have lengths $(4, 4, 4, 4, 3, 4, 3, 4)$, refining the original four bands on layers 2–31. A binary action either retains native support or swaps its boundary page in the chosen band, preserving K. The primary set is

$$\mathcal{A}_4^{(8)} = \{a \in \{0,1\}^8 : \textstyle\sum_i a_i = 4\}, \qquad |\mathcal{A}_4^{(8)}| = 70.$$

Both intercept-plus-unary and degree-two least-squares models use all 70 observations with the same uniform measure. Their ranks are 8/28, not 70; neither interpolates an arbitrary surface. These are finite-set polynomial projections, not the orthogonal product-lattice O1/O2 components of the complete Cartesian audit. Tying adjacent band choices recovers 16 coarse four-band programs; six lie in the primary set, so the collection contains 80 distinct actions per source. The remaining coarse actions are execution bridges, not extra labels for either primary fit.

Four changed bands need not mean the same number of changed layers. The prespecified secondary subset retains 40 actions changing exactly 15 layers; its fit ranks are 7/21. It is not a replacement feasible set selected from the outcomes, and cannot rescue the primary.

Choice arithmetic and inference. Each source's tokenwise native-reference KL matrix is fitted using the frozen centered binary design and multi-right-hand-side least-squares solve, then averaged

over the same 32 positions. Each projection selects the first exact lexicographic minimizer. Actual mean KL defines its regret against the best action in the same feasible set. No tolerance changes a tie or reselects a route. The estimand is

$$\mu = \tfrac{1}{2}\mathbb{E}_G[r_U - r_P] + \tfrac{1}{2}\mathbb{E}_{PG}[r_U - r_P].$$

The sole success criterion is a strictly positive lower endpoint of the two-sided 95% source-stratified percentile bootstrap interval: 50,000 draws, seed 202609193240, 80 source draws per corpus, with paired regrets carried together. No token or action is an additional independent observation. The original frozen $N = 160$ addresses pilot source heterogeneity and precision, not numerical-error cancellation across hardware. The equal-corpus Welch–Satterthwaite interval is a prespecified sensitivity, not a gate.

All 160 sources and all conditions completed before effect analysis. The primary gain is 2.0821×10^{-4} with interval $[0.9082, 3.8437] \times 10^{-4}$, satisfying the criterion. Welch sensitivity is $[0.5519, 3.6123] \times 10^{-4}$, also positive. Mean regret falls from 3.4785×10^{-4} to 1.3964×10^{-4}: a 59.86% ratio of means, not a mean of source-specific percentages or improvement over native Quest.

Table 49: Fresh native confirmation and prespecified secondary subset. All regret entries and intervals are in units of 10^{-4}. Only SAME70 determines confirmation; the changed-layer subset is secondary.

Set	U (10^{-4})	P (10^{-4})	Gain [95% CI] (10^{-4})	Reduction
SAME70	3.479	1.396	2.082 $[0.908, 3.844]$	59.86%
Changed15 (40)	3.448	1.232	2.216 $[1.033, 4.005]$	64.26%

Average gain, heterogeneous choices. The source gains are positive/negative/zero in 93/41/26 cases. Exact optimum hits increase only from 32/160 to 35/160: pairwise corrects 22 unary misses but introduces 19. Mean regret improvement is therefore distinct from either input-wise optimality or a confirmed hit-rate increase. The mean source-wise range-normalized regret falls $0.14240 \to 0.07215$; these are averages of within-source ratios, not ratios of the population means.

Table 50: Source heterogeneity for SAME70 and the prespecified secondary subset. Corpus rows are descriptive, not independent confirmations. Regrets/gains use units of 10^{-4}; hit counts and gain signs use document units.

Set	Corpus	N	U/P (10^{-4})	Gain (10^{-4})	Exact hits U/P	$+/-/0$
SAME70	All	160	3.479/1.396	2.082	32/35	93/41/26
SAME70	GovReport	80	4.208/1.345	2.863	15/15	45/21/14
SAME70	PG19	80	2.749/1.448	1.301	17/20	48/20/12
Changed15 (40)	All	160	3.448/1.232	2.216	20/35	94/35/31
Changed15 (40)	GovReport	80	4.179/1.242	2.936	10/20	47/17/16
Changed15 (40)	PG19	80	2.718/1.222	1.495	10/15	47/18/15

The largest positive source contributes 28.33% of total positive gain; the effect is not uniform. Precomputed equal-corpus leave-one-out means remain positive, ranging over $[1.3837, 2.2205] \times 10^{-4}$. These are descriptive point estimates, not leave-one-out confidence bounds. No source is removed. The artifact retains every source and the prespecified correction-direction, insufficient-correction and competing-minimum categories, including adverse rerankings. The positive 40-action secondary comparison does not change the primary decision.

Execution proof and reproduction. Every fine action repeats; all 16 tied coarse programs per source have exact bridges. Native/no-op, base-cache, support, own-KV and teardown checks pass. The complete collection has 12,800 action receipts and 2,560 coarse bridge checks, all bound to source, worker and execution hashes. CPU validation checks each finite 32-value FP32-representable KL vector, reconstructs all sources, and independently compares raw 0/1 and centered binary projection designs. Projection-identity tolerances do not change authoritative exact argmins. The prescribed local CPU analysis environment is fixed separately from the H100 execution: Python 3.14.6, NumPy

2.4.6, SciPy 1.18.0, aarch64 and one BLAS/OpenMP thread. There is no selection between analysis substrates.

The anonymous `artifact.native_confirmation` command replays saved endpoints, inference and tables with sanitized source/execution provenance. Full signatures and runtime attestations record checks that cannot be rerun from saved scalars alone; CPU replay does not reconstruct unavailable hidden states or certify the complete historical source registry independently. The earlier failed bridges and GB10/H100 pilots remain in the artifact.

Scientific scope. This result establishes prospective *average hindsight choice value* against feasible-set-aware unary under the specified Llama/H100 profile and source population. The entire surface is observed before either projection, so this is not cheap acquisition or an online routing algorithm. Both methods choose among four-of-eight-band perturbations; neither is shown to outperform unperturbed native Quest, whose self-KL is zero. The study does not establish hardware invariance, benefit on every input, another checkpoint, generated quality, or an internal cause of interactions. It also does not prove that the earlier GB10 pilot was uncertain solely because its sample was small.

Q A PROSPECTIVE CALIBRATION TEST

A new prospective policy test. After 16-document development, we froze a separate 80-document study: 40 GovReport CRS-train reports and 40 PG-19 books selected from train-shard0 *union* shard1 at the same pinned revision. The extension was specified before collection because only 19 eligible shard0 books remained. Fixed-order selection and independent source reconstruction exclude recorded allocations and exact/normalized document and prompt duplicates; all selected PG-19 books happen to lie in shard1. Freshness is relative to recorded history, not pretraining or semantic similarity. Neither the development cohort nor the earlier 160-document KL study's failed primary is relabeled or rescued.

Selection and execution. The fixed Qwen3-4B-Instruct-2507 and Llama-3.1-8B-Instruct checkpoints use 4,096-token prompts, four bands, $K = 1{,}456$, contiguous/interleaved support libraries, and fixed deterministic H100 execution. Probes consume a prompt-only dense greedy guide, beginning with the dense-prefill argmax; candidates follow that guide with private caches. Hybrid acquires candidates from the prompt-mass anchor by one forward greedy coordinate sweep over the four three-choice bands: query both alternatives at each current state and adopt the best only if first-token KL improves by more than 10^{-12}, yielding the anchor plus eight distinct alternatives. It then selects the best observed four-token mean KL; matched G4 uses that mean for both acquisition and choice, while direct G1/G4 controls repeat forward/reverse cyclic sweeps until their distinct-candidate cap or an unchanged sweep, returning the best observed action (lowest action ID within 10^{-12} of the minimum). No pairwise (P_2) surrogate is fitted. G1/G4 denote one/four guide steps: G1 candidates run serially, G4 candidates in four-token blocks. Forward/reverse denote band order; 11 is the candidate budget. All policies, libraries, and repetitions are committed before accessing each source's future 32 corpus targets. Evaluation then teacher-forces those targets: output 0 is common dense prefill, outputs 1–4 use the selected route, and outputs 5–31 return to the *original* anchor with one continuous private cache, without reset or splicing. All policies also receive a held-route evaluation.

Primary result. Both model runs completed all 80 sources and integrity checks before analysis. The sole primary averages mean32 dense-reference KL over both models and libraries within each source. The family-stratified, source-paired percentile bootstrap uses 20,000 draws (PCG64 seed 202609266901); cells, tokens, and repetitions add no independent units. Anchor minus returned-hybrid KL is **0.0043745 nats/token**, with two-sided 95% interval $[0.0014722, 0.0088637]$, satisfying the frozen lower-bound-above-zero criterion. Pooled KL falls from 0.0870795 to 0.0827050: **5.02%**, a ratio-of-means reduction, with positive gains on 56/80 sources. No sample extension or outcome-adaptive policy change was permitted.

Table 51: Complete returned-policy panel. KL gains and intervals are in 10^{-3} nats/token; positive favors the policy over anchor, or hybrid over the control, respectively. Only the anchor-row contrast is primary; other intervals are nominal, not simultaneous.

Policy	Anchor gain	Control − hybrid [95% CI]	Stage (s)
Anchor	0	$4.375\ [1.472, 8.864]$	0.000015
Hybrid G1/G4 choose9	4.375	—	0.5303
Matched G4/G4 choose9	4.591	$-0.216\ [-0.578, 0.143]$	0.5303
G1 forward converged	4.022	$0.353\ [-0.930, 2.209]$	0.8044
G1 reverse converged	3.944	$0.431\ [-1.014, 2.538]$	0.7601
G4 forward11	4.583	$-0.208\ [-0.638, 0.241]$	0.6041
G4 reverse11	4.107	$0.268\ [-1.106, 2.171]$	0.6038
G1 forward11	3.606	$0.768\ [-0.415, 2.558]$	0.5637
G1 reverse11	4.151	$0.224\ [-0.591, 0.984]$	0.5616

Measured cost and decision consequence. Stage times average two timed repetitions in randomized policy order, after discarded warmup. Copies, dense-guide/reference construction, readout, and selection are charged; common prefill/anchor and post-clock audit/storage are excluded. Hybrid costs 0.5303 s, 5.92%/5.57% less than G1 forward/reverse11; paired savings are 33.40 ms $[28.92, 37.53]$/31.28 ms $[27.39, 35.11]$ (nominal 95%). This supports a lower-calibration-cost candidate with positive anchor-relative fidelity. It is not TTFT or end-to-end speedup: relative to the anchor itself, calibration adds approximately 0.5303 s. The matched G4 method has essentially the same measured cost, with no resolved cost advantage for hybrid. Converged G1 continues coordinate search until convergence.

Scope of the positive claim. All seven hybrid–search mean32 KL intervals include zero: neither superiority nor equivalence/noninferiority to direct controls is established. Anchor-relative NLL gain is 0.00124 $[-0.00305, 0.00593]$ nats/token, so overall likelihood improvement is unresolved; the G1 forward11 NLL contrast is nominally positive (0.00303 $[0.00081, 0.00540]$), not a replacement primary. Held minus returned hybrid KL is -0.00213 $[-0.00799, 0.00373]$: significance of the returned anchor contrast does not establish a benefit of returning rather than holding. Thus the result demonstrates prospective fidelity improvement for a fixed policy selected without future targets, not causal lifetime benefit, generated task quality, or pairwise-model necessity. Prefix/anchor identity and deterministic-repeat checks pass; independent primary/cost arithmetic agrees within 5.56×10^{-17}. Artifact replay uses saved endpoints and measured times; it does not regenerate native model outputs.