How Active Readout Paths Shape KV-Cache Sensitivity

Anonymous authors
Paper under double-blind review

Abstract

Across 64 shared QASPER document caches on Qwen2.5-1.5B, changing the question produces a median absolute same-entry log-sensitivity change of 0.573. We ask what a score must observe to predict the effect of a fixed KV perturbation. Controlled query substitution isolates reader dependence; an audited tail case gives a 227.45-fold change with storage held fixed. Under radial perturbations, static vector magnitude gives a mean within-reader position-rank correlation of -0.051, squared current-reader routing reaches 0.747, and the first-order last-query response (S4) reaches 0.778. In an eight-document matched perturbation-family control, S4 falls from 0.780 under radial perturbations to 0.403 under unit-Gaussian directions. In a separate seven-document length control, it falls from 0.807 at 512 tokens to 0.580 at 8,192 tokens; routing remains close. Tracking the same entry across questions requires more path information: on a matched path sample, complete-suffix differentiation raises layer-3 tracking from 0.100 for the last-query downstream-Fisher diagnostic to 0.698, with little gain at layer 23. Actual joint deletion reveals a parallel depth pattern: measured-sensitivity selection reduces output divergence beyond routing at layer 3, while no tested score gives a resolved improvement over routing at layer 23. Current routing supports position ranking; broader path coverage improves early-layer tracking, and joint deletion locates a corresponding early-layer gap in output preservation.

1 Introduction

A language model can revisit the same cached document to answer several questions. The stored keys and values remain fixed, while each question changes the computation that reads them. Attribution and cache selection summarize these entries with scores based on vector magnitude, attention history, or the current query. Interpreting such a score requires knowing how the effect of a fixed perturbation depends on that reading computation.

In our natural-reader experiment, the same document entry is perturbed under four different questions. The median absolute change in its log sensitivity is 0.573 across the tested question pairs. Controlled query substitution isolates this dependence further by changing only the query activation while keeping the stored perturbation fixed. In the largest audited case, this produces a 227.45-fold sensitivity change (Fig. 1A). The two effects rank at the 1.46th and 99.69th percentiles of one fixed self-reader effect distribution.

How much of the reading computation must a score observe to predict that effect? Activation patching, causal tracing, and interchange interventions connect internal changes to model outputs (Meng et al., 2022; Geiger et al., 2022; Zhang & Nanda, 2024). Downstream Fisher pullbacks provide an established geometry for those changes (Martens, 2020; Wang et al., 2026). We use internal interventions and output-Fisher geometry to study paired reader changes at a fixed KV perturbation.

The attention-explanation debate concerns when attention weights support an explanation of a prediction (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Transformer analyses extend the question to transformed-vector norms and query–key feature interactions (Kobayashi et al., 2020; Kamath et al., 2025). Our target is the measured sensitivity to a specified KV perturbation. An attention weight is one factor in that response; alignment, value contrast, and downstream/output geometry determine what it carries to the output.

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.

KV-cache methods already exploit information about readers: H2O retains heavy hitters and recent tokens (Zhang et al., 2023), SnapKV uses an observation window (Li et al., 2024), and QUEST selects cache pages using the current query (Tang et al., 2024). KVzip uses context reconstruction to build reusable, query-agnostic caches (Kim et al., 2025). Query visibility itself can change method rankings (Luo et al., 2026). These studies establish query-dependent cache utility; we study single-entry local sensitivity and its attention-path structure. Appendix A develops these connections.

A score can observe different readers and different parts of their computation. We call these dimensions *reader access* and *path coverage*: which query is observed, and which stages and suffix positions of the KV response the score follows. Our contributions are:

1. We measure same-entry reader dependence in natural questions and isolate it with fixed-storage substitutions and controlled attention interventions.
2. We separate gains from current-reader access and richer local responses: routing nearly matches the last-query response in position ranking, across perturbation families and prefix lengths with different absolute recovery.
3. We test path coverage by restoring earlier suffix responses and their cross terms, recovering much of the early-layer tracking that the last-query diagnostic loses. Actual joint deletion follows the same depth asymmetry in output preservation: measured-sensitivity selection improves on routing at layer 3, with no resolved improvement in output preservation at layer 23 (Sections 5.2–5.3).

2 Measuring a Memory's Effect Under Different Readers

We measure the output effect of perturbing one historical key or value. Let p_r be the baseline next-token distribution under reader r, and let $p_r^{\pm}$ follow perturbations $\pm\epsilon\delta$ to the same stored entry. The measured KV sensitivity is

$$I_r^{(\epsilon)}(K,V,\delta) = \frac{D_J(p_r,p_r^{+}) + D_J(p_r,p_r^{-})}{2\epsilon^2}, \qquad D_J(p,q) = D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p). \tag{1}$$

Our primary experiments use $\epsilon = 0.01$; empirical summaries abbreviate this finite-scale target as I_r. For a smooth computation, the infinitesimal coefficient is

$$I_r^{(0)} = \lim_{\epsilon\to 0} I_r^{(\epsilon)} = \delta^{\top} J_{\mathrm{KV},r}^{\top} F(p_r) J_{\mathrm{KV},r}\delta, \qquad F(p) = \mathrm{diag}(p) - pp^{\top},$$

where $J_{\mathrm{KV},r}$ differentiates the full intervened computation. The local path analysis uses this coefficient; the perturbation-scale audits are reported in Appendices E and F.

The controlled experiment fixes (K,V,δ) and substitutes a naturally computed query activation from another prompt at the target layer. The donor query is inserted after query projection, while source-position RoPE, the source cache, and all model parameters remain unchanged. Downstream activations are recomputed under the substituted query. The source prefix is cached before either reader is evaluated. This intervention isolates the read operation while preserving the memory being read.

In the natural-reader experiment, a document is cached once and several questions are evaluated against independent clones of that cache. Each perturbation is applied before the complete question suffix is run, so the reader condition includes the question's full computation. Across either design, the same-entry effect change is

$$\Delta_{\mathrm{pair}} \log I = \log I_B - \log I_A. \tag{2}$$

The paired difference isolates reader-dependent sensitivity changes at a fixed stored entry and perturbation.

3 From Fixed Storage to a Changing Effect

3.1 Controlled substitutions at fixed storage

The controlled comparison separates a change in the reader from a change in the memory. Among 340 strict-audit candidates, the largest change occurs for a historical key at layer 23 and cache

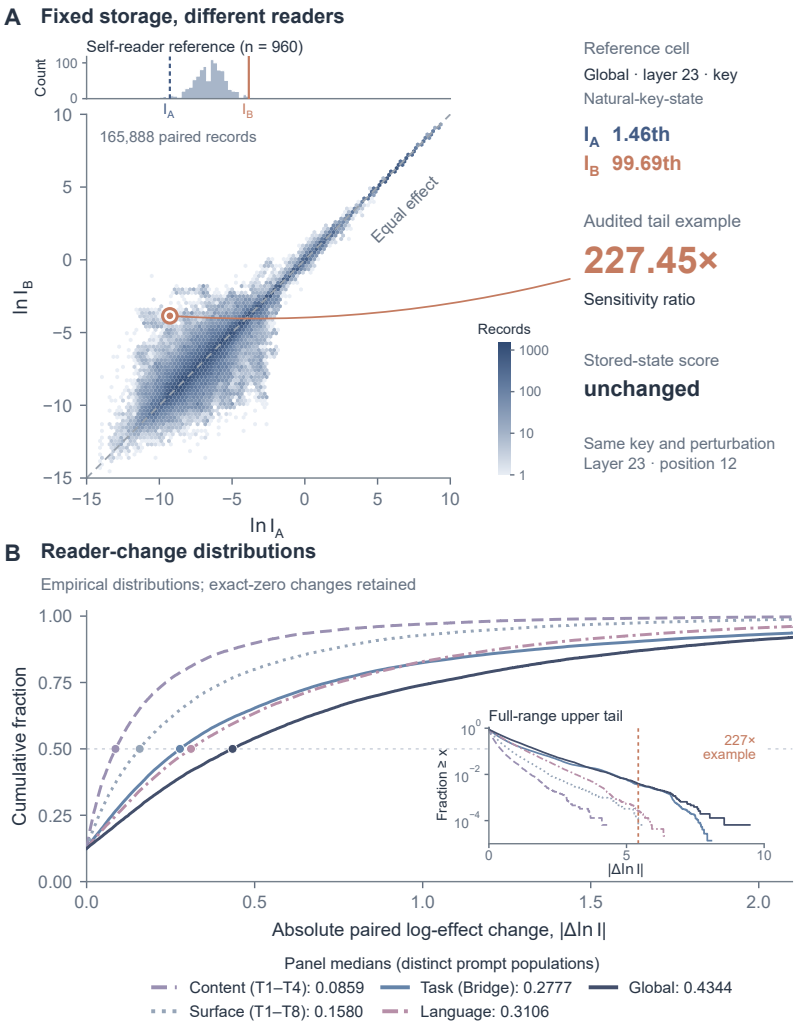

Figure 1: **The same stored KV perturbation can have different effects under different readers. A,** All 165,888 pairs fix storage and perturbation. Source-reader effects are horizontal; substituted-reader effects are vertical. Hexagon colors show logarithmic record counts; the diagonal denotes equality. The top histogram is the one-dimensional distribution of 960 self-reader reference effects (global, layer 23, key, natural-key-state), not the full hexbin marginal. Both I_A (1.46th) and I_B (99.69th) are evaluated on its common effect axis; I_B is the vertical coordinate in the main plot. The $227.45\times$ tail example is the largest change among 340 strict-audit candidates. **B,** Paired log-effect magnitude CDFs include zeros; dots mark medians and the inset shows the full tail. Core Content covers T1–T4; Surface covers T1–T8 and preserves task, proposition, and language. Bridge instruction-head changes hold context and language fixed within each pair; the panel spans 32 numeric variants of one transfer-context template. Panel populations differ (Appendices C and D).

position 12 (Fig. 1A). Its measured sensitivity is 9.3474×10^{-5} under reader A and 2.1260×10^{-2} under reader B, a 227.45-fold change. Against the same 960-effect self-reader reference, the entry moves from the 1.46th to the 99.69th percentile. The reference fixes panel, layer, channel, and perturbation direction type (Appendix C). The KV-only predictor stays at the 11.4th percentile of its static reference population while the measured effect changes.

Fixed-input reader dependence also appears in controlled experiments on Qwen2.5-0.5B, Qwen2.5-1.5B, and GPT-2 small (Qwen Team, 2025; Radford et al., 2019). To connect this activation substi-

tution to changes in a reader's request, we compare surface wording and instruction heads within controlled prompt sets.

3.2 Reader changes within Core and Bridge

The reader can change even when the facts it is asked to read stay fixed. Figure 1B summarizes the content, surface, Bridge, language, and global reader-change panels. Within the Controlled Core, surface changes preserve semantic instance, task, and language, giving a pooled-record median absolute same-entry log-effect change of 0.1580. Within Bridge, instruction-head changes over a fixed context and language give a pooled-record median of 0.2777. Bridge uses 32 numeric variants of a single transfer-context template. These are separate within-population controls, with different contexts, task families, and prompt lengths.

The Bridge aggregate combines instruction pairs with different effect changes. B1 asks for an initial count, B2 a post-transfer count, B3 a comparison, and B4 structured output. The median of reader-pair group medians is 0.1303 for the two counting heads (B1–B2). The five pairs involving comparison or structured output give 0.3153–0.3728 (Fig. 2A; Table 5).

Within the shared Core T1–T4 inventory, Surface gives a median shift of 0.1552 and Content 0.0859. Semantic preservation alone does not order these changes. Table 6 gives the construction and family breakdown.

4 Which Reader Information Recovers Position Rankings?

The controlled comparisons show that a fixed entry's effect changes with the reader. Whether observing that reader improves position ranking is a separate empirical question. We start from the local response.

4.1 Four factors shape the local response along the attention path

For one attention head, $o = \sum_j a_j v_j$, where $a = \mathrm{softmax}(qK^\top/\sqrt{d})$. A small value perturbation changes its output by $\delta o_V = a_i \delta v_i$; a key perturbation changes it by

$$\delta o_K = \frac{a_i}{\sqrt{d}}(q^\top \delta k_i)(v_i - o).$$

Query alignment and value contrast enter the key response, while routing scales both responses. Downstream computation converts either response into a change in the final distribution.

Let J_{down} map the intervention-layer head output to final logits, including W_O, residual addition, later blocks, final normalization, and the language-model head. Its output-Fisher pullback is

$$G_{\mathrm{down}} = J_{\mathrm{down}}^\top F(p_r) J_{\mathrm{down}}. \tag{3}$$

This downstream/output geometry combines propagation through J_{down} with the reader-specific output metric $F(p_r)$ (Martens, 2020; Wang et al., 2026). For the single-query, single-head path, the local coefficients are

$$I_V^{(0)} = a_i^2 \|\delta v_i\|^2_{G_{\mathrm{down}}}, \tag{4}$$

$$I_K^{(0)} = \frac{a_i^2}{d}(q^\top \delta k_i)^2 \|v_i - o\|^2_{G_{\mathrm{down}}}. \tag{5}$$

Targeted interventions test how each factor changes measured sensitivity.

Query alignment. Equal-norm key perturbations in the active-query row space yield a median 3.05×10^5 sensitivity ratio over paired numerical-null directions (256 pairs; 95% prompt-cluster CI $[1.80, 5.73] \times 10^5$). Thus equal-size key perturbations can differ in their visibility to the active query.

Attention routing. For value perturbations, the reader can change the response strength while the stored perturbation stays fixed. Across 256 paired comparisons, $\Delta \log a_i^2$ and measured $\Delta \log I$ have Spearman correlation 0.686; the through-origin slope is 0.962 (95% CI [0.869, 1.053]; Fig. 2B). Observing the current attention weight therefore connects a fixed value perturbation to its changing effect.

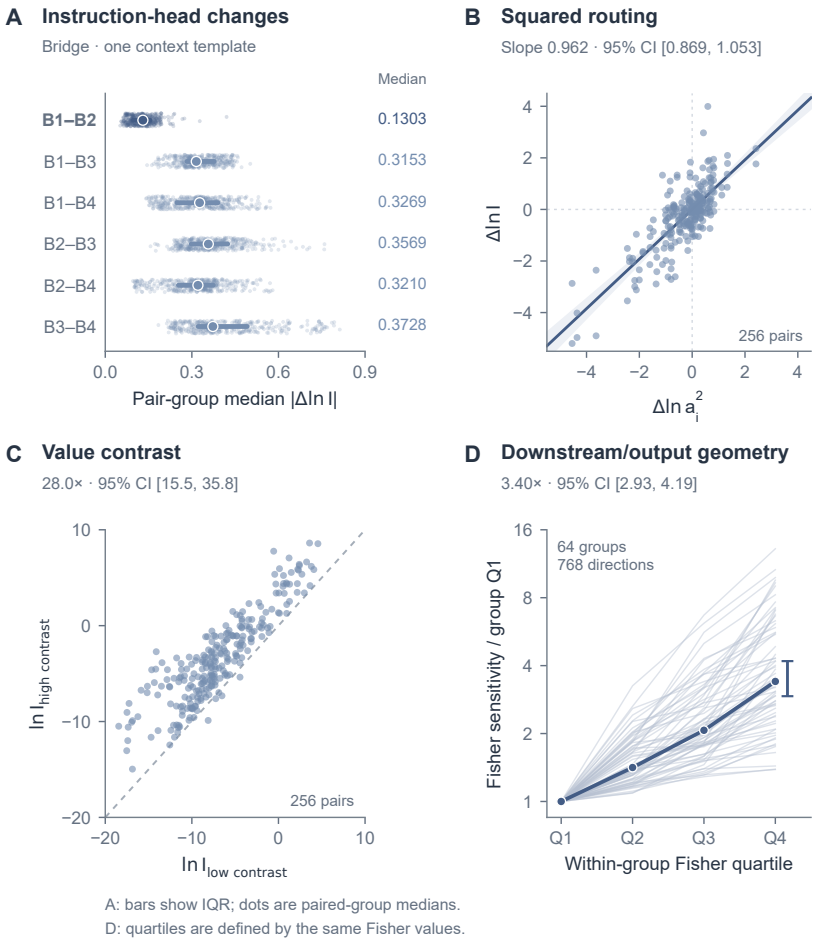

Figure 2: **Controlled interventions probe factors that shape the local KV response. A,** Bridge instruction-head pairs, using 32 numeric variants of one context template. Small points summarize 32 effect records for each of 384 reader-pair groups by their median; large points and bars show the across-group median and interquartile range. B1–B2 (the two count queries) gives 0.1303, compared with 0.3153–0.3728 for the other pairs within this fixed construction. **B,** Equal-norm value perturbations: 256 paired changes in sensitivity and squared routing. The through-origin slope is 0.962; shading shows its 95% CI [0.869, 1.053]. **C,** Paired low- and high-contrast sensitivities at a fixed reader and key perturbation; the diagonal denotes equal effects. The median paired ratio is $28.0\times$ (95% prompt-cluster CI [15.5, 35.8]; 256 pairs). **D,** Directions are sorted by downstream/output Fisher sensitivity within each group; quartile medians are normalized by that group's Q1. Thin lines show all 64 groups (768 equal-norm directions); the thick line shows their median. The Q4/Q1 median is $3.40\times$ (95% group-bootstrap CI [2.93, 4.19]). The alignment intervention and full protocols appear in Appendices D and E.

Value contrast. Holding the reader, key perturbation, and attention weight fixed while moving the cached value between low- and high-contrast conditions changes measured sensitivity by a median factor of 28.0 (95% prompt-cluster CI [15.5, 35.8]).

Downstream/output geometry. Equal-norm directions injected before W_O differ in categorical-Fisher sensitivity through the remaining network. Across 64 prompt/layer groups, the median highest-to-lowest quartile ratio is $3.40\times$ (95% group-bootstrap CI [2.93, 4.19]).

Contrast effects include the accompanying change in baseline head output and any change in downstream geometry (Appendix E). We next ask how much each factor adds when a natural question reads the cache in a larger model.

4.2 From stored magnitude and read history to current routing

Qwen2.5-1.5B base reads four natural questions over each of 64 shared QASPER document caches (Dasigi et al., 2021). We perturb all 512 retained-prefix positions in FP32 using radial scaling, $z_i \mapsto (1 \pm 0.01)z_i$. Each question and answer cue form a 9–32-token suffix, run against an independent clone of the document cache after the perturbation is applied. Across question pairs sharing the entry, layer, channel, and perturbation, the median absolute log-effect change is 0.573.

For a given question, which positions are most sensitive? We rank the 512 positions within each of 4,096 document/reader/layer/channel sets. Static vector magnitude (S0) gives a mean Spearman correlation of -0.051. Document-prefill attention history (S1), which records how the document was read before the question arrived, reaches 0.283. The active question supplies a much stronger signal: squared current-reader routing (S2) reaches 0.747 (Fig. 3A).

S2 changes which reader the score observes: the final suffix token rather than document-prefill history. Squaring its attention weights follows the local coefficients in Eqs. 4–5. This reveals where the current reader attends, before accounting for what the values contribute to the head output.

4.3 Adding the gate and last-query response

We next restore the remaining attention-local structure. For the final suffix token T, let a_{hi} be the attention weight on document position i in query head h, and let $\delta o_{T,h,i}$ be the first-order head-output response to the chosen KV perturbation. Then

$$S_2(i) = \sum_h a_{hi}^2, \qquad S_4(i) = \left\| W_O \,\mathrm{concat}_h\big(\delta o_{T,h,i}\big) \right\|_2^2 . \tag{6}$$

The head responses use the value or key expressions in Section 4.1, with each query head reading its shared GQA key/value head. The analytic gate S3 adds the direction norm for values or the query projection and softmax response for keys. The first-order last-query response S4 further includes value contrast and mixing through W_O, measuring the change in post-projection residual coordinates. All scores use the unperturbed model state, with no coefficients fitted to the QASPER effects.

The analytic gate reaches .755 and S4 .778 [.772, .784]. Moving from S1 to S2 adds .464; extending the last-query local response from S2 to S4 adds .031. Both still omit earlier suffix paths. The split check gives S4 = 0.777 on the previously analyzed 32-document Split A and 0.779 on the prospectively frozen, non-overlapping 32-document Split B.

Alignment, contrast, and downstream geometry each change sensitivity substantially under controlled interventions, yet S4's attention-local additions improve natural position ranking by only .031 over current routing. Mechanistic relevance and incremental predictive value are therefore separate grounds for including a factor.

4.4 Perturbation family and prefix length

The perturbation specifies what a reader-dependent score must predict. Two separate matched comparisons change the ranking task: the perturbation family and the document-prefix length (Fig. 3B).

In the direction control, eight documents, eight layers, both channels, and all 512 positions are matched across 512 decision sets. Radial directions retain stored-vector norms; tuple-seeded Gaussian directions have unit norm, so both orientation and magnitude vary.

On this sample, S4 falls from 0.780 under radial scaling to 0.403 under unit-Gaussian directions. S2/S3 fall from 0.747/0.755 to 0.391/0.398. The S4–S2 increment decreases from 0.033 [0.031, 0.035] to 0.011 [0.010, 0.013]. Routing's near match with S4 persists as both lose predictive strength.

The length comparison uses seven long documents, layers 3 and 23, and 64 stratified positions per prefix (512-token S4 baseline: 0.807; the primary 0.778 pools 64 documents, eight layers, and all

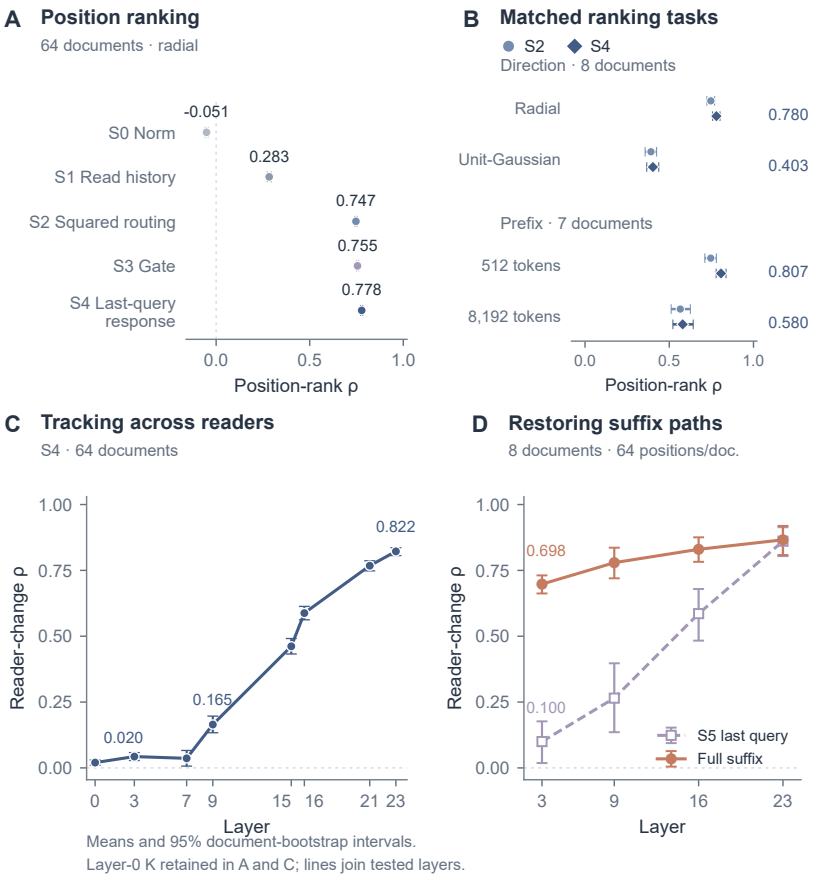

Figure 3: **Current routing supports position ranking; restoring suffix paths improves early-layer tracking. A,** Primary radial position ranking: S2 recovers most of the signal captured by S3/S4. **B,** Separate matched perturbation-family (eight documents, eight layers, all 512 positions) and length (seven long documents, layers 3/23, 64 stratified positions per prefix) comparisons. Labels give S4 estimates. Radial directions have stored-vector norms; Gaussian directions have unit norm. **C,** S4 effect-change tracking remains weak through layer 9 and rises sharply later. **D,** On eight documents and 64 fixed positions per document, complete-suffix differentiation adds earlier responses and cross terms to the last-query downstream-Fisher diagnostic S5. Both series track the same finite radial-sensitivity target. Means and 95% document-bootstrap CIs are shown; lines join tested layers. A/C use all 64 documents and retain layer-0 K (Appendix F.1). Full path estimates and increments appear in Table 17.

512 positions). Extending the prefix to 8,192 tokens on the same documents and questions lowers S2/S4 from 0.746/0.807 to 0.565/0.580 (Table 19). Each length ranks positions sampled across its own prefix. The S4–S2 increment narrows from 0.061 to 0.014. Both comparisons show that recovery strength depends on the ranking task, while routing captures most of the signal recovered by the fuller response.

5 What the Last Query Cannot See

5.1 Last-query tracking weakens in early layers

A position ranking need not track one entry across questions. Holding each entry fixed, we correlate predicted and measured log-sensitivity changes. S4 gives a mean document-level correlation of 0.440 [0.425, 0.454], versus 0.436 for squared routing and 0.434 for the analytic gate. The richer local response barely changes this aggregate.

The layer breakdown exposes a much larger difference (Fig. 3C). S4 gives correlations of 0.020, 0.043, 0.036, and 0.165 at layers 0, 3, 7, and 9, then 0.462, 0.588, 0.768, and 0.822 at layers 15, 16, 21, and 23. Each estimate averages 64 document-level correlations; the frozen layer schedule and 95% intervals from 5,000 document-cluster bootstrap replicates appear in Table 12.

With the scoring rule, radial protocol, and questions unchanged, tracking remains weak through layer 9 and rises sharply in the later tested layers. The aggregate correlation of 0.440 conceals this transition.

5.2 Restoring omitted paths recovers early-layer tracking

The complete cache intervention reaches the output through every suffix position that reads the perturbed entry:

$$g_i = \sum_{t=1}^{T} J_{\mathrm{post,t}} \delta r_{t,i}, \qquad I_i^{(0)} = g_i^\top F(p_r) g_i. \tag{7}$$

Here $\delta r_{t,i}$ is the post-W_O response at position t, and $J_{\mathrm{post,t}}$ propagates it to the final logits. Write $g_{T,i} = J_{\mathrm{post,T}} \delta r_{T,i}$, $g_{<T,i} = g_i - g_{T,i}$, and $F = F(p_r)$. Then

$$I_i^{(0)} = \underbrace{g_{T,i}^\top F g_{T,i}}_{S_5(i)} + 2 g_{T,i}^\top F g_{<T,i} + g_{<T,i}^\top F g_{<T,i}. \tag{8}$$

The first term is the last-query downstream-Fisher diagnostic S5. The other terms contain earlier suffix responses and their interaction with the final-query response. S4 observes an earlier stage still: the unpropagated squared norm of $\delta r_{T,i}$. The complete derivative demonstrates recovery with broader path access. It uses downstream computation as a measurement instrument, not as a candidate cache-selection score. We test this omission on 64 prospectively fixed positions per document, using the eight S5 documents and all four readers.

Relative to S5, the complete derivative raises layer-3 tracking from .100 to .698, and layer-9 tracking from .265 to .779. The paired gains are .598 [.529, .679] and .514 [.386, .654], against .244 [.169, .327] at layer 16 and .005 [-.001, .010] at layer 23 (Fig. 3D; Table 17). Earlier paths recover much of S5's layer-3 deficit, with much smaller gains at layer 23. Matched-layout 1.5B/7B runs reproduce this asymmetry, resolving small layer-23 gains (Appendix G).

The depth contrast persists against the differentiated target $I_i^{(0)}$, where S5 tracks reader changes at .137 in layer 3 and .990 in layer 23. Finite differences retain predictable signal (.698 at layer 3), but median $I^{(0)}/I^{(.01)}$ ranges from .253 to .441, with worse agreement at smaller steps (Table 18). The differentiated target separates the path-coverage finding from this magnitude mismatch.

5.3 From sensitivity rankings to joint deletion

The last-query view raises a practical question: does it also leave room to improve output preservation when entries are removed? At a 5% budget, squared current-reader routing captures 81.9% of measured sensitivity mass, versus 49.0% for prefill history and 3.8% for static norms (uniform random: 5.1%). S4 captures 83.0%, adding just 1.1 percentage points over routing (Table 1).

Table 1: Mean sensitivity mass (%) captured by selecting 26 of 512 positions in each primary radial decision set. Oracle ranks measured sensitivity; uniform random selection has expectation $26/512$. Full budgets appear in Table 15.

S0 norm	S1 history	S2 routing	S4 response	Random	Oracle
3.8	49.0	81.9	83.0	5.1	91.2

On eight documents, joint deletion retains 26 of 512 K/V positions at layer 3 or 23 and recomputes the suffix, leaving other layers intact. Selectors are S2, $S_{4,K} + S_{4,V}$, and measured $I_K + I_V$. The last is an additive oracle formed from measured single-entry sensitivities under the intact cache, not an oracle for the output after joint deletion. Averaging the two layers, output Jeffreys divergence is .0231 for S2 versus .1719 for uniform random retention (lower is better; Table 20).

Layer 3 leaves room for better selection. Ranking by measured local sensitivity reduces divergence from .0378 for S2 to .0272, a paired reduction of .01060 [95% CI .00635, .01510]. At layer 23, local-sensitivity selection (.00848) gives no resolved improvement over S2 (.00843). S4 gives no resolved improvement over S2 at layer 3 and is slightly worse at layer 23: the divergence increase is .00038 [.00009, .00068].

Measured local sensitivity is obtained from finite single-entry cache interventions that recompute the complete suffix, and therefore reflects earlier suffix paths omitted by S2 and S4. Its resolved layer-3 advantage and unresolved layer-23 gain follow the Full–S5 depth pattern (Section 5.2), though oracle access and path coverage are not separable here.

6 Discussion

Which property matters most depends on the question asked of a score. Position ranking rewards reader access: observing the current query's routing recovers most of the signal captured by richer local descriptions. Tracking a fixed entry across questions gains substantially from restoring earlier suffix paths, and joint deletion shows the same asymmetry in output preservation. Good position ranks alone settle neither how well a score tracks changing effects nor how well it preserves the output under deletion.

For cache design, current routing is a useful starting point when queries are available (Zhang et al., 2023; Li et al., 2024; Tang et al., 2024). At layer 3, richer local responses still leave deletion headroom. Exact path propagation incurs downstream computation; the design challenge is to approximate information missed by last-query scores cheaply enough for cache selection. Whether such a score closes the deletion gap without measured effects remains open.

7 Limitations

The primary natural-reader panel uses Qwen2.5-1.5B base on 64 document prefixes truncated to 512 tokens, with a matched eight-document 7B replication (Appendix G). The 8,192-token follow-up covers seven documents and two layers. Other families and longer contexts remain untested in these natural-reader panels. Mechanism interventions center on Qwen2.5-0.5B; fixed-storage substitutions are also tested in Qwen2.5-1.5B and GPT-2 small. Those substitutions can combine cache and query activations outside their natural joint distribution. The 32-observation GQA controls are descriptive (Appendix E). Joint deletion changes one layer at a time; it does not evaluate a complete compression policy or answer accuracy.

The finite target and its infinitesimal coefficient differ materially: median $I^{(0)}/I^{(.01)}$ is .266/.441/.253/.309 at layers 3/9/16/23; smaller steps reduce reader-change agreement. FP32 repeatability does not ensure a calibrated derivative (Table 18). A one-document, one-reader audit gives all-logits/final-logits magnitude ratios of 3.4–4.2, suggesting that output layout may contribute to the finite-to-infinitesimal gap. Position ranks are more stable than magnitudes, with Spearman correlations of .858–.943 across four 32-position layer/channel cells, but are not layout invariant (Appendix G). Frozen and matched-layout results remain separate. Perturbation-family comparisons also vary direction norms.

Bridge is a controlled prompt set with 32 numeric variants of one context template. Its v0.1 reference answers contain a typed-transfer accounting inconsistency; these answers are neither model inputs nor measurement targets. We interpret the panel as instruction-text dependence (Appendix D.1).

8 Conclusion

Predicting the effect of a fixed KV perturbation requires information about the reader that consumes it. Squared current-reader routing recovers most of the position-ranking signal captured by the tested local descriptions. The quality of this ranking varies across perturbation families and prefix lengths. Tracking the same entry across questions exposes information omitted by the last-query diagnostic; restoring earlier suffix paths recovers much of the lost tracking at the early tested layers. Joint deletion locates a corresponding early-layer gap in cache selection. The relevant unit of explanation is the stored state together with the active path that consumes it.

AI USE STATEMENT

Generative AI assisted with code development, data analysis and interpretation, and manuscript revision, and provided input on research questions and experimental design. Reported numerical results come from model runs and statistical analyses. The authors take responsibility for the final content.

ETHICS STATEMENT

The study analyzes publicly available pretrained language models, a public research-paper question-answering dataset, and synthetic prompts. It does not collect new human-subject data or evaluate decisions about individuals. The accompanying derived results are provided subject to source-model and dataset licenses, with identifying local-execution details excluded from the release view.

REPRODUCIBILITY STATEMENT

The estimands, interventions, model families, frozen hyperparameters, decision-set construction, and uncertainty procedures are specified in the main text and Appendices B–F. The supplementary material contains frozen measurements, statistical summaries, and numerical audit results, with a guide to the included files. Implementation and analysis code will be made publicly available upon acceptance.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of EMNLP*, pp. 4895–4901, 2023. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/.

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/.

Alessio Devoto, Maximilian Jeblick, and Simon Jégou. Expected attention: KV cache compression by estimating attention from future queries distribution. *arXiv preprint arXiv:2510.00636*, 2025. URL https://arxiv.org/abs/2510.00636.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Harish Kamath, Emmanuel Ameisen, Isaac Kauvar, Rodrigo Luger, Wes Gurnee, Adam Pearce, Sam Zimmerman, Joshua Batson, Thomas Conerly, Chris Olah, and Jack Lindsey. Tracing attention computation through feature interactions. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attention-qk/index.html.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W. Lee, Sangdoo Yun, and Hyun Oh Song. KVzip: Query-agnostic KV cache compression with context reconstruction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5585. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4eaa4b8f2d08edb3f0af990d56134ea-Abstract-Conference.html.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of EMNLP*, pp. 7057–7075, 2020. doi: 10.18653/v1/2020.emnlp-main.574. URL https://aclanthology.org/2020.emnlp-main.574/.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a452a7c6c463e4ae8fbdc614c6e983e6-Abstract-Conference.html.

Daming Luo, Christy Liang, and Junyu Xuan. How query visibility changes KV-cache compression rankings: A matched-budget audit. *arXiv preprint arXiv:2607.11942*, 2026. URL https://arxiv.org/abs/2607.11942.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://www.jmlr.org/papers/v21/17-678.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115v2*, 2025. URL https://arxiv.org/abs/2412.15115v2.

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

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of ACL*, pp. 2931–2951, 2019. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

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, 2024. URL https://proceedings.mlr.press/v235/tang24l.html.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. FishBack: Pullback Fisher geometry for optimal activation steering in transformers. *arXiv preprint arXiv:2605.17231*, 2026. URL https://arxiv.org/abs/2605.17231.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract.html.

A Full Related-Work Context

A.1 Intervention-based mechanistic interpretability

Causal tracing, activation patching, interchange interventions, and related methods estimate how internal components affect model outputs (Meng et al., 2022; Geiger et al., 2022; Zhang & Nanda, 2024). These methods motivate our direct internal perturbations. We follow the measured sensitivity to a *fixed* KV perturbation across readers and ask which local descriptions predict it, under which perturbation families and at which layers.

A.2 Representation summaries and local geometry

Integrated Gradients attributes model predictions to input features (Sundararajan et al., 2017). Fisher geometry characterizes local changes in predictive distributions (Martens, 2020; Wang et al., 2026). For a downstream map f, the pullback relation

$$M_z = J_f(z)^\top M_y J_f(z)$$

already shows that output-relevant geometry depends on the downstream computation. FishBack explicitly pulls the output Fisher metric through the downstream Jacobian and uses the induced geometry for activation steering. We use this relation as established framework rather than a new geometric derivation. Our experiments connect fixed-memory reader dependence to controlled interventions on the active path and recovery under natural questions.

A.3 Query-aware attention attribution

Attention-weight studies test whether weights explain output behavior (Jain & Wallace, 2019; Serrano & Smith, 2019). Transformer analyses incorporate transformed-vector norms (Kobayashi et al., 2020). QK attribution traces attention patterns to query-side and key-side feature interactions (Kamath et al., 2025). We ask how a fixed KV perturbation's measured sensitivity changes with the reader. The attention-explanation debate also makes the explanatory question and evaluation protocol explicit (Wiegreffe & Pinter, 2019). Under natural variation, we test which local descriptions recover position ranking and fixed-entry effect changes, and how their predictive strength depends on perturbation family and depth.

A.4 KV-cache selection and compression

H2O, Scissorhands, SnapKV, Expected Attention, and related methods use attention history, observation-window queries, or expected future readers to compress KV caches (Zhang et al., 2023; Liu et al., 2023; Li et al., 2024; Devoto et al., 2025; Tang et al., 2024). QUEST directly models query-dependent token criticality; KVzip instead builds query-agnostic reusable caches, and a matched-budget audit shows that query visibility can change compression rankings (Kim et al., 2025; Luo et al., 2026). These methods establish query-dependent cache utility as context for our study of single-entry causal effects.

B Complete Experimental Protocol

B.1 Models, checkpoints, and frozen artifacts

The local intervention experiments use Qwen2.5-0.5B and Qwen2.5-1.5B (Qwen Team, 2025), and GPT-2 small (Radford et al., 2019). Analyses use the frozen outputs provided in the supplementary material. Controlled query-swap, factor, mechanism, and final QASPER analyses use CUDA, seed 1729, FP32 model/cache arithmetic, and primary $\epsilon = 0.01$; the GPT-2 export uses its recorded 64-token, two-position configuration. QASPER evaluates an unperturbed batch with exactly the same batch shape as each perturbation batch, then forms log-softmax, Jeffreys divergence, and symmetric $+\epsilon/-\epsilon$ accumulation in FP32.

Model and tokenizer version identifiers, together with recorded software and hardware settings, will accompany the code release.

Observed repeatability. The frozen FP32 controlled task run includes three separately captured source occurrences for each of 1,536 unique prompts. The runner recomputes unmodified-reader sensitivity for each occurrence with the same direction, position, layer, and channel. The 24,576 unique prompt/entry measurements yield 73,728 repeat comparisons, with median and maximum absolute log-effect differences both zero. This records repeatability under the frozen configuration.

B.2 AUTOREGRESSIVE CACHE INTERVENTION

In the controlled panels, the model processes all tokens except the final token and stores the resulting prefix KV cache. The final token is evaluated against this source cache. One historical key or value at a selected layer and cache position is perturbed by $+\epsilon\delta$ and $-\epsilon\delta$; the prefix is not recomputed. The primary target is Eq. 1. The symmetric construction cancels the leading odd-order finite-difference term. Jensen–Shannon divergence is computed from the same forward outputs.

B.3 CONTROLLED READER SUBSTITUTION

The source prompt supplies the KV cache and perturbation. A second prompt supplies a naturally computed post-q_{proj} query activation at the target layer. This activation is inserted into the source computation while source-position RoPE, source KV tensors, perturbation, projection weights, and all other parameters remain fixed. GPT-2 uses the corresponding post-c_{attn} query activation and has no RoPE. The operation is a controlled cross-context composition and may produce query–cache combinations outside the natural joint activation distribution. The query is captured before RoPE and rotated at the source position after insertion: the donor's final-token RoPE phase is not transplanted. Its raw query can still reflect length and position through the preceding donor computation. The separate RoPE probes explicitly alter the applied phase and therefore do not measure the phase change of an ordinary query swap.

B.4 REPRESENTATION-ONLY CONTROLS AND OPERATOR LADDER

Representation-only predictors receive stored-state information but no active query, attention weight, or downstream feature. They include global and K/V-typed L2 summaries, positive diagonal metrics, shared low-rank variants, fixed metadata features, and frozen nonlinear regressors. The GPT-2 capacity control additionally receives 1,536 squared K/V coordinates. Its aggregate Ridge uses standardized features and $\alpha = 10$. The coordinate GBM is a fixed `HistGradientBoostingRegressor` with learning rate 0.05, 300 iterations, 15 maximum leaf nodes, 20 minimum samples per leaf, L2 regularization 1.0, and seed 1729.

The operator ladder restores perturbation magnitude, query magnitude, query–key logit change, attention routing, softmax-Jacobian sensitivity, and the squared attention-output response before o_{proj}:

$$KV \to +Q \to +QK^{\top} \to +A \to +J_{\mathrm{softmax}} \to +V.$$

Coefficients are fit on an independent calibration split; evaluation prompts and held-out layers are excluded from fitting. This calibrated controlled ladder is not fitted to QASPER. The QASPER all-position panel uses the coefficient-free ladder in Appendix F: S3 retains query alignment and attention sensitivity, S4 adds value contrast and output projection, and S5 adds last-query downstream-Fisher geometry on the frozen diagnostic subset.

B.5 STATISTICS

We report median absolute reader change, log-RMSE, Spearman correlation, sign agreement, calibration slope, and relative error reduction. Sign agreement is

$$n^{-1} \sum_{i} \mathbf{1}[\mathrm{sign}(\Delta_i^{\mathrm{true}}) = \mathrm{sign}(\Delta_i^{\mathrm{pred}})].$$

For controlled panels, with $y_{sr} = \log \max(I_{sr}, 10^{-12})$ and the same floor applied to predictions, the level error is

$$\mathrm{RMSE}_{\mathrm{level}} = \left[\frac{1}{2N} \sum_{s=1}^{N} \sum_{r \in \{A,B\}} (\widehat{y}_{sr} - y_{sr})^2 \right]^{1/2}.$$

Both readers enter the sum; the held-out summary restricts it to natural directions at layers 3, 9, 15, and 21. Paired tracking is $\rho(\widehat{y}_{sB} - \widehat{y}_{sA}, y_{sB} - y_{sA})$, a rank statistic whose magnitude does not quantify numerical calibration. Relative level log-RMSE reduction is $(\mathrm{RMSE}_{\mathrm{static}} - \mathrm{RMSE}_{\mathrm{candidate}})/\mathrm{RMSE}_{\mathrm{static}}$. Controlled query-swap summaries use prompt-cluster bootstrap; calibration slopes use pair-index bootstrap. QASPER uses document-cluster bootstrap. Small GQA/RoPE panels are descriptive. Static reader change is exactly zero by construction, and pooled reader-variance shares are never interpreted as paired-change effect sizes.

C Strict Counterexample Audit

The audit operates on the frozen 165,888-row controlled table. Every row uses a naturally computed, non-random key or value perturbation direction. The static score is **prediction_static**, the K/V-typed L2 predictor fit on an independent calibration split. Percentiles use average ranks, with larger values denoting greater importance. Static percentiles are computed within (panel, channel, natural direction); realized percentiles are computed within (panel, layer, channel, natural direction) for self and swapped readers.

A record satisfies the frozen predicate iff

$$P_{\mathrm{static}} \le 0.25, \qquad \max(P_{\mathrm{self}}, P_{\mathrm{swap}}) \ge 0.90, \qquad |\log I_{\mathrm{swap}} - \log I_{\mathrm{self}}| \ge \log 5.$$

Thus the event combines a bottom-quartile static score, a top-decile realized effect for at least one reader, and at least a five-fold symmetric reader ratio.

Table 2: Strict-audit counts and top-decile membership flips in the complete frozen table.

Panel	Records	Strict candidates	Top-10 flips
Content	15,360	5	418
Surface	15,360	2	714
Bridge	73,728	90	4,716
Language	46,080	154	3,654
Global	15,360	89	1,446
Total	165,888	340	10,948

Table 2 gives the complete panel counts. Of the 340 strict candidates, 331 are key interventions and nine are value interventions; 322 (94.7%) also flip top-decile membership. The panel counts are 154 language, 90 Bridge, 89 global, five content, and two surface cases. Table 3 summarizes the strict-candidate distribution.

Table 3: Distribution of the 340 strict candidates. Rank shift is the absolute difference between self- and swap-reader realized percentiles.

Quantity	Min	Q25	Median	Q75	Q95	Max
Reader-effect ratio	5.02	6.32	9.57	15.72	38.98	227.45
Realized rank shift	.019	.341	.428	.592	.798	.939
Static percentile	.012	.080	.113	.175	.234	.248
High realized percentile	.900	.935	.956	.980	.997	1.000

The strongest case has pair index 188, source/reader prompt indices 376/377, layer 23, cache position 12, and a natural-key-state direction. The stored direction norm squared is 384.2397. Its effects are $9.3473704332 \times 10^{-5}$ and $2.1260164544 \times 10^{-2}$, producing the unrounded ratio 227.4454. Static, self-realized, and swap-realized percentiles are 0.1143, 0.0146, and 0.9531; attention weights are 0.0180 and 0.0253. The 340/165,888 frequency is an audit frequency conditional on the predicate, not a population estimate.

One fixed reference for both effects. For the displayed example, let $\mathcal{R}_A$ contain the 960 self-reader effects in the same global-panel, layer-23, key, natural-key-state stratum. We evaluate both

effects against its empirical cumulative distribution:

$$\widehat{F}_A(x) = \frac{1}{960} \sum_{I_{A,j} \in \mathcal{R}_A} \mathbf{1}[I_{A,j} \le x].$$

There are 14 reference effects at or below I_A, and 957 at or below I_B. Thus $100\widehat{F}_A(I_A) = 1.4583$ and $100\widehat{F}_A(I_B) = 99.6875$, reported as the 1.46th and 99.69th percentiles in the Introduction and Fig. 1. The source effect occurs once in this reference, and the substituted effect has no exact match, so these values also agree with the average-rank convention. The 95.3125th percentile instead uses the substituted-reader distribution.

The existence result is not confined to the exact audit thresholds. Varying one cutoff at a time gives 149/340/944 candidates for static-percentile maxima 0.10/0.25/0.40; 539/340/199 for realized-percentile minima 0.80/0.90/0.95; and 1,079/340/159 for minimum reader-effect ratios $2\times$ /$5\times$ /$10\times$. These counts characterize threshold sensitivity of the audit frequency; establishing fixed-storage reader dependence requires only one fixed-input, different-target case.

D Controlled Prompt Sets and Factor Panels

The RC-PromptBench v0.1 controlled prompt set contains a 960-prompt Core:

$$12 \text{ languages} \times 8 \text{ task regimes} \times 5 \text{ instances per task} \times 2 \text{ surface forms}.$$

A Core semantic instance is one task-specific canonical proposition or problem, shared across its languages and surface forms. There are five instances within each of eight task families, giving 40 distinct semantic IDs in total. The task regimes are contextual retrieval, compositional relation, arithmetic, logical derivability, evidence conflict, constrained summarization, structured transformation, and code execution. The main mechanism condition uses a language-invariant answer lexicon and fixed suffix; token counts are recorded with the model tokenizer.

Each language version is rendered by hand-curated controlled templates from a language-independent canonical proposition structure. Automated checks verify proposition-field equality within each Core semantic ID and context invariance across Bridge instruction heads. These template realizations have not received independent native-speaker review, so the language-swap magnitude may combine language, tokenization, and realization-quality effects. The matched-length analysis below addresses token count but not translation naturalness or grammaticality.

Content swaps hold language, task, and surface fixed; surface swaps hold language, task, and semantic content fixed; language swaps use a balanced degree-three graph while holding task, content, and surface index fixed. The Bridge panel is constructed independently of Core: 32 shared contexts $\times 4$ instruction heads $\times 12$ languages give $1{,}536$ prompts. A Bridge semantic instance means one shared context, with the same ID across all heads and languages; these 32 contexts are not a subset of the 40 Core instances. The global panel uses unrestricted readers while retaining source KV and perturbation.

Table 4: Frozen controlled factor-sweep summary. Operator ρ tracks paired log-effect changes; reductions concern level log-RMSE relative to the KV-only static predictor, using both readers.

Panel	Groups	Records	Median $\lvert\Delta \log I\rvert$	ρ	All	Held-out
Content	480	15,360	.0859	.7317	48.41%	63.50%
Surface	480	15,360	.1580	.7487	47.14%	62.19%
Bridge	2,304	73,728	.2777	.7487	44.28%	58.35%
Language	1,440	46,080	.3106	.7965	47.12%	62.32%
Global	480	15,360	.4344	.8214	47.25%	62.55%

Table 4 summarizes the five panels. The frozen adapter takes the first 480 distinct unordered content pairs. In its deterministic ordering these cover T1–T4 (20 Core instances), whereas the surface panel covers T1–T8 (40 instances). The common-family restriction in Table 6 aligns the Core source population. Table 5 reports the Bridge head-pair breakdown with both group and pooled-record summaries. Restricting language pairs to tokenizer-count differences of at most ten retains 843

pairs and 26,976 records; median change remains 0.2937, Spearman 0.7982, and all-record error reduction 47.04%.

Calibration slopes for surface, language, content, Bridge, and global panels are 1.0698, 1.0386, 1.0758, 1.0469, and 1.0571; corresponding R^2 values are 0.715, 0.737, 0.680, 0.749, and 0.807. A crossed drop-one descriptive model explains 33.19% of absolute paired-change variance. Partial R^2 values are 0.2828 for layer, 0.0286 for factor, 0.0052 for channel, 0.0040 for source task, 0.0031 for source language, and 0.0004 for semantic ID; these are not additive variance components. Key/value median-change ratios are 1.12, 1.13, 1.13, 1.13, and 1.07. For value interventions, twice the change in log attention weight has slopes 0.957–1.230, Spearman 0.720–0.787, and sign agreement 0.803–0.843 across factors.

Bootstrap resampling uses prompt/source or semantic clusters rather than individual translated prompts. The frozen quality audit confirms finite values, complete expected layers and channels, no literal padding markers, derivability wording for logical tasks, and unique scores in sorting tasks.

D.1 Within-population controls and descriptive summaries

The Core surface panel contains 480 bidirectional pair groups (15,360 records) over 40 semantic instances. The Bridge panel contains 2,304 pair groups (73,728 records) over 32 numeric variants of one transfer-context template. Each pair group contains two orientations, eight layers, and two channels. Surface pairs preserve semantic content, task, and language; Bridge pairs preserve context and language while changing instruction heads. These are controls within separate prompt populations. The Core surface panel is not a paraphrase control for the Bridge contexts. The reported medians describe the fixed prompt constructions and do not estimate variation across new instruction templates.

Overall panel values use the pooled-record median

$$\text{median}\,|\log(\max(I_B, 10^{-12})) - \log(\max(I_A, 10^{-12}))|.$$

For the Bridge head-pair breakdown, we also summarize the distribution across reader-pair groups. Let x_{gj} be the absolute log-effect change for record j in group g, and let G_h contain the 384 groups for head pair h. The group summary is

$$M_h^{\text{group}} = \text{median}_{g \in G_h}\left(\text{median}_{j=1,\ldots,32}\, x_{gj}\right).$$

It differs from the median over all $384 \times 32 = 12{,}288$ records. Table 5 reports both; the 0.1303 B1–B2 value is M_h^{group}. B1 and B2 ask initial and post-transfer count questions; B3 asks for a comparison and B4 for structured output.

Table 5: Bridge instruction-head pairs. Each row has 384 reader-pair groups and 12,288 effect records from 32 numeric variants of one context template in 12 languages. Columns report group and pooled-record medians within this single-template construction.

Head pair	Median of group medians	Pooled-record median
B1–B2	.1303	.1292
B1–B3	.3153	.3134
B1–B4	.3269	.3136
B2–B3	.3569	.3556
B2–B4	.3210	.2958
B3–B4	.3728	.3784

Common Core task families. Restricting the surface records to T1–T4 retains 240 pair groups and 7,680 records; the content panel has 480 groups and 15,360 records. Both cover the same 20 semantic instances and 480 source prompts. Their source-state keys (prompt, layer, channel, position, and direction family) coincide; each has exactly twice as many content records as surface records. Thus the same source population enters with proportional weights. The pooled medians are .1552 for surface and .0859 for content (Table 6). The pooled ordering persists under the common task inventory. Content substitutions replace facts within a fixed template; surface substitutions change instruction wording while preserving semantics. Semantic preservation can therefore coexist with a larger change in measured KV sensitivity. Each family has five semantic instances, so the family-level estimates are descriptive; T4 has the opposite sample ordering.

Table 6: Core controls on the same source inventory. Each task family has five semantic instances, 120 content pair groups, and 60 surface pair groups. All entries are pooled-record medians of absolute same-entry log-effect change.

Core families	Semantic instances	Content	Surface
T1	5	.1233	.2360
T2	5	.0437	.1628
T3	5	.0982	.2012
T4	5	.1092	.0788
T1–T4 pooled	20	.0859	.1552

Token-count restriction. Table 7 applies the existing source–donor token-gap criterion separately within each panel. On T1–T4, it gives surface .1590 and content .0845. The full Surface and Bridge source-token medians after filtering remain 46 and 64, respectively: this condition does not match those two prompt populations. The language result reproduces the frozen 843-pair, .293706 check.

Table 7: Within-panel restriction to source–donor token-count differences of at most ten. Every pair group contributes 32 effect records.

Panel	All pairs	Retained pairs	All median	Retained median
Content, T1–T4	480	464	.0859	.0845
Surface, T1–T4	240	225	.1552	.1590
Surface, T1–T8	480	455	.1580	.1582
Bridge	2,304	1,696	.2777	.2621
Language	1,440	843	.3106	.2937

Bridge answer metadata. The frozen context transfers type-R units from A to B, while the reference answers increase B's type-U count. Under the literal typed-transfer proposition, this makes 384 B2, 72 B3, and 384 B4 prompt answer keys inconsistent with the stated facts. The prompt builder supplies only context, instruction, and answer prefix; the reference answers are neither model inputs nor targets for Eq. 1. We preserve the frozen prompts and effects and interpret the panel as dependence on the actual instruction heads. Its answer metadata does not validate task execution. The observed same-reader repeatability result is reported with the general protocol in Appendix B.

E Mechanism, Operator Ladder, and Robustness

E.1 Targeted mechanism interventions

For alignment, equal-norm key directions are projected into the row space of all active queries in one GQA group or into its numerical null space. Routing uses equal-norm value perturbations under natural reader swaps. Contrast holds the key perturbation, reader, norm, and attention weight fixed while moving the target cached value to low- or high-contrast conditions. The downstream/output-geometry experiment injects equal-norm directions immediately before the attention output projection and scores them with exact categorical-Fisher JVPs to final logits (Table 8).

Changing the target value in the contrast experiment also changes the baseline attention output and can change its downstream pullback metric. The reported effect ratio is the total response to that manipulation.

For one GQA group G, alignment energy is

$$E_{\text{align}} = \|Q_G \delta k_G / \sqrt{d}\|_2^2 = \sum_{h \in G} (q_h^\top \delta k_G / \sqrt{d})^2.$$

A numerical-null-space direction has zero first-order key-logit change; the reported 9.0×10^{-12} is finite-precision residue. For routing, the slope 0.962 has bootstrap CI [0.869, 1.053]. For contrast, the median contrast ratio is 21.1, the median effect ratio 28.0, and sign agreement 97.3%. For

Table 8: Targeted mechanism interventions in Qwen2.5-0.5B.

Factor	Controlled comparison	Main statistic	Sample
Alignment	query-row vs numerical-null key directions	energy 403.9 vs 9.0×10^{-12}; effect 3.05×10^{5}	256 pairs
Routing	equal-norm value changes across readers	slope .962; $\rho = .686$	256 pairs
Contrast	low/high value contrast, fixed reader/key change	slope .736; $\rho = .749$; effect $28.0\times$	256 pairs
Downstream	equal-norm pre-projection directions	within-group Q4/Q1 $3.40\times$	64 groups, 768 rows

downstream/output geometry, exact Fisher medians are 0.0151 (Q1) and 0.0695 (Q4), while median pre-projection norm is 6.1383 in every quartile. The median within-group Q4/Q1 ratio is 3.40; these quartiles summarize the spread of Fisher sensitivity across equal-norm directions.

Uncertainty from the frozen paired records. For contrast, the median of 256 high/low measured-sensitivity ratios is 28.0226. Resampling the eight source prompts, retaining all layers, positions, and KV groups within each prompt, gives a 95% percentile interval [15.5, 35.8] from 5,000 replicates (seed 1729). Resampling individual pairs instead gives [16.5, 39.0]. The alignment panel has 256 pairs from the same eight prompts: median measured sensitivities are 6.74×10^{-3} in the query row space and 2.64×10^{-8} in the numerical null space. The median paired ratio is 3.05×10^{5}, with prompt-cluster interval $[1.80 \times 10^{5}, 5.73 \times 10^{5}]$. This ratio describes the recorded finite-precision null residual, not the ratio to an exactly invisible direction; none of its denominators is clipped by the 10^{-12} analysis floor.

For downstream/output geometry, the 768 directions comprise 12 equal-norm directions in each of 64 prompt/layer groups. The statistic is the median across groups of the ratio between the within-group Q4 and Q1 medians of exact Fisher sensitivity: 3.4044. A 5,000-replicate group bootstrap gives [2.93, 4.19]; resampling the eight prompts together with all their layers gives [2.98, 3.95]. These intervals describe dispersion in the frozen direction panel. Quartiles were defined by sorting the same Fisher values, so Q4 exceeding Q1 in all 64 groups follows from their construction when values are distinct.

E.2 Complete operator-ladder summary

Table 9 reports relative log-RMSE reduction over the KV-only rung as attention computation is restored. Intermediate rungs need not improve monotonically because each frozen feature set is separately calibrated; the final $+V$ rung consistently improves all three models.

Table 9: Operator-ladder improvement over KV-only (%). NH denotes natural held-out layers.

Model	Split	+Q	+QK	+A	+J	+V
Qwen2.5-0.5B	all	3.74	6.41	17.55	17.67	35.01
	held-out	-0.85	2.92	22.55	22.67	47.47
	NH	-1.15	6.72	30.83	31.08	57.75
Qwen2.5-1.5B	all	8.46	-7.71	-4.91	4.15	44.47
	held-out	8.33	-10.70	-7.01	3.93	52.75
	NH	7.86	-6.45	-3.72	16.61	63.93
GPT-2 small	all	3.18	7.86	9.45	12.22	20.16
	held-out	3.76	8.72	10.48	13.65	21.03
	NH	4.58	12.56	14.21	18.53	24.51

E.3 Capacity, locality, output distance, and architecture

The GPT-2 calibration artifact contains 1,024 rows and 1,536 squared K/V coordinates. The controlled operator log-RMSE is 1.7402 overall and 1.7148 on held-out layers. Aggregate Ridge gives

1.8787/1.8902; coordinate GBM gives 1.9723/2.1468. The resulting operator advantages are 7.38–11.77% overall and 9.28–20.12% held out. Adding all squared coordinates to the fixed Ridge raises RMSE to 2.5009/2.6595, a sensitivity result potentially affected by regularization mismatch rather than evidence against rich static features.

Across 384 fixed input-direction combinations and six perturbation sizes, median effects are 0.002257, 0.002255, 0.002268, 0.002265, 0.002262, and 0.002262 for $\epsilon = 0.001, 0.0025, 0.005, 0.01, 0.02, 0.04$. Relative to 0.001, median drift is 1.18% at 0.01, 1.28% at 0.02, and 1.68% at 0.04; tail deviation grows at 0.04. Jensen–Shannon rescoring keeps static reader change exactly zero and true reader change nonzero. Reader-change correlations are 0.499/0.548/0.596 for Qwen2.5-0.5B and 0.360/0.403/0.425 for Qwen2.5-1.5B across all/held-out/natural-held-out controlled splits.

Fixed-input reader dependence is also observed in Qwen2.5-0.5B, Qwen2.5-1.5B, and GPT-2 small. Their pooled reader shares of level-$\log I$ variance are 4.1%, 6.9%, and 19.2%; these are level-decomposition quantities, not paired-change sizes.

E.4 Exploratory GQA and RoPE controls

Qwen2.5-0.5B has 14 query heads and two KV heads, or seven query heads per GQA group (Ainslie et al., 2023). Replacing one query group while holding source KV and perturbation fixed gives median absolute log ratios 0.355 and 0.433 for the two groups; full reader replacement gives 1.031. Within-group permutations give 0.553 and 0.516; cross-group controls give 0.968 and 1.591. Each condition contains 32 observations. This full-swap value differs from the main factor-sweep global median (0.4344) in reader pairs, positions, directions, and denominator; it is not a second estimate of the same quantity.

A synthetic RoPE probe (Su et al., 2021) changes only effective positional phase. Relative to native source-position phase, median absolute log changes are 1.890 when phase is cancelled, 0.969 with reader-position phase, 0.747 at half the source angle, and 1.468 at double the source angle. Each mode has 32 observations. These panels define components of the active readout state and are not population effect-size estimates.

F Natural Multi-reader QASPER All-position Audit

The QASPER all-position panel contains 64 documents, four natural question readers per document, eight layers, both K and V channels, and all 512 retained document positions. It produces 2,097,152 primary FP32 measurements, 4,096 within-reader decision sets, and 3,145,728 fixed stored-KV reader pairs. Split A is the exact 32-document pilot panel; Split B is the next deterministic, non-overlapping set of 32 eligible documents and was frozen before outcomes were inspected. Table 10 gives the full protocol.

An answerable QA has at least one annotation not marked unanswerable that contains a nonempty free-form answer, one or more extractive spans, or a yes/no label. After filtering in the dataset's stored QAS order, the first four answerable QAs define the readers. Candidate documents are serialized from nonempty fields with one newline between pieces, tokenized with special tokens, required to retain at least eight tokens and one nonspecial position, and sorted by the pair (token count after 512-token truncation, document ID). Split A takes the first 32 candidates and is checked against the pilot IDs; Split B takes candidates 33–64. The combined question template uses the tokenizer's right-truncation rule at 64 tokens, although the 256 selected readers span only 9–32 tokens and no truncation occurs.

Question tokens participate in the forward computation but are not candidates for intervention. All 64 source documents contain at least 512 model tokens before truncation (minimum 1,601); every candidate is a real retained document token, with no padding, question, answer-cue, or synthetic special-token position included. The active reader query is the final model token of the suffix ending in `Answer:`. The natural reader condition therefore includes the full question-conditioned residual and downstream state. No rung has fitted coefficients; S0–S4 are computed from the unperturbed model state, and no perturbed logits, Jeffreys targets, or QASPER causal target enters score construction. The reader-independent rungs are channel-specific direction norm $S_0 = \|\delta z_i\|_2^2$ and document-

Table 10: Frozen QASPER all-position protocol.

Component	Frozen choice
Model/numerics	Qwen2.5-1.5B base; eager attention; FP32 model, cache, intervention, logits, and divergence; TF32 disabled
Data/splits	QASPER dev v0.3; previously analyzed first 32 eligible documents as Split A; prospectively frozen next 32 as non-overlapping Split B
Eligibility/order	At least four answerable QAs; at least eight tokenized document tokens and one nonspecial retained position; ascending (512-truncated token count, document ID)
Prefix/readers	Nonempty title, abstract, section names, and paragraphs joined by one newline; first four answerable QAs in source order; `Question: [question]` then newline and `Answer:`; 64-token cap (observed 9–32, so none truncated); final suffix token is the active query
Cache scope	Document prefix prefilled once; perturb a cloned document cache, then run the entire question suffix; 512 document positions are candidates
Layers/channels	Layers 0, 3, 7, 9, 15, 16, 21, 23; key and value
Direction/target	Primary radial $z_i \mapsto (1 \pm 0.01)z_i$; symmetric Jeffreys divergence divided by ϵ^2
Batch/null	Separate baseline, plus, and minus forwards with identical batch size 32 and cache layout; second unperturbed same-layout forward
Recovery	Static norm, document history, squared current-reader routing, analytic gate, first-order last-query response, and last-query downstream-Fisher diagnostic
Inference	5,000 document-cluster bootstrap replicates; exact $k/512$ random expectation at 5/10/20/50%
Robustness	Eight frozen documents for $\epsilon = 0.005, 0.02$, one tuple-seeded unit-Gaussian direction, and S5 at layers 3, 9, 16, 23

prefill history $S_1 = \sum_{t>i,h} a^{\mathrm{doc}}_{h,t,i}$. The squared current-reader routing score is $S_2 = \sum_h a^2_{hi}$. For the analytic gate,

$$S_3^V = \sum_h a^2_{hi} \|\delta v_{g(h),i}\|_2^2,$$

$$S_3^K = \sum_h \|J_{\mathrm{softmax},h}[:, i]\|_2^2 \big(\gamma_{\mathrm{attn}}\, q_h^\top \delta k_{g(h),i}\big)^2,$$

where $g(h)$ maps query heads to shared KV heads and γ_{attn} is the model's attention scaling. Q and cached K use the RoPE-positioned coordinates passed to attention. The first-order last-query response uses

$$\delta o_h^V = a_{hi}\delta v_{g(h),i}, \qquad \delta o_h^K = a_{hi}(\gamma_{\mathrm{attn}} q_h^\top \delta k_{g(h),i})(v_{g(h),i} - o_h).$$

Let $\delta u_{T,i} = \mathrm{concat}_h(\delta o_h)$ and $\delta r_{T,i} = W_O \delta u_{T,i}$, with T denoting the final suffix token. Then $S_4 = \|\delta r_{T,i}\|_2^2$. The last-query downstream-Fisher diagnostic is

$$S_5 = S_{5,\mathrm{last}} = \delta r_{T,i}^\top J_{\mathrm{post,T}}^\top F(p) J_{\mathrm{post,T}} \delta r_{T,i}, \qquad J_{\mathrm{post,T}} = \frac{\partial \mathrm{logits}}{\partial r_T}.$$

The code injects the direction after o_{proj}, with zero injection at all earlier suffix tokens, and computes the final-logit JVP by autodifferentiation. Thus $J_{\mathrm{post,T}}$ begins after W_O; for the same path the head-output Jacobian is $J_{u_T} = J_{\mathrm{post,T}} W_O$, consistent with Eq. 3.

The measured intervention instead perturbs the cloned document cache before running the entire question suffix. Earlier question states are recomputed under the perturbed cache. For a single intervened layer, write its response at every suffix token as $\delta r_{t,i}$. The full directional derivative and its Fisher coefficient are

$$g_i = \sum_{t=1}^{T} J_{\mathrm{post},t} \delta r_{t,i}, \qquad I_i^{(0)} = g_i^\top F(p) g_i.$$

This quadratic form includes cross-token terms. S5 retains only the $t = T$ path, so its correlation with the finite-ϵ target assesses that path-specific diagnostic. The complete coefficient satisfies $I_i^{(\epsilon)} = I_i^{(0)} + O(\epsilon^2)$ in the smooth, numerically resolved regime. We evaluate it separately in Appendix F.2; S5 retains its last-query definition.

The diagnostic subset contains four documents from each split, selected by a frozen deterministic ranking of seed, purpose, split, and document ID before outcomes were observed. Layers 3, 9, 16, and 23 were pre-specified to span depth while bounding JVP cost.

Recovery estimands and uncertainty. The two recovery tasks use different comparisons. Position ranking holds the document d, reader r, layer ℓ, and channel c fixed and compares 512 entries:

$$\rho^{\mathrm{pos}}_{d,r,\ell,c} = \mathrm{Spearman}_i\big(S_{d,r,\ell,c,i}, I_{d,r,\ell,c,i}\big). \tag{9}$$

The position-ranking mean averages these decision sets. Effect-change tracking instead fixes an entry and compares its effects across question pairs. For each document, it correlates the predicted and measured changes over reader pairs, layers, channels, and positions:

$$\rho^{\Delta}_{d} = \mathrm{Spearman}_{(r_A,r_B),\ell,c,i}\big(\Delta_{\mathrm{pair}} \log S, \Delta_{\mathrm{pair}} \log I\big). \tag{10}$$

Tracking means average document correlations; layer estimates restrict the same calculation to one layer. Population sizes follow from the frozen protocol (Table 10). Confidence intervals resample whole documents in 5,000 bootstrap replicates, retaining their readers and positions together.

For the variance decomposition, the fixed intervention is indexed by $s = (d, \ell, c, i, \delta)$, comprising document, layer, channel, retained-prefix position, and frozen radial perturbation; r indexes the four natural question readers associated with document d. Define $y_{sr} = \log(\max(I_{sr}, 10^{-30}))$, $\mu_s = R_s^{-1} \sum_r y_{sr}$, and $e_{sr} = y_{sr} - \mu_s$. The reported floor is inactive in the primary data: 0 of 2,097,152 records hit 10^{-30}, and the minimum observed importance is 2.99×10^{-8}. The reported decomposition is

$$V_{\mathrm{state}} = \mathrm{Var}_s(\mu_s), \qquad V_{\mathrm{reader}} = |S|^{-1} \sum_s R_s^{-1} \sum_r e_{sr}^2,$$

with shares computed relative to $V_{\mathrm{state}} + V_{\mathrm{reader}}$. Pooled shares are 97.83% between intervention sites and 2.17% within-site reader variation; median $|\Delta \log I|$ across all reader pairs is 0.573. V_{state} pools variation in reader-averaged measured log effects across documents, layers, channels, and positions, with the perturbation fixed at each site. Stored-vector features are not varied independently of those site indices, so this term does not estimate a separate storage contribution. The decomposition contrasts site variation with reader variation within sites, rather than storage with computation. Conditioning on layer and channel gives reader shares from 7.1% to 38.6%; these remain level shares, not paired-change effect sizes.

Table 11: Natural recovery ladder. Within-reader entries are means of per-set Spearman correlations; paired and centered entries are means of per-document Spearman correlations. Intervals and paired increments use the same document aggregation, with 5,000 document-bootstrap replicates. All aggregates retain layer-0 K; its locality audit and exclusion sensitivity appear in Appendix F.1.

Rung	Within reader	Fixed-KV change	Reader centered
S0 static norm	-.051 [-.060,-.043]	–	–
S1 history	.283 [.274,.294]	–	–
S2 squared current-reader routing	.747 [.741,.753]	.436 [.421,.450]	.435 [.419,.451]
S3 gate	.755 [.750,.761]	.434 [.419,.448]	.432 [.416,.448]
S4 last-query response	.778 [.772,.784]	.440 [.425,.454]	.438 [.423,.453]

Paired document-cluster bootstrap gives within-reader increments of .008 [.008,.009] for S3–S2, .022 [.022,.023] for S4–S3, and .031 [.030,.032] for S4–S2. For fixed stored-KV reader change, the corresponding document-mean increments are -.002 [-.004,.000], .006 [.005,.007], and .004 [.002,.006]. Thus S3/S4 progressively improve within-reader ranking but add little to reader-induced effect-change recovery beyond squared routing. As a secondary aggregation, pooling all reader pairs gives S2/S3/S4 correlations .435/.431/.437; pooled centered correlations are .435/.432/.438. The intervals in Table 11 apply to document means, not these pooled values.

Split A/B S4 within-reader correlations are .777/.779; fixed stored-KV document-mean correlations are .443/.436.

The primary analysis retains the pilot's eight-layer schedule, frozen before the all-position run. This nonuniform subset spans early, middle, and later blocks of the 28-layer model. Each layer contributes 64 documents and $64 \times \binom{4}{2} \times 2 \times 512 = 393{,}216$ reader-pair records (Table 12). The six

question pairs share four readers within each document; uncertainty is estimated over 64 document clusters using 5,000 bootstrap replicates.

Table 12: Fixed-entry reader-change tracking by layer (S4). Intervals are the frozen 95% document-bootstrap intervals for the mean of per-document Spearman correlations. Both channels are retained, including layer-0 K.

Layer	Documents	Reader-pair records	Mean ρ	95% CI
0	64	393,216	.020	[.010, .030]
3	64	393,216	.043	[.028, .058]
7	64	393,216	.036	[.007, .066]
9	64	393,216	.165	[.133, .197]
15	64	393,216	.462	[.433, .491]
16	64	393,216	.588	[.562, .613]
21	64	393,216	.768	[.748, .786]
23	64	393,216	.822	[.806, .836]

On the matched S5 subset, document-mean fixed stored-KV correlation changes from .486 (S4) to .516 (S5), paired difference .030 [.001,.063]; reader-centered correlation changes from .486 to .525, difference .039 [.009,.079]. S5 does not improve within-reader ranking (.794 to .786).

Layer-specific last-query downstream-Fisher diagnostic. The layer analysis joins S5 reader-pair records to the primary S4 records by document, layer, channel, position, and reader pair. All 196,608 measured log-effect changes agree across the two exports. Using the original analysis functions gives document-mean correlations of 0.4855217 (S4) and 0.5159410 (S5).

For each document/layer, we compute Spearman correlation between predicted and measured paired log-effect changes over $6 \times 2 \times 512 = 6{,}144$ records, then average the eight document correlations. We bootstrap paired per-document S5–S4 differences with 5,000 replicates and seed 1729. Intervals are pointwise, with documents as the resampling unit. The aggregate above pools layers within each document before computing correlation; it is not an average of the layer means.

Table 13: Layer-specific effect-change tracking in the matched eight-document subset. S5 is the last-query downstream-Fisher diagnostic. The interval applies to the paired document-mean increment. Each layer contains 49,152 reader-pair records.

Layer	S4 mean ρ	S5 mean ρ	S5–S4	95% CI	Documents
3	.069	.142	.073	[-.001, .166]	8
9	.157	.242	.085	[.027, .149]	8
16	.581	.595	.013	[-.011, .043]	8
23	.828	.867	.039	[.020, .060]	8

The layer-9 and layer-23 intervals in Table 13 indicate a contribution from post-final-query downstream geometry on this subset; layer-3 and layer-16 increments are less resolved. Tracking remains weak at the early tested layers after adding this geometry. The complete derivative, including earlier suffix paths, is evaluated separately in Appendix F.2.

Table 14 shows that both channels recover within-reader rankings and fixed stored-KV changes. Value is stronger for the mean within-reader target, whereas key and value reader-change correlations are similar. Removing the failed layer-0 K cell strengthens every key summary; the frozen primary aggregate continues to retain that cell.

Mean importance Gini is .956 (median .975; IQR [.951,.985]). We report MassEfficiency@k by averaging the per-set ratios $M_{S4}@k/M_{\text{oracle}}@k$. This quantity differs from the ratio of aggregate mean masses in Table 15. Mean efficiency is .883 at 5%, .911 at 10%, .936 at 20%, and .972 at 50%.

Uniform random selection has expected mass $k/512$, independent of importance concentration. With concentrated importance, recovering dominant positions can yield high mass without accurately ordering positions with small effects. S4 exceeds S0/S1 at every reported budget, while its

Table 14: Supplementary K/V channel audit from frozen outputs. Within-reader entries are mean Spearman correlations; reader-change entries are pooled fixed-entry Spearman correlations. The last two rows report S4 top-5% recall and retained-prefix sensitivity mass.

Metric	Key	Value	Key excluding layer 0
Within-reader S2	.719	.775	.791
Within-reader S3	.718	.793	.806
Within-reader S4	.762	.794	.835
Reader-change S2	.436	.431	.486
Reader-change S3	.431	.433	.492
Reader-change S4	.441	.434	.501
S4 Recall@5%	.626	.704	.711
S4 mass@5%	.796	.864	.903

Table 15: QASPER all-position selection, averaged over decision sets. For budget b, $k = \lceil 512b \rceil$. Let A_k be the score's top-k positions and O_k the top-k positions by measured sensitivity. Then $\mathrm{Recall@}k = |A_k \cap O_k|/k$. Retained-prefix mass is the sensitivity summed over A_k, divided by its sum over all 512 retained positions. Uniform random selection has expectation $k/512$ for both metrics.

Score/metric	5%	10%	20%	50%
S4 Recall	.665	.691	.726	.812
S2 Recall	.624	.654	.695	.793
S1 Recall	.196	.259	.363	.603
S0 Recall	.052	.104	.199	.487
Random expectation	.051	.102	.201	.500
S4 mass	.830	.877	.916	.965
S2 mass	.819	.872	.914	.962
S1 mass	.490	.530	.588	.693
S0 mass	.038	.071	.134	.297
Oracle mass	.912	.948	.973	.992

increment over current routing (S2) is modest (.830 versus .819 at 5%). Within-reader rank correlations complement mass capture by evaluating ordering across all positions.

Split A/B S4 mass differs by at most 1.1 percentage points across the four budgets. Dividing importance by squared perturbation norm preserves raw position ranking at mean Spearman .966; S4 remains .763 for the normalized target.

F.1 Precision and epsilon audit

Local cache interventions are not reliable in BF16 at this scale. On the matched pilot panel, BF16/FP32 reader-change correlation is .056 with 52.8% sign agreement. The all-position panel therefore uses FP32 through divergence and batch-matched baselines. Its Jeffreys same-forward null is zero in all 2,097,152 primary records, indicating no observed repeat-forward difference. Locality is assessed by the cross-epsilon comparisons below.

On eight frozen documents, $\epsilon = .005/.02$ versus .01 gives record-level rank correlations .946/.975, mean within-set correlations .833/.903, fixed stored-KV reader-change correlations .763/.835, and top-k overlaps .858–.910/.909–.926. Absolute log effects are less stable than ranks. Layer-0 K is the single failed locality cell: its rank correlations are .194/.256 and median plus/minus asymmetry is approximately one. The other 15 layer/channel cells have rank correlations at least .807/.896. Frozen primary aggregates retain layer-0 K, but it is interpreted as a nonlinear finite-perturbation corner rather than a stable derivative. Excluding this cell gives S2/S3/S4 within-reader correlations .783/.799/.813 and fixed stored-KV document means .460/.464/.469. The corresponding pooled reader-change correlations are .459/.461/.466. The pooled reader share becomes 14.06%; S4 top-5% Recall/mass become .708/.882, compared with .665/.830 in the frozen aggregate.

Matched perturbation-family comparison. The tuple-seeded unit-Gaussian control contains eight frozen documents and 512 decision sets. S0/S1/S2/S3/S4 mean within-reader correlations are -.001/.113/.391/.398/.403; the corresponding pooled reader-change correlations for S2/S3/S4 are .142/.149/.150. Table 16 matches radial and Gaussian measurements on document, reader, layer, and channel. Correlations use the original decision-set estimator; intervals resample the eight document means with 5,000 replicates and seed 1729.

Table 16: Matched perturbation-family comparison: mean within-reader position-rank Spearman correlation, with 95% document-bootstrap intervals. Both families use the same eight documents and 512 decision sets.

Score	Radial	Unit-Gaussian
S2 squared current-reader routing	.747 [.723, .769]	.391 [.357, .425]
S3 analytic gate	.755 [.733, .776]	.398 [.362, .432]
S4 first-order last-query response	.780 [.758, .801]	.403 [.366, .438]
S4–S2 paired increment	.033 [.031, .035]	.011 [.010, .013]

Gaussian directions have unit norm; radial directions retain the stored-vector norm. The control therefore tests perturbation-family dependence without matching direction magnitude. The matched radial S4 pooled reader-change correlation is .434, a separate estimand from the within-reader correlations in Table 16.

The secondary FP32 Jensen–Shannon estimate is below its numerical resolution: median measured sensitivity is 3.02×10^{-4}, while median same-forward null is 4.03×10^{-4}; only 36.1% of records exceed their null. Jeffreys–JS mean within-set correlation is .072. Gold evidence is fully inside the retained prefix for six readers and partially inside for 16; the any-overlap subset gives S4 .765 within reader and .423 for pooled fixed stored-KV changes. The main population remains all readers and measures natural-question readout over a retained prefix rather than QASPER answer accuracy.

F.2 Complete suffix-path partition

We specified this extension after inspecting the primary and S5 results, then froze its positions and analysis before the new forwards. It reuses the eight S5 documents, all four readers, both channels, and layers 3, 9, 16, and 23. Within each document's 512-token prefix, we divide positions into 64 consecutive eight-position strata and select one position per stratum by a deterministic ranking with seed 20260909. The resulting positions are shared across layers, channels, and readers. This gives 16,384 measurements and 24,576 fixed-entry reader-pair records, including 768 reader-pair records per document/layer. No records are removed by effect size, locality, or predictor performance.

Intervention and path partition. For each sampled radial KV perturbation, we construct the analytic post-W_O response $\delta r_{t,i}$ at every suffix position. The calculation retains the model's GQA head sharing, RoPE-positioned queries, key-dependent value contrast, and output mixing. Autodifferentiation then propagates either the final-position response alone or all suffix responses through the remaining computation, using the unperturbed document cache. Subtracting the two logit JVPs gives $g_{<T,i} = g_i - g_{T,i}$. We save the three terms in Eq. 8: $g_{T,i}^\top F g_{T,i}$, $2g_{T,i}^\top F g_{<T,i}$, and $g_{<T,i}^\top F g_{<T,i}$. The cross term is signed, so the individual terms are not independent fractions of importance mass.

The model, cache, local response, and JVP use FP32 with eager attention and TF32 disabled. Last-only and full-path directions are evaluated together in batches of four independent examples, retaining only the final logits. Fisher contractions use the resulting probabilities and centered logit directions, accumulated in float64. This changes the summation precision, while the model and directional derivatives remain FP32.

Validation. At the first two selected positions in every document/layer/channel cell for reader 0, we also differentiate the KV intervention directly and run an independent earlier-suffix-only injection. Across 128 checks, the maximum relative logit-JVP discrepancy between direct KV and all-suffix differentiation is 1.65×10^{-5}. Earlier-only JVPs agree with $g_i - g_{T,i}$ to 1.51×10^{-5}, relative to $\|g_i\|_2 + \|g_{T,i}\|_2$. The largest relative residual in the three-term Fisher identity is 1.18×10^{-7}. Across

all new records, the maximum relative differences from frozen S4 and S5 values are 4.36×10^{-6} and 2.49×10^{-5}, respectively. Batched last-only and full-path baselines have identical logits.

Reader-change tracking. The target remains the original FP32 finite-difference sensitivity at $\epsilon = .01$. We reuse the original reader-pair construction and per-document Spearman estimator, now on the sampled positions. Table 17 reports the mean of eight document correlations at each layer; Fig. 4 displays the paired document increments. All increments compare the same documents, positions, channels, and six reader pairs. Confidence intervals use 5,000 paired document-cluster bootstrap replicates with seed 20260909; they are pointwise intervals. Complete JVP values are computational diagnostics of the first-order effect, requiring downstream access. They are not scores evaluated with only local reader information.

Table 17: Reader-change tracking on the new, matched path-partition sample. S4 is the first-order last-query response; S5 adds final-query downstream Fisher geometry; Full includes every suffix path and cross term. Each layer uses eight documents and 6,144 reader-pair records. The interval applies to the paired document-mean Full–S5 increment.

Layer	S4	S5	Full	Full–S5	95% CI
3	.024	.100	.698	.598	[.529, .679]
9	.160	.265	.779	.514	[.386, .654]
16	.580	.586	.830	.244	[.169, .327]
23	.818	.861	.866	.005	[-.001, .010]

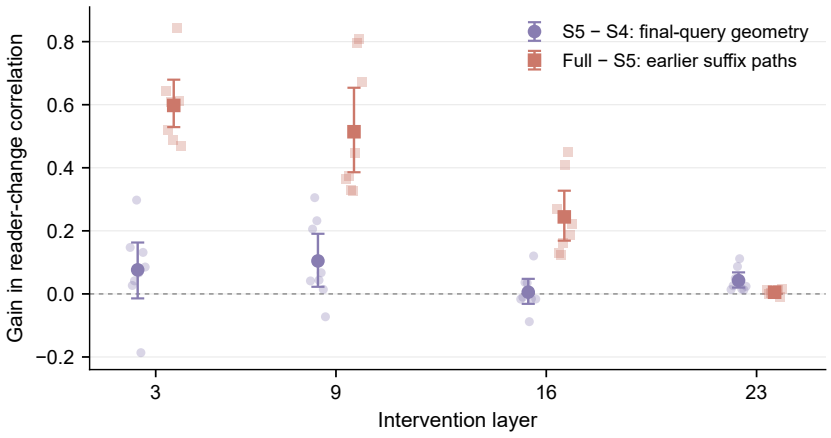

Figure 4: **Document-level gains from extending the observed path.** Paired increments from S4 to S5 restore final-query downstream geometry; increments from S5 to Full add earlier suffix responses and their cross terms. Faint marks show each of eight documents; solid marks show their means with 95% paired document-bootstrap intervals. Markers are offset for visibility. The corresponding absolute correlations appear in Fig. 3D and Table 17.

Finite-difference agreement. The complete JVP is the first-order limit of the measurement target. Its agreement with finite-ϵ effects therefore checks the measurement regime; the informative path comparison removes earlier suffix responses while keeping the final-query downstream geometry. Against the exact full coefficient, S5 has reader-change correlations .137, .284, .710, and .990 at layers 3, 9, 16, and 23. The weak early-layer tracking of S5 thus also occurs with a target computed by differentiation.

Full-JVP agreement with the finite target remains imperfect. An exploratory reanalysis of the existing $\epsilon = .005, .01, .02$ records on these sampled positions gives Table 18. The median ratio $I^{(0)}/I^{(.01)}$ is .266, .441, .253, and .309 at the four layers. Smaller steps move the measurements

further from the first-order coefficient, consistent with a contribution from finite numerical resolution. The layer-3 correlation of .698 at $\epsilon = .01$ demonstrates predictable reader-change signal in that finite target, but rank agreement cannot bound its magnitude error or exclude a substantial numerical contribution. Zero same-reader repeat differences alone do not establish that these small input-dependent responses are numerically resolved. We retain the original finite target and all sampled records in Table 17; the additional epsilon comparison is a follow-up audit, not a replacement primary analysis.

Table 18: Reader-change tracking and finite-difference resolution on the matched path-partition sample. Entries are means of eight document-level Spearman correlations. S5 versus the exact full coefficient isolates last-query path completeness from finite-difference target mismatch. The other columns compare the complete JVP against existing finite measurements at three step sizes.

Layer	S5 vs. exact full	Full vs. $\epsilon = .005$	Full vs. $\epsilon = .01$	Full vs. $\epsilon = .02$
3	.137	.567	.698	.807
9	.284	.652	.779	.873
16	.710	.742	.830	.891
23	.990	.813	.866	.917

F.3 Ranking on longer document prefixes

We extend the natural-reader ranking protocol to every QASPER development document with at least four answerable questions and at least 8,192 tokens under the same tokenizer. Seven documents meet this rule: `1601.02403`, `1703.10344`, `1802.06024`, `1807.07279`, `1908.05441`, `1912.13337`, and `2002.01984`. Their original lengths range from 8,195 to 20,222 tokens. Five also belong to the original 64-document panel. We select the first four answerable questions in source order; their suffixes contain 10–25 tokens, below the unchanged 64-token cap.

The same documents are evaluated at prefix lengths 512 and 8,192, with layers 3 and 23 and both K/V channels. For each document and length, 64 equal-width position strata contribute one position chosen by a deterministic ranking with seed 202609091. Those positions are shared across readers, layers, and channels. We freeze all choices before the new forwards. The longer prefix covers a larger position population, so this is a matched-document protocol comparison rather than an intervention that changes length while holding each sampled stored entry fixed.

The model remains Qwen2.5-1.5B base in FP32 with eager attention and no TF32. Prefix prefill uses 256-token chunks at both lengths. Radial perturbations retain $\epsilon = .01$, and the whole question suffix is recomputed. All reported measurements use batch size one; earlier batch-four records remain archived and are excluded from the final comparison for execution-layout consistency. Every perturbed forward uses a batch-matched intact baseline, with an identical unperturbed repeat for each reader/prefix.

The completed panel contains 14,336 measurements and 224 decision sets of 64 positions. Within each set, we compute Spearman correlation with the finite sensitivity target, average readers and channels within document, and then average document means. Pooled results also average the two layers within each document. Table 19 reports 5,000 document-bootstrap intervals. No observations are removed by measured effect size or recovery. All unperturbed repeats have identical log probabilities.

The finite target has 0 nonpositive values among the 14,336 measurements. The archive also gives paired S4–S2 and long–short increments, per-document estimates, and the secondary fixed-entry reader-change analysis. S0 is constant across readers at a fixed entry, so its reader-change correlation is undefined. Pooled S4 exceeds S2 by .061 [.054, .068] at 512 tokens and .014 [.009, .018] at 8,192 tokens. The paired long–short change in S4 is -.227 [-.307, -.148]. These comparisons describe ranking over the sampled prefix populations under the same finite-scale measurement; they do not isolate numerical resolution from the effect of additional context.

Table 19: Within-reader position-rank recovery at two prefix lengths on the same seven long documents. Entries are means with 95% document-bootstrap CIs. Each reader/layer/channel decision set ranks 64 stratified positions.

Layers	Score	512 tokens	8,192 tokens
3 and 23	S0	-0.041 [-0.089, 0.001]	-0.129 [-0.158, -0.108]
	S2	0.746 [0.711, 0.780]	0.565 [0.511, 0.625]
	S3	0.780 [0.750, 0.812]	0.573 [0.518, 0.633]
	S4	0.807 [0.778, 0.837]	0.580 [0.522, 0.642]
3	S0	-0.044 [-0.106, 0.014]	-0.116 [-0.152, -0.076]
	S2	0.577 [0.502, 0.650]	0.402 [0.339, 0.473]
	S3	0.633 [0.567, 0.700]	0.419 [0.357, 0.491]
	S4	0.669 [0.608, 0.731]	0.428 [0.363, 0.501]
23	S0	-0.037 [-0.099, 0.024]	-0.142 [-0.196, -0.090]
	S2	0.916 [0.901, 0.929]	0.729 [0.668, 0.784]
	S3	0.927 [0.912, 0.942]	0.727 [0.665, 0.783]
	S4	0.945 [0.931, 0.959]	0.731 [0.666, 0.791]

F.4 Finite joint deletion at a fixed layer

The additive sensitivity-mass audit in Table 1 ranks single-entry radial effects with the cache intact. To evaluate the consequence of deleting a set, we fix a follow-up protocol using the same eight documents as the original S5 subset, all four readers, and layers 3 and 23. Each trial intervenes at one layer. It physically retains the top 26 of 512 prefix positions in both K and V, shared across KV heads, and deletes the other 486. All other layers retain their full document cache. Every question-suffix token is then recomputed, with the original absolute positions and causal order. Thus the 5% budget applies at the intervened layer.

All selection scores use the intact cache under the same question. S2 is squared current-reader routing. To select one shared K/V mask, we sum the separate key and value S4 scores, or the separate measured key and value sensitivities. The latter is an oracle for that additive local target, not for the output after joint deletion. Ties are resolved by ascending position. Each document/reader/layer cell also uses 20 independently tuple-seeded uniform subsets of size 26.

The primary outcome is Jeffreys divergence between the intact and post-deletion final next-token distributions. We average random replicates within each cell, then readers within each document. Table 20 gives the mean of eight document means with 5,000 document-bootstrap intervals (seed 202609091). The pooled values in the main text also average the two layers within each document. All 64 reader/layer cells and 1,472 post-deletion measurements are retained.

Table 20: Actual joint deletion at one layer, retaining 26 of 512 prefix positions. Entries are mean output Jeffreys divergence with 95% document-bootstrap CIs; smaller values preserve the intact output better. Each layer uses eight documents and four readers.

Selection score	Layer 3	Layer 23
S2 routing	.0378 [.0158, .0622]	.00843 [.00573, .01099]
S4 response	.0381 [.0156, .0630]	.00881 [.00594, .01156]
Local sensitivity	.0272 [.0090, .0482]	.00848 [.00576, .01121]
Random	.2470 [.1456, .3757]	.09679 [.08073, .11259]

The paired S4–S2 divergence increment is .00032 [-.00101, .00175] at layer 3 and .00038 [.00009, .00068] at layer 23. S4 therefore gives no improvement over S2 in this test; its small layer-23 difference is in the opposite direction. Local-sensitivity selection improves on S2 at layer 3 by .01060 [.00635, .01510], while their layer-23 difference is unresolved. The archive also reports Jensen–Shannon divergence, top-1 agreement, the retained local mass, and every selected position.

Implementation checks. Model weights, cache, and forwards use FP32, eager attention, and no TF32; output divergences are accumulated in float64. We use identical final-logit computation for

intact and deleted caches. All-retained controls and repeated baselines have identical logits. For the first reader in each document/layer cell, physical compaction agrees with masking the deleted positions out of the softmax to within 1.62×10^{-5} in maximum absolute logit difference. Recomputed S2/S4 values match the frozen scores. These checks validate the specified layerwise intervention; no memory, latency, or task-accuracy result is inferred from it.

G SCALE REPLICATION ON QWEN2.5-7B

Matched design. Before the first forward for each model, we recorded the documents, layers, positions, estimands, and bootstrap procedure in a local protocol file. The replication uses the eight frozen S5 documents, all four readers, 512-token prefixes, and the same 64 stratified positions per document as the path audit. Both models have 28 layers, so this comparison tests width and parameter scale within the Qwen2.5 family, with depth held fixed. Document and question token IDs match the frozen inputs exactly. S0–S4 are evaluated at layers 3, 9, 16, and 23; S5 and the complete suffix-path derivative are evaluated at layers 3 and 23.

We rerun Qwen2.5-1.5B base alongside Qwen2.5-7B base on the same RTX PRO 6000 GPU, using FP32, eager attention, no TF32, and radial steps of .01. Finite-difference batches contain 32 positions and compute only final-token logits; each JVP batch contains eight positions for each of the two path conditions. Each model yields 16,384 records (eight documents $\times$ four readers $\times$ four layers $\times$ two channels $\times$ 64 positions), including 8,192 with path scores, with no outcome filtering.

Estimands and uncertainty. Position recovery correlates scores and measured sensitivity across 64 positions within each document, reader, layer, and channel. We average these correlations within documents, then across eight documents. Tracking uses the original six oriented reader pairs and pools channels and positions within each document/layer, giving 768 pairs per document. All intervals use 5,000 document-cluster bootstrap replicates (seed 20260910). Paired increments resample the same documents for both scores. These matched-subset estimates are separate from the primary 64-document, eight-layer, all-position results.

Position recovery and path coverage. Current routing provides most of the position-ranking signal in both models (Table 21). S2 and S4 reach .750 and .792 in 7B, versus .773 and .802 in 1.5B on the same subsample. In 7B, history reaches .256 and static norm -.062; the increase from S1 to S2 is .494, while moving from S2 to S4 adds .042.

Both models show weaker early-layer tracking and stronger tracking at the later tested layers (Table 22). Restoring earlier suffix paths raises layer-3 tracking substantially in each model. The layer-23 increment is much smaller, although its interval excludes zero in both matched-layout runs (Table 23). The frozen estimate, .005 [-.001, .010], and the matched-layout estimates thus agree on the small magnitude at layer 23, while differing in whether the gain is resolved. This comparison supports the depth asymmetry; it does not locate a common transition layer across models.

Table 21: Matched position recovery. Entries are document-mean Spearman correlations with 95% document-bootstrap intervals over four layers and 64 positions.

Score	Qwen2.5-1.5B	Qwen2.5-7B
S0	-0.101 [-0.148, -0.048]	-0.062 [-0.088, -0.034]
S1	0.322 [0.280, 0.360]	0.256 [0.225, 0.283]
S2	0.773 [0.737, 0.804]	0.750 [0.726, 0.775]
S3	0.784 [0.751, 0.813]	0.776 [0.753, 0.798]
S4	0.802 [0.770, 0.830]	0.792 [0.769, 0.814]

Table 22: Fixed-entry reader-change tracking on matched documents and positions. S5 observes the final query token; Full includes all suffix paths. Entries are means with 95% document-bootstrap intervals.

Layer	Score	Qwen2.5-1.5B	Qwen2.5-7B
3	S4	0.032 [-0.016, 0.087]	0.071 [0.013, 0.125]
3	S5	0.123 [0.046, 0.195]	0.199 [0.140, 0.255]
3	Full	0.815 [0.783, 0.844]	0.908 [0.888, 0.923]
9	S4	0.172 [0.091, 0.251]	0.271 [0.171, 0.383]
16	S4	0.624 [0.524, 0.716]	0.574 [0.424, 0.693]
23	S4	0.854 [0.801, 0.897]	0.855 [0.811, 0.896]
23	S5	0.939 [0.909, 0.965]	0.924 [0.903, 0.943]
23	Full	0.947 [0.916, 0.971]	0.937 [0.917, 0.955]

Table 23: Paired Full–S5 tracking increments with 95% document-bootstrap intervals. Earlier suffix paths add substantially more tracking information at layer 3.

Layer	Qwen2.5-1.5B	Qwen2.5-7B
3	0.692 [0.606, 0.775]	0.710 [0.658, 0.767]
23	0.008 [0.005, 0.011]	0.013 [0.008, 0.019]

Numerical checks and execution layout. Identical-layout baseline repeats agree exactly. Direct KV and complete suffix-injection JVPs agree within a maximum relative logit-JVP error of 3.70×10^{-5}. Recomputed 1.5B S4, S5, and Full values agree with their frozen counterparts to within 10^{-4} maximum relative error across the matched records. The 7B signed path partition uses FP64 subtraction and Fisher accumulation; model forwards and JVPs remain FP32. This arithmetic refinement resolves a 1.15×10^{-7} relative partition residual encountered with FP32 subtraction and leaves S5 and Full unchanged. The initial partial records remain in the archive.

Finite differences are more sensitive to execution layout. In a follow-up audit on the first frozen document and reader, we evaluate 32 fixed positions at layers 3 and 23 in both channels. Computing all suffix logits produces different finite differences from computing only final logits, even at the same batch size of 32. Pooling both channels within each layer, the median all-logits/final-logits ratio is 4.16 at layer 3 and 3.44 at layer 23. Repeating the all-logits calculation at batch sizes 4 and 16 also changes the values. The matched model comparison uses one common layout; the original frozen results remain separate.

Across the same 32 fixed positions per layer/channel cell, all-logits versus final-logits Spearman correlations are .934 for layer-3 keys, .858 for layer-3 values, .943 for layer-23 keys, and .933 for layer-23 values. The corresponding median all-logits/final-logits sensitivity ratios are $3.06\times$, $5.00\times$, $2.56\times$, and $4.10\times$, respectively. Position ranks are therefore more stable than absolute finite-difference magnitudes, but are not layout invariant. These correlations are descriptive diagnostics over 32 fixed positions per cell, computed from existing records without pooling or additional model forwards.

Scope. The replication tests parameter-scale extension within one model family at fixed depth, on eight documents and 512-token prefixes. It does not establish cross-architecture, long-context, or discrete-eviction generalization. The supplementary material includes replication measurements, statistical summaries, and numerical audit results.