THE SHADOW OF A FEW KEYS: LOCAL LOW-DIMENSIONAL STRUCTURE IN LONG-CONTEXT ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient long-context attention rests on an implicit chain of assumptions: a low-dimensional query–key interaction should admit a small separable representation, that representation should give cheap and accurate attention, and accurate attention should preserve the model's answers. We test this chain link by link in two frozen language models, using local coordinates adapted to each block of queries. The first link fails: better local directions sharply reduce the error of an exact-exponential diagnostic, yet finite-state kernel summaries on the same coordinates recover only a small part of that gain. We call this discrepancy the *geometry–realization gap* and then locate it. A small set of high-mass keys, under one percent of the history, suffices to rebuild nearly all of the useful local geometry, and computing exactly on those same keys closes most of the gap, an effect that survives controls for projection error, attention sinks and the choice of support. We call this the *few-key shadow*. Two boundaries follow. A low-dimensional remainder often pays off under matched storage but never under matched arithmetic in our tests, and the support that reconstructs average geometry is far smaller than what the answer-determining queries need, so a frozen fresh-input evaluation still falls short of task preservation on three models. The results separate geometric compressibility, computational realization and task fidelity, and show that the first does not certify the other two.

1 INTRODUCTION

Attention over the history is the quadratic part of long-context inference, and two families of methods promise relief by naming a structure to exploit. Sparse attention keeps a few keys (Zhang et al., 2023; Li et al., 2024; Lai et al., 2025; Yuan et al., 2025; Lu et al., 2025); low-rank and kernel methods keep a few directions and compute through them with a separable feature map (Katharopoulos et al., 2020; Choromanski et al., 2021; Nauen et al., 2025; Arora et al., 2024; Kacham et al., 2024); hybrids combine the two (Chen et al., 2021; Zhang et al., 2026; Wang et al., 2026; McDermott et al., 2026). Behind the second family lies a chain of implications that is rarely tested link by link: a low-dimensional query–key interaction should admit a small separable representation, that representation should give cheap and accurate attention, and accurate attention should preserve the model's answers. This paper tests the chain in a setting built to favor it, and finds each link failing in a specific, measurable way.

For each block of queries we adapt a basis to the key distribution tilted toward the block's first query, and we separate two decisions that are usually evaluated together (Zhang et al., 2024; Farzam et al., 2026): which directions to retain, and how to realize the exponential interactions along them. The first link breaks here. At rank eight, replacing a frozen global basis by the local one lowers the error of the exact-exponential diagnostic by .0225 and .0219 on Llama and Qwen, but the error of a separable quadratic summary on the very same coordinates by only .0025 and .0055, and higher-order realizations recover much of the rest only with thousands of states. We call the difference between a realization and the diagnostic on the same coordinates the *geometry–realization gap* (Section 3). Section 3 also shows that output error is governed by the anchor-weighted kernel error rather than by the largest pointwise error, so the question is where the weighted error is concentrated.

It is concentrated on the same few keys that identify the useful geometry. A basis rebuilt from the highest-mass $1/128$ of the history captures almost all the variance available to the full-history basis

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

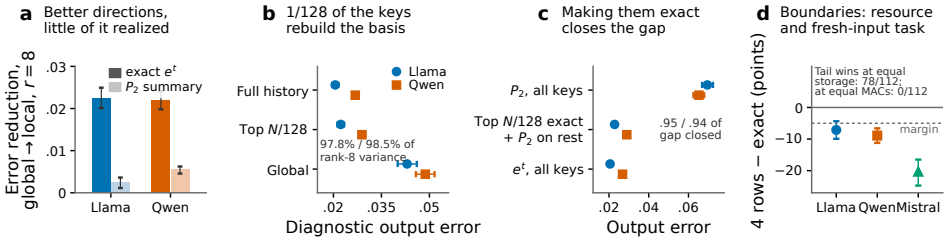

Figure 1: **Overview of the main results** (test split). (a) At rank eight, the local basis reduces the exact-exponential error far more than the P_2 summary's: the geometry–realization gap. (b) A basis from the top $N/128$ keys nearly matches the full-history basis and captures about 98% of its rank-eight variance. (c) Making those keys exact, with P_2 elsewhere, closes about 95% of the gap (a full-information intervention). (d) A low-dimensional remainder wins at equal storage and never at equal arithmetic, and frozen fresh inputs miss the five-point margin on three models.

of the same dimension and, in both models and in most of their 1,824 attention heads, nearly the same output diagnostic, whereas equal-sized uniform sets and the entire low-mass tail do not. And in a full-information intervention, evaluating just those keys exactly and keeping the polynomial for the remaining 99% of the history closes about 95% of the gap between the summary and the diagnostic (Section 4). We call this the *few-key shadow*: a small high-mass support identifies directions along which the rest of the history also varies.

The shadow changes the budget question (Figure 1). Rather than asking a low-order kernel to represent every interaction, spend the budget on exact interactions with a compact high-mass set and place the rest in one tilted mean state inside the same normalization. Both ingredients have precedents (Chen et al., 2021; Mitchell & Kersting, 2025; Jones et al., 2026; Li et al., 2026b; Hoshi et al., 2026; Li et al., 2026a; Huang et al., 2026); the anchor gate is a test of the explanation. It beats the standalone summaries we test at matched storage and arithmetic, and it settles a question that "sparse versus low-rank" cannot: a low-dimensional term on the remainder usually wins under a storage constraint and, in every tested configuration, loses under an arithmetic one (Section 5).

The last link breaks in full models (Section 6). Selecting keys from several query rows restores most of what a single block-first selector loses, with and without look-ahead. But the support that reconstructs average local geometry from $1/128$ of the history is not the support that the question rows need: holding their error at the level of the full-model runs takes about a quarter of the history, and a frozen fresh-input evaluation, on a third model as well, still misses a five-point non-inferiority margin. Operator fidelity and decision fidelity are different achievements (Yang & Li, 2026), and this paper measures the distance between them.

Scope. We study a structural mechanism rather than a new sparse–low-rank primitive or a faster decoding system: the restricted-basis and exact-set experiments are full-information diagnostics, the gated representation still scans the history for every block with the full KV cache retained, and we claim neither subquadratic prefill nor task equivalence. The support is reusable across a block without being invariant, and on one model much of it is the attention sink (Sections 4 to 6).

The paper makes three contributions. **Geometry versus realization:** on the same coordinates, better local directions cut the exact-exponential error far more than any tested finite-state realization (Section 3). **The few-key shadow:** a basis rebuilt from $1/128$ of the history recovers nearly all the useful variance in most of 1,824 heads, and a full-information intervention localizes about 95% of the realization gap on those keys (Section 4). **Resource and task boundaries:** a richer remainder wins under matched storage and in no tested case under matched arithmetic, and full-model checks, including frozen fresh inputs, locate where local support stops serving the answer (Sections 5 and 6).

2 Setting and Notation

Scope. Our operator study uses captured queries, keys and values from Llama-3.2-3B (Grattafiori et al., 2024) and Qwen3-8B-Base (Yang et al., 2025) on five long-context tasks: FWE, NIAH and VT in the style of RULER (Hsieh et al., 2024), and HotpotQA and 2WikiMQA from LongBench (Yang et al., 2018; Ho et al., 2020; Bai et al., 2024). Model weights remain frozen; captured queries and keys retain RoPE and model-specific normalization. Inputs contain up to 16K tokens; disjoint training,

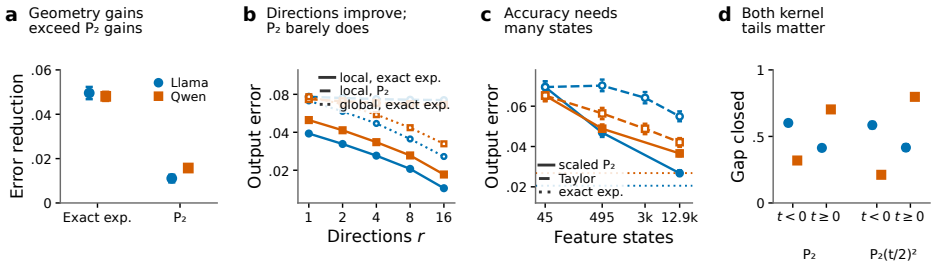

Figure 2: **A small kernel summary realizes only part of the gain from better directions.** (a) Frozen global $r = 1$ to local $r = 8$: exact-exponential error falls much more than P_2 error (paired, test). (b) Over r, the diagnostic keeps improving while P_2 barely moves (development). (c) At local $r = 8$, approaching the diagnostic (dotted) takes many more states; a state holds $d_v + 1$ numbers. (d) Restoring e^t on only negative or only positive interactions closes a model-dependent share of the P_2 and $P_2(t/2)^2$ gaps.

calibration and test splits (the first two form the development split) leave 20 test items per model. We capture two query heads fixed by index in every layer, repeat the central geometric checks on all heads, and evaluate the complete model with every head modified.

A local coordinate system. Fix a head with scale β and a block of 32 queries following N history keys. Its first query q_a defines $s_a(j) = \beta q_a^\top k_j$, the history distribution $p_a(j) = e^{s_a(j)} / \sum_{\ell=1}^{N} e^{s_a(\ell)}$, and the tilted mean $\bar{k} = \sum_j p_a(j) k_j$. Writing $u_i = \sqrt{\beta} R^{-1}(q_i - q_a)$ and $w_j = \sqrt{\beta} R(k_j - \bar{k})$ gives the exact identity

$$\beta q_i^\top k_j = s_a(j) + \underbrace{\beta (q_i - q_a)^\top \bar{k}}_{c_i} + u_i^\top w_j. \tag{1}$$

Here R is the symmetric square root of the eigenvalue-floored training second moment of query displacements, fixed per head and layer; it rescales query displacements without whitening the keys. The factor e^{c_i} must be retained when history competes with exact causal keys inside the block.

Directions versus implementation. Let $C = \sum_j p_a(j)(k_j - \bar{k})(k_j - \bar{k})^\top$ and $J = \sum_j p_a(j) w_j w_j^\top = \beta R C R$. The local basis U_r contains the leading r eigenvectors of J, estimated separately for each window; training-frozen and random bases provide controls in the same coordinates. Thus *low-dimensional structure* refers to this tilted covariance rather than to the rank of the full attention matrix. For the projected interaction $t_{ij} = u_i^\top U_r U_r^\top w_j$, the *exact-exponential diagnostic* evaluates $e^{t_{ij}}$ key by key; the separable implementation instead uses $P_2(t) = 1 + t + t^2/2$. At $r = 8$, the latter has $1 + r + r(r+1)/2 = 45$ feature states. Both preserve s_a, c_i and exact in-block interactions, isolating the choice of interaction kernel.

Evaluation. Operator error is $\|\hat{o}_i - o_i\|_2 / V_{\max}$, where $V_{\max} = \max_j \|v_j\|_2$ over the legal prefix, with matched-arithmetic references. This L2/Vmax metric differs from output-relative error and task performance. End-to-end tests report task scores and layerwise output-relative errors on the approximate run's own inputs instead of accumulated drift from an exact model. Task-weighted comparisons use item-level uncertainty; stored numbers, multiply-accumulates and key-pair counts are separate resource accounts and imply no measured speedup. Appendices A, B and G give the construction, protocols, statistical conventions (Field & Welsh, 2007; Wellek, 2010) and cost ledger.

3 Right Directions, Limited Returns

If a few directions retain the important interactions, can a small summary turn them into accurate attention? We separate these two questions using the same projected coordinates of Section 2. The *exact-exponential diagnostic* evaluates e^t without approximating the nonlinearity. It tests what projection alone changes but still touches every projected interaction, and it serves as a reference rather than an error floor (Appendix C). The executable alternative replaces e^t by the separable, positive quadratic $P_2(t) = 1 + t + t^2/2$, the second-order Taylor kernel used by TaylorShift (Nauen et al., 2025).

The directions help more than the summary. On the same test rows, moving from the frozen global $r = 1$ basis to the local $r = 8$ basis reduces mean diagnostic error by .0496 on Llama and .0481 on Qwen. The corresponding P_2 reductions are .0111 and .0158, both significant under paired cluster bootstrap (Figure 2(a)). Better directions do *not* make the operator worse on average; the discrepancy is between the gains. Comparisons at fixed $r = 8$ and local rank sweeps give the same ordering: from $r = 1$ to $r = 16$ the diagnostic error falls by more than half while P_2 falls by only 6% on Llama and 10% on Qwen (Figure 2(b), Appendix C).

A finite-budget obstacle rather than an impossibility. Keeping the local $r = 8$ coordinates fixed, we vary the realization over polynomial, Chebyshev and random-feature families (Figure 2(c), Appendix C). With at most 495 states the best non-oracle realization, $P_2(t/2)^2$, lowers the mean error from .0695 to .0469 on Llama and from .0653 to .0488 on Qwen, closing only .46 and .43 of the gap to the diagnostic for an elevenfold state count; the fourth-order Taylor polynomial with the same states does no better. Sign-specific counterfactuals show that excessive weight on negative and insufficient weight on positive interactions both contribute, in model-dependent proportion (Figure 2(d)).

Larger summaries recover more: the degree-eight $P_2(t/4)^4$ closes .873 and .746 of the gap but needs 12,870 states, 1.66 million numbers per block against 2.29 and 2.32 million for the full history's keys and values. The issue is how much representation a positive separable kernel needs here rather than whether one can be accurate; worst-case lower bounds exist for nonnegative feature maps and bounded states (Opazo, 2026; Haris & Onak, 2025), and we measure the cost on the models' own coordinates rather than proving one.

What the output error depends on. A pointwise guarantee $P(t) \approx e^t$ at every key is sufficient but not necessary. Let a realization give history key j the unnormalized weight $e^{s_a(j)+c_i} g_{ij}$ with $g_{ij} \geq 0$ and keep the legal block keys exact, and let $\Gamma_i = \sum_j p_a(j)\,|g_{ij} - e^{u_i^\top w_j}|$ be its anchor-weighted kernel error.

Proposition 1 (Weighted kernel error)**.** *With $M_i = \sum_j p_a(j) e^{u_i^\top w_j}$ and h_i the share of row i's causal mass on the history, $\|\hat{P}_i - P_i\|_1 \leq 2h_i\Gamma_i/M_i$ and $\|\hat{o}_i - o_i\|_2 \leq 2h_i V_{\max}\Gamma_i/M_i$ (proof in Appendix C).*

Each key enters Γ_i in proportion to its anchor mass, and a key evaluated exactly leaves it, whereas a uniform guarantee is expensive (Aggarwal & Alman, 2022): on the range of t of the median window, $|P(t)e^{-t} - 1| \leq .1$ needs degree 9 and 10, hence at least 11,440 and 19,448 coordinates for an exact separable realization at $r = 8$ (Remark 2). The bound is loose by an order of magnitude but orders rows by error (Appendix C), and it says where to look: where the anchor places its mass. Before increasing the representation again, we therefore ask: *which history keys supply that local geometry, and are they where the polynomial fails?* Section 4 answers both.

4 The Shadow of a Few Keys

If good local directions do not readily become a good finite-state kernel, what do they represent? We test a *shadow* explanation: much of the effective local geometry can be reconstructed from a small set of keys carrying high anchor mass, and making those keys exact removes most of the realization gap of Section 3.

4.1 Rebuilding the local basis from a small support

For each window, let $\mathcal{H}$ contain the top $m = N/128$ history keys under p_a. Replace the full tilted covariance J by $J_{\mathcal{H}} = \sum_{j\in\mathcal{H}} p_a(j)\, w_j w_j^\top$ and take its leading eight eigenvectors as the *restricted basis* $B_{\mathcal{H}}$. Only the covariance summands change: weights keep their scale, deviations keep the full-history center $\bar{k}$, R stays fixed from training, and the same exact-exponential diagnostic runs. It is a full-information diagnostic: finding $\mathcal{H}$ and $\bar{k}$ reads the whole history, and every projected interaction is still evaluated.

Despite using only $1/128$ of the summands, the restricted basis gives errors of .0223 and .0289 on Llama and Qwen, close to the full-history basis at .0205 and .0268 and far below the frozen global basis at .0430 and .0487 (Figure 3(a)). Centering at the selected keys' own mean retains most of the advantage over the global basis, so the full-history center is not carrying the result. Across all 1,824

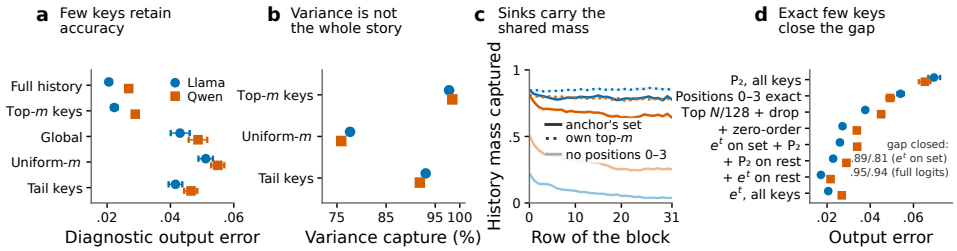

Figure 3: **A few high-mass keys reconstruct useful local geometry, and making them exact closes most of the realization gap** ($m = N/128$, test split). (a) Diagnostic error of bases built from different key sets. (b) Median variance capture relative to the optimal rank-eight basis: the tail basis captures much variance yet errs more. (c) Share of each query's history mass inside the anchor's top-m set (solid) and its own top-m set (dotted), by row; faint lines exclude sequence positions 0–3. (d) Error against P_2 and e^t on all keys: positions 0–3 exact, or the top $N/128$ keys exact with the remainder dropped, summarized by a zero-order state, or given P_2 (with full logits or only e^t on the exact keys) or e^t on the same eight directions.

head–layer pairs, the restricted-basis error exceeds the full-history error by at most .003 in 74% of Llama pairs and 72% of Qwen pairs, and beats the global basis in 97% and 96%; the exceptions are heads with diffuse anchor distributions (Appendix D).

4.2 Recovering useful variance rather than an identical subspace

The reconstruction has a geometric counterpart. Evaluate the restricted basis on the *full* covariance through $\mathcal{C}_{\mathcal{H}} = \mathrm{tr}(B_{\mathcal{H}}^{\top} J B_{\mathcal{H}}) / \sum_{\ell=1}^{8} \lambda_\ell(J)$. Its median is 97.8% on Llama and 98.5% on Qwen; these percentages refer to the variance captured by the *optimal rank-eight basis* rather than to all variance in J (Figure 3(b)). In this sense the top $1/128$ is *geometry sufficient* at level $\varepsilon \approx .02$ in the median window. Selected keys recover directions along which unselected keys also vary, and their share of directional energy is below their share of anchor mass; the largest principal angle (Björck & Golub, 1973) between the two subspaces is often large while the leading direction aligns, because rotation near the eighth-eigenvector cutoff costs little variance (Appendix D).

4.3 Testing the connection between directions and keys

Not every small set works: equal-sized uniform sets and the next-ranked keys give much worse bases, as does the entire low-mass tail, which captures much variance yet raises the error to .0416 and .0465, close to the global basis (Figure 3(a,b)). Sampling keys in proportion to p_a nearly matches the top-m basis and mass-matched samples work as well (Appendix D): what the shadow needs is a small set carrying most of the anchor mass. A few-key representation's error also rises and falls with the few-direction diagnostic across windows, with partial rank correlations of .808 and .809 after controlling for measured difficulty (Appendix E).

4.4 Making the same keys exact closes most of the gap

We test whether the keys that reconstruct the geometry also account for the realization error of Section 3, since a basis preserves their interactions and those are the ones most likely to leave the range in which P_2 tracks e^t (Figure 3(d)). Give the top $N/128$ keys their full logits, as the gate of Section 5 does, and keep P_2 on the same eight local directions for the remaining 99% of the history: the error falls from .0695 and .0653 to .0228 and .0290, against .0205 and .0268 for the exact exponential on every projected interaction. Measured against that diagnostic, $1/128$ of the keys thus closes .954 and .944 of the realization gap, more than the degree-eight polynomial with 12,870 states (Appendix C). With the same set, dropping the rest gives .0377 and .0450, the zero-order state of Section 5 .0272 and .0338, and e^t on the rest .0172 and .0216; what the polynomial still misses is spread over the tail (Appendix C). The correction restores the omitted directions on those keys as well as the exponential, and the two separate (Appendix C.2): keeping the projection and replacing only $P_2(t)$ by e^t on $\mathcal{H}$ closes .890 and .815 of the gap (.0259 and .0339), .93 and .86 of the full-logit reduction, and the omitted directions on those keys account for the remaining .0031 and .0050. Equal-sized uniform and next-ranked exact sets close at most .12 of the gap, sets drawn by

anchor mass .900 and .869, and a per-row oracle of the m largest weighted kernel errors 1.08 and 1.20: the top-mass set is close to the best m keys.

Reusable within limits. The support serves a block unequally, and attention sinks (Xiao et al., 2024; Gu et al., 2025) carry much of the shared mass: the first four positions of the sequence hold a median .83 of a query's history mass on Llama and .36 on Qwen, and .97 and .55 of the mass that the anchor's set captures (Figure 3(c)). Sinks form part of the shadow without being the whole of it (Appendix C.2). Positions 0–3 alone, made exact with P_2 elsewhere, close .316 and .418 of the gap; with them exact in every arm, the rest of $\mathcal{H}$ still lowers the error by .0314 and .0205, .641 and .533 of the gap. A basis rebuilt from the top $N/128$ keys outside positions 0–3 still captures .969 and .980 of the rank-eight variance and beats the global basis by .0158 and .0166, though its error of .0273 and .0321 exceeds the sink-included basis by .0050 and .0032; making those keys exact closes .662 and .654 of the gap. On Llama the sinks are the larger part of the support; on neither model are they the explanation, and keeping only the first positions exact, as sink-preserving designs do (Zhu et al., 2026b; Su & Yuan, 2025), closes a third to two fifths of the gap. Fixed-row comparisons also show the anchor's selection going stale as queries move (Appendix F).

5 From Local Geometry to Resource Allocation

The shadow suggests spending the budget on exact interactions with high-mass keys before building a richer summary; we instantiate this with an *anchor gate*. Let $\mathcal{H}$ contain the m history keys with largest anchor mass and $\mathcal{O}$ the remaining history. The gate evaluates $\mathcal{H}$ and the legal in-block keys exactly, while retaining

$$S_0 = \sum_{j \in \mathcal{O}} e^{s_a(j)}, \qquad S_1 = \sum_{j \in \mathcal{O}} e^{s_a(j)} v_j. \tag{2}$$

The remainder enters the same softmax as a virtual entry with value $\hat{\mu} = S_1/S_0$ and logit $\log S_0 + \beta(q_i - q_a)^\top \bar{k}$, so both numerator and denominator retain its contribution. The full block state includes the tilted key mean $\bar{k}$: $d + d_v + 1 = 257$ numbers, besides the exact keys, values and indices; the gate serves as a test of the explanation rather than as a new hybrid (Zhang et al., 2026; Hooper et al., 2025).

What the summary must preserve. For a nonempty remainder, let π_i and μ_i be its true mass and conditional value mean, $\hat{\pi}_i$ its reconstructed mass, and ν_i the conditional mean over *all* exact terms: $\mathcal{H}$ plus the legal in-block keys. Masses refer to the complete causal softmax.

Proposition 2 (Remainder error)**.** *The gate satisfies*

$$\hat{o}_i - o_i = \pi_i(\hat{\mu} - \mu_i) + (\hat{\pi}_i - \pi_i)(\hat{\mu} - \nu_i).$$

The first term measures a stale remainder value; the second measures a wrong balance between remainder and exact contributions. A small state helps when it tracks both, and a small remainder mass alone does not make it accurate (Appendix H). The remainder is also the easier target: second-order tilted moments track its log-normalizer with slopes 1.02 and 1.04 but the whole history's with slopes .38 and .60 (Appendix H.3); centering it at its own mean removes the first-order mass error and lowers test error at every budget (Appendix H), which the full-model runs below do not yet use. At $m = N/128$, adding this state reduces test L2/Vmax error from .0377/.0450 to .0272/.0338 (Llama/Qwen; Figure 3(d)). These are *same-set* rather than equal-cost comparisons: at small budgets, spending the summary's arithmetic on more exact keys can be better.

Does the explanation survive resource matching? Yes, for the tested representations. The local eight-direction P_2 summary stores 6,957 numbers per block. The same storage buys the gate 26 exact keys and its zero-order state, yielding .0293/.0382 error instead of .0695/.0653. Across the tested local and frozen-global bases and $r \in \{1, 2, 4, 8\}$, the gate is more accurate than standalone P_2 at matched storage and within matched matrix-product arithmetic budgets. On the final test windows, every separable realization of Section 3 with more than 45 states is less accurate than the gate at some budget that stores fewer numbers per block, and the 45-state ones at every budget (Figure 4(c), Appendices C.1 and G).

The conclusion changes when low-dimensional states serve *only the remainder*. Across two selectors, seven exact-set budgets, exact-set and frozen-global bases, and four ranks, the P_2 tail beats a zero-order gate enlarged to the same storage in 78 of 112 configurations, with paired intervals favoring it

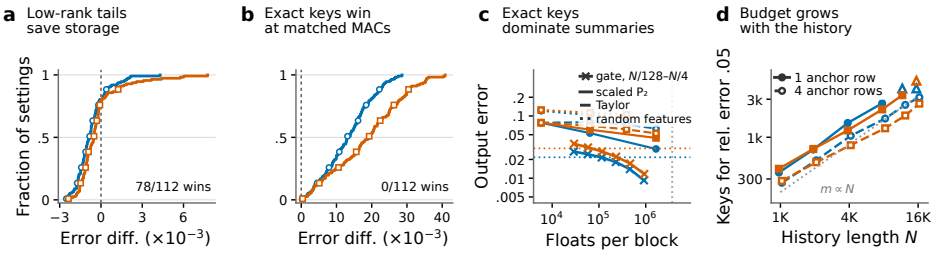

Figure 4: **Two resource constraints, two answers; exact keys dominate every tested standalone summary.** (a, b) Cumulative fraction of 112 configurations by the error of a low-dimensional tail minus that of a zero-order gate enlarged to the same storage (a) or arithmetic (b); negative favors the tail, a win needs paired intervals below zero on both models, and the scales differ. (c) Error against numbers stored per block (block representation rather than KV-cache memory; final test windows): the gate at $N/128$ to $N/4$ against the separable realizations of Section 3, counted as states times $d_v + 1$ without their eight directions and center, a count that favors the summaries; dotted lines mark the diagnostic and the full history. (d) Exact keys that hold a relative error of .05 as the history grows; triangles mark lengths the grid did not reach (Appendix G).

on both models. At matched multiply-accumulates, it wins none: more exact keys are a better use of that arithmetic (Figure 4(a,b)). A low-dimensional remainder can thus be storage-efficient without being arithmetic-efficient, under the tested 32-query, per-block rebuild schedule; sparse and low-rank terms remain complementary (Chen et al., 2021), and their value depends on the constrained resource.

What the allocation does not change. The saving is a constant factor rather than an order of growth. The exact budget that holds a relative error of .05 grows about in proportion to the history up to 16K on Llama and nearly so on Qwen, with fitted exponents between .78 and .98 (Figure 4(d)), and more slowly beyond 16K on longer synthetic inputs (Appendix G.6). Every block still reads the whole history for its anchor rows and remainder state, so a favorable MAC count is not a favorable byte count: a fused two-pass kernel at the main operating point would move 5.1 to 5.4 times the bytes of a dense tiled kernel (Dao et al., 2022) per query (Appendix G.3), and our unfused implementation is slower than dense attention (Appendix G.4).

6 The Geometry–Task Gap

Support sufficient for average local geometry need not serve the queries that determine the answer; this section tests that last implication. We insert the representation of Section 5 into every head and layer, using 32-query blocks and $m = N/4$ history keys. Each model processes 200 development inputs of up to 16K tokens. E4 equally weights NIAH and FWE accuracy and HotpotQA and 2WikiMQA F1, with 40 inputs per task; floor-level VT remains a structural check (Appendix I).

The budget at the question-row target. The budget is far above the $N/128$ of the geometric diagnostic because the target is different and stricter (Figure 5(a), Appendix G). The diagnostic tolerates an L2/Vmax error of .03 averaged over all rows of a window, which at most about a hundred keys meet up to 16K. The prefill rows of the question tokens, the *question rows*, are harder: at $N/4$ their relative error is .124 and .157 with the block-first anchor and .058 and .047 with four anchor rows; holding it at the full-model level takes about a quarter of the history with four rows and is out of reach for a single anchor within $N/4$. This is an operating-point measurement for one selector, metric and target rather than a lower bound on task-relevant support (Nawrot et al., 2026; Luo et al., 2026; Shi & Gu, 2026).

Finding support for later queries. A block-first query can miss keys needed by the question later in the block. We score the history with queries at positions $0, 10, 21, 31$ and select the m keys with highest maximum probability across them; the remainder retains the block-first tilt. Four rows recover 13.1 of the 15.3 lost E4 points on Llama and 21.5 of 26.5 on Qwen (Table 1(a)). NIAH accuracy rises from .525 to .950 and from .200 to .850, respectively, Qwen's FWE accuracy rises from .775 to .975, and the other tasks move within their intervals (Figure 5(b,c)): refreshing the selection signal beats reusing one query's support. The six Qwen NIAH inputs that still fail have the same mean

Table 1: **Selection improves task scores, but does not guarantee preservation.** Points; brackets: 95% task-stratified paired-bootstrap intervals (2,000 resamples). (a) Development E4. Cost counts key pairs rather than runtime; drop uses $71N/256$ keys to match the four-row mean arm's MAC account instead of its key-pair count. Other mean arms use $N/4$; † marks look-ahead. (b) Fresh, frozen task averages: E4 plus TREC for Llama/Qwen; HotpotQA, NIAH and TREC for Mistral. n counts primary-endpoint inputs. The four-row arm is primary; causal is secondary.

(a) Development		Llama-3.2-3B		Qwen3-8B-Base	
Selector / remainder	Cost	E4	Δ [95% CI]	E4	Δ [95% CI]
Exact attention	1.000	58.9	—	72.0	—
1 row / mean	.300	43.6	−15.3 [−22.4, −8.4]	45.5	−26.5 [−31.8, −20.7]
4 rows† / mean	.393	56.6	−2.3 [−7.0, +2.5]	67.0	−5.0 [−8.9, −1.2]
4 rows, causal / mean	.393	53.8	−5.1 [−11.4, +1.0]	64.0	−8.0 [−12.5, −3.7]
4 rows† / drop	.405	55.1	−3.8 [−9.6, +2.1]	67.6	−4.4 [−8.4, −0.9]

(b) Fresh inputs	n	4 rows†: Δ [95% CI]	Causal: Δ [95% CI]
Llama-3.2-3B	807	−7.1 [−9.9, −4.4]	−12.3 [−15.4, −9.4]
Qwen3-8B-Base	808	−8.9 [−11.2, −6.6]	−12.0 [−14.3, −9.6]
Mistral-7B-v0.3	368	−20.3 [−24.7, −16.5]	−18.5 [−22.7, −14.4]



Figure 5: **Full models: selection matters, and fresh inputs show where the loss is.** (a) Relative error on the question rows of the final block against the exact budget at $N \approx 16$K (operator level, 24 synthetic sequences); dotted lines mark the full-model question-row error at $N/4$. (b, c) Development task scores of exact attention and the one-row and four-row selectors at $m = N/4$, with paired 95% intervals. (d) Four rows minus exact on the fresh inputs of Table 1(b), by task; dashed: the five-point margin.

question-row error, .063 to three decimals, as the 34 answered correctly, so average question-row error does not flag them (Appendix I).

Recovery without look-ahead. The four-row selector uses later prompt queries to select earlier rows' keys (prefix-perturbation tests confirm this). A progressive variant adds each anchor only when reached, rebuilding the selected set for the following sub-block. It recovers 10.2 and 18.5 E4 points without look-ahead. The remaining gaps from the four-row arm are 2.8 and 3.0 points, with intervals including zero (Appendix I). These are development comparisons rather than non-inferiority tests.

What the remainder contributes. At fixed m, the zero-order remainder lowers question-row error relative to dropping unselected keys. At equal MACs, its mean-minus-drop E4 differences are $+1.5$ $[-3.1, +6.5]$ and -0.6 $[-3.1, +1.7]$ points: a consistent task benefit is not established. Local errors compare against exact attention on each gated layer's own inputs instead of the exact model's trajectory. Better local fidelity need not preserve the final decision.

Fresh inputs sharpen the boundary. We froze an additional evaluation on fresh inputs, a new task family and Mistral-7B-v0.3 (Jiang et al., 2023) before running the gated arms (Appendix K). On Mistral, four rows improve over one by 13.3 [8.8,17.8] points, so the selection benefit extends beyond the development models. However, the four-row representation misses the frozen five-point non-inferiority margin on all three models, with losses of 7.1, 8.9 and 20.3 points (Table 1(b); Holm correction across models (Holm, 1979)). The losses are largest on NIAH for Qwen and Mistral and on TREC and FWE for Llama, and the new TREC family (Li & Roth, 2002) loses at least nine

points on all three models (Figure 5(d)); the frozen task sets differ across models and do not support cross-model rankings.

At longer contexts, Llama's 64K NIAH/FWE average loses 20 points, while Qwen's 32K scores are confounded by answer truncation under the fixed generation cap (Appendix J).

7 Related Work

From geometry to a kernel. Linear attention and Performer obtain compact running states through separable feature maps (Katharopoulos et al., 2020; Choromanski et al., 2021), TaylorShift uses a polynomial approximation (Nauen et al., 2025), and theory relates approximation complexity to logit magnitude and interaction dimension (Alman & Song, 2023; Sarlos et al., 2023; Haris & Onak, 2025; Opazo, 2026; Sui & Zhang, 2026); we separate geometric compression from kernel realization experimentally on the same coordinates. Keys are locally low-rank (Hwang, 2026), and low-rank bases can be built from a subset of tokens (Han et al., 2026; Bhojanapalli et al., 2021); Section 4 identifies which subset suffices for the anchor-tilted covariance and shows that the same keys carry the realization gap.

Selecting support for several queries. SnapKV votes with an observation window (Li et al., 2024), QuoKA selects keys for each prefill chunk from a subset of its queries (Jones et al., 2026), and MInference and FlexPrefill estimate sparse patterns from the last prompt queries or pooled query blocks (Jiang et al., 2024; Lai et al., 2025). Our block selectors belong to this family.

Exact interactions and summarized remainders. Scatterbrain corrects a low-rank kernel estimate exactly at selected sparse positions (Chen et al., 2021); improved clustered attention selects shared keys by query centroids and keeps centroid weights elsewhere (Vyas et al., 2020); MuSe combines exact retrieval with exponentially tilted cluster summaries (Mitchell & Kersting, 2025); SLA and ELSAA compute a sparse high-weight part exactly and a low-rank rest linearly, ELSAA under one denominator (Zhang et al., 2026; Heidari et al., 2026); ResKV and MomentKV reconstruct omitted contributions from compact statistics (Zhan et al., 2026; Li et al., 2026b); concurrent systems route high-error chunks to exact attention and summarize the rest under one normalization (Li et al., 2026a; Huang et al., 2026; Hoshi et al., 2026), and Support Basis computes large entries exactly and a polynomial on the rest (Aliakbarpour et al., 2026). Neither mixing exact and approximate interactions nor tilting a remainder toward a representative query is new (Appendix L); what they do not report, and Sections 4 and 5 measure, is which support reconstructs the geometry, where a compact realization fails, and how the allocation ranks under matched storage and arithmetic.

8 Conclusion

We asked whether the low-dimensional structure of long-context attention can be turned into cheap and faithful computation, and we tested the chain of assumptions behind that hope link by link. In two frozen language models, good local directions do not translate into a good finite-state kernel, which we call the geometry–realization gap. The gap is generated by a few high-mass keys: they rebuild the useful local basis and, when computed exactly, remove most of the realization error, an effect that survives projected-only, sink and support controls. Exact support plus a summary then beats the tested standalone summaries under matched storage and arithmetic, while a richer low-dimensional remainder pays off under storage and in none of the tested arithmetic-matched configurations. The connection stops at the block: the support guarantees neither invariant sharing, task equivalence nor acceleration, and the support that reconstructs average geometry is far smaller than the one the question rows need (Appendices F, I and J). Future work should test the mechanism on further architectures and on fresh traces, use the remainder-centered state in full models, and design selectors that serve the question rows within the budget that the local geometry suggests.

AI USAGE DISCLOSURE

In this work, we used generative AI tools to assist in implementing our methods and to help write data processing scripts for cleaning and reformatting our datasets. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or assist with proofs, propose or refine hypotheses, design research methodology or experiments, support qualitative or thematic data analysis, or interpret results; translation is not applicable to this work. Additionally, we used generative AI tools to assist with writing and debugging other software code, and to edit the manuscript to improve readability (e.g., grammar, phrasing, and clarity). We have reviewed all AI-assisted work. All AI-polished text was read and checked by the authors to ensure it accurately reflects our ideas, claims, and results. AI-assisted code, including method implementations and data processing scripts, was reviewed, tested, and verified for correctness by the authors, including checking processed outputs against the raw data. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The appendix documents everything needed to reproduce the results. Appendix A gives the models and checkpoint revisions, the construction of the synthetic tasks (generator seed, filler source, record identifiers of the LongBench items), the train, calibration and test splits, the captured heads, the numerical conventions of the operator-level pipeline, the decoding settings and scoring functions of the full-model runs, and the statistical conventions (task-stratified cluster bootstrap, sign-flip tests, non-inferiority margin). Appendix B states all assumptions and gives complete proofs of the propositions and remarks, with the empirical tightness of every bound. The kernel realizations, the restricted-basis constructions, the resource accounting and the exact-set controls are specified in Appendices C, C.2, D and G, including the rules that were frozen before each control experiment was run. The full-model protocol, the causal selectors and the prospectively frozen fresh-input evaluation, with its registered endpoint, task sets and sample sizes, are in Appendices I and K. Every number in the paper is produced by archived analysis code from saved per-window and per-input result files, and the source of each reported number is recorded next to the text that reports it.

REFERENCES

Amol Aggarwal and Josh Alman. Optimal-Degree Polynomial Approximations for Exponentials and Gaussian Kernel Density Estimation. In *37th Computational Complexity Conference (CCC 2022)*, volume 234 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 22:1–22:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. doi: 10.4230/LIPIcs.CCC.2022.22. arXiv:2205.06249v1.

Maryam Aliakbarpour, Vladimir Braverman, Junze Yin, and Haochen Zhang. Support Basis: Fast Attention Beyond Bounded Entries. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 325–333. PMLR, 2026. arXiv:2510.01643v2.

Josh Alman and Zhao Song. Fast Attention Requires Bounded Entries. In *Advances in Neural Information Processing Systems*, volume 36, pp. 63117–63135. Curran Associates, Inc., 2023. doi: 10.52202/075280-2755.

Josh Alman and Zhao Song. Fast RoPE Attention: Combining the Polynomial Method and Fast Fourier Transform. *arXiv preprint arXiv:2505.11892*, 2025.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1763–1840. PMLR, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.183.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? In *Conference on Language Modeling*, 2025. arXiv:2504.02732v4.

Srinadh Bhojanapalli, Ayan Chakrabarti, Himanshu Jain, Sanjiv Kumar, Michal Lukasik, and Andreas Veit. Eigen Analysis of Self-Attention and its Reconstruction from Partial Computation. *arXiv preprint arXiv:2106.08823v1*, 2021.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973. doi: 10.1090/s0025-5718-1973-0348991-3.

Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying Sparse and Low-rank Attention Approximation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17413–17426. Curran Associates, Inc., 2021.

Zhuoming Chen, Ranajoy Sadhukhan, Zihao Ye, Yang Zhou, Jianyu Zhang, Niklas Nolte, Yuandong Tian, Matthijs Douze, Leon Bottou, Zhihao Jia, and Beidi Chen. MagicPIG: LSH Sampling for Efficient LLM Generation. In *International Conference on Learning Representations*, 2025. arXiv:2410.16179v4.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking Attention with Performers. In *International Conference on Learning Representations*, 2021.

Timothy Chu, Josh Alman, Gary Miller, Shyam Narayanan, Mark Sellke, and Zhao Song. Metric Transforms and Low Rank Representations of Kernels for Fast Attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47014–47068. Curran Associates, Inc., 2024. doi: 10.52202/079017-1491.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359. Curran Associates, Inc., 2022. doi: 10.52202/068431-1189.

Chandler Davis and W. M. Kahan. The Rotation of Eigenvectors by a Perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. *arXiv preprint arXiv:2405.04434v5*, 2024.

DeepSeek-AI. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *arXiv preprint arXiv:2512.02556v1*, 2025.

Aditya Desai, Kumar Krishna Agrawal, Shuo Yang, Alejandro Cuadron, Luis Gaspar Schroeder, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. vAttention: Verified Sparse Attention. *arXiv preprint arXiv:2510.05688*, 2025.

Harry Dong, Xinyu Yang, Zhenyu Zhang, Zhangyang Wang, Yuejie Chi, and Beidi Chen. Get More with LESS: Synthesizing Recurrence with KV Cache Compression for Efficient LLM Inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11437–11452. PMLR, 2024. arXiv:2402.09398v2.

Ky Fan. On a Theorem of Weyl Concerning Eigenvalues of Linear Transformations I. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949. doi: 10.1073/pnas.35.11.652.

Amirhossein Farzam, Hossein Mobahi, Nolan Andrew Miller, and Luke Sernau. Data-Aware Random Feature Kernel for Transformers. *arXiv preprint arXiv:2603.04127v1*, 2026.

Yuan Feng, Junlin Lv, Haoyu Guo, Yukun Cao, S Kevin Zhou, and Xike Xie. CriticalKV: Optimizing KV Cache Eviction from an Output Perturbation Perspective. In *International Conference on Machine Learning*, 2026. arXiv:2502.03805v2.

C. A. Field and A. H. Welsh. Bootstrapping Clustered Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

Yizhao Gao, Zhichen Zeng, DaYou Du, Shijie Cao, Peiyuan Zhou, Jiaxing Qi, Junjie Lai, Hayden So, Ting Cao, Fan Yang, and Mao Yang. SeerAttention: Self-distilled Attention Gating for Efficient Long-context Prefilling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 55846–55869. Curran Associates, Inc., 2025. doi: 10.52202/085713-1869.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783v3*, 2024.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When Attention Sink Emerges in Language Models: An Empirical View. In *International Conference on Learning Representations*, 2025. arXiv:2410.10781v2.

Han Guo, Songlin Yang, Tarushii Goel, Eric P. Xing, Tri Dao, and Yoon Kim. Log-Linear Attention. In *International Conference on Learning Representations*, 2026. arXiv:2506.04761v3.

Ankit Gupta and Jonathan Berant. Value-aware Approximate Attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9567–9574. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.753.

Shreya Gupta, Boyang Huang, Barna Saha, Yinzhan Xu, and Christopher Ye. Subquadratic Algorithms and Hardness for Attention with Any Temperature. In *International Conference on Learning Representations*, 2026. arXiv:2505.14840v1.

Jialong Han, You Wu, and Kewei Tu. S^4R: Selective Sampling, Subspaces, and Sparse Reconstruction for Compressed Long-Context KV Caching. *arXiv preprint arXiv:2608.00528v2*, 2026. Accepted at AACL-IJCNLP 2026 (arXiv comment).

Themistoklis Haris and Krzysztof Onak. Compression Barriers in Autoregressive Transformers. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2757–2785. PMLR, 2025. arXiv:2502.15955v1.

Mahdi Heidari, Mohammad Mahdi Rahimi, and Jaekyun Moon. ELSAA: Efficient Low-Rank and Sparse Attention Approximation for Training Transformers. *arXiv preprint arXiv:2607.20214v1*, 2026.

Franz A. Heinsen and Leo Kozachkov. Self-Attention at Constant Cost per Token via Symmetry-Aware Taylor Approximation. *arXiv preprint arXiv:2602.00294*, 2026.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Coleman Hooper, Sebastian Zhao, Luca Manolache, Sehoon Kim, Michael W. Mahoney, Sophia Shao, Kurt Keutzer, and Amir Gholami. Multipole Attention for Efficient Long Context Reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 37382–37405. Curran Associates, Inc., 2025. doi: 10.52202/085713-1255.

Yasuto Hoshi, Daisuke Miyashita, and Jun Deguchi. Residual-Mass Accounting for Partial-KV Decoding. *arXiv preprint arXiv:2604.05438v2*, 2026.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *Conference on Language Modeling*, 2024.

Zhen Huang, Ruizhe Yao, Danyi Liu, Xinrui Chen, Shuwei Li, Siru Zhong, Zijian Cao, Yushan Lai, Mingming Guo, Weijie Zheng, and Haohuan Fu. CompKV: Compensation-Aware KV Selection for Long-Context LLM Inference. *arXiv preprint arXiv:2609.26300v1*, 2026.

Junsung Hwang. LOCKS: Page-Local Compact Key Summaries for Efficient Long-Context Decoding. *arXiv preprint arXiv:2607.24555v2*, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825v1*, 2023.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. MInference 1.0: Accelerating Pre-filling for Long-Context LLMs via Dynamic Sparse Attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52481–52515. Curran Associates, Inc., 2024. doi: 10.52202/079017-1663.

Dalton Jones, Junyoung Park, Matthew Morse, Mingu Lee, Chris Lott, and Harper Langston. QUOKA: Query-Oriented KV Selection For Efficient LLM Prefill. *arXiv preprint arXiv:2602.08722*, 2026.

Sahil Joshi, Agniva Chowdhury, Amar Kanakamedala, Ekam Singh, Evan Tu, and Anshumali Shrivastava. RACE Attention: A Strictly Linear-Time Attention Layer for Training on Outrageously Large Contexts. In *International Conference on Learning Representations*, 2026. arXiv:2510.04008v5.

Praneeth Kacham, Vahab Mirrokni, and Peilin Zhong. PolySketchFormer: Fast Transformers via Sketching Polynomial Kernels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22748–22770. PMLR, 2024. arXiv:2310.01655v3.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020.

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, and Xun Zhou. FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference. In *International Conference on Learning Representations*, 2025.

Dong Le, Thong Nguyen, Cong-Duy Nguyen, and Anh Tuan Luu. Don't Read Everything: A Curvature-Conditioned Query for Linear Attention. *arXiv preprint arXiv:2606.01294v2*, 2026. EMNLP 2026 camera-ready (arXiv comment).

Siyu Li, Dong Wang, Jie Zhou, Wei Li, Yang Xu, and Sijie Song. CEDAR: Error-Bounded Residual Routing for Efficient Long-Context Attention. *arXiv preprint arXiv:2609.07237v1*, 2026a.

Xin Li and Dan Roth. Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, volume 1, pp. 1–7. Association for Computational Linguistics, 2002. doi: 10.3115/1072228.1072378.

Yu Li, Binxu Li, and Tian Lan. MomentKV: Closing the Directional Gap in KV Cache Eviction for Long-Context Inference. *arXiv preprint arXiv:2606.01563*, 2026b.

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, pp. 22947–22970. Curran Associates, Inc., 2024. doi: 10.52202/079017-0722.

Chaofan Lin, Jiaming Tang, Shuo Yang, Hanshuo Wang, Tian Tang, Boyu Tian, Ion Stoica, Song Han, and Mingyu Gao. Twilight: Adaptive Attention Sparsity with Hierarchical Top-p Pruning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 119406–119432. Curran Associates, Inc., 2025. doi: 10.52202/085713-3986.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, Zhiqi Huang, Huan Yuan, Suting Xu, Xinran Xu, Guokun Lai, Yanru Chen, Huabin Zheng, Junjie Yan, Jianlin Su, Yuxin Wu, Yutao Zhang, Zhilin Yang, Xinyu Zhou, Mingxing Zhang, and Jiezhong Qiu. MoBA: Mixture of Block Attention for Long-Context LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17790–17815. Curran Associates, Inc., 2025. doi: 10.52202/085713-0602.

Daming Luo, Christy Liang, and Junyu Xuan. How Query Visibility Changes KV-Cache Compression Rankings: A Matched-Budget Audit. *arXiv preprint arXiv:2607.11942v1*, 2026.

Luke McDermott, Robert W. Heath Jr., and Rahul Parhi. LoLA: Low-Rank Linear Attention with Sparse Caching. *Transactions on Machine Learning Research*, 2026. arXiv:2505.23666v3.

Rupert Mitchell and Kristian Kersting. Multipole Semantic Attention: A Fast Approximation of Softmax Attention for Pretraining. *arXiv preprint arXiv:2509.10406*, 2025.

Tobias Christian Nauen, Sebastian Palacio, and Andreas Dengel. TaylorShift: Shifting the Complexity of Self-attention from Squared to Linear (and Back) Using Taylor-Softmax. In *Pattern Recognition*, volume 15306 of *Lecture Notes in Computer Science*, pp. 1–16. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-78172-8_1. ICPR 2024.

Piotr Nawrot, Jianing Li, Renjie Huang, Sebastian Ruder, Kelly Marchisio, and Edoardo Ponti. The Sparse Frontier: Sparse Attention Trade-offs in Transformer LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38667–38701. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1926.

Kasper Nordenram and Amelie Dittmann. ClusterAttention: A training-free speedup of bidirectional attention. *arXiv preprint arXiv:2608.26965*, 2026.

Vicente Opazo. Three Tokens Force Exponential Feature Rank in Nonnegative Kernel Attention. *arXiv preprint arXiv:2608.11427v1*, 2026.

Theodore J. Rivlin. *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory*. John Wiley & Sons, 2nd edition, 1990. ISBN 0471628964.

Tamas Sarlos, Xingyou Song, David Woodruff, and Qiuyi Zhang. Hardness of Low Rank Approximation of Entrywise Transformed Matrix Products. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52568–52582. Curran Associates, Inc., 2023. doi: 10.52202/075280-2289.

Utkarsh Saxena, Gobinda Saha, Sakshi Choudhary, and Kaushik Roy. Eigen Attention: Attention in Low-Rank Space for KV Cache Compression. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15332–15344. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.899.

Jack Shi and Jerry Gu. Trust the Mass: Forced Weights in KV-Cache Eviction. *arXiv preprint arXiv:2608.25230v2*, 2026.

Prajwal Singhania, Siddharth Singh, Shwai He, Soheil Feizi, and Abhinav Bhatele. Loki: Low-rank Keys for Efficient Sparse Attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 16692–16723. Curran Associates, Inc., 2024. doi: 10.52202/079017-0532.

Jiwon Song, Dongwon Jo, Beomseok Kang, and Jae-Joon Kim. CompactAttention: Accelerating Chunked Prefill with Block-Union KV Selection. *arXiv preprint arXiv:2605.16839v1*, 2026.

Zunhai Su and Kehong Yuan. KVSink: Understanding and Enhancing the Preservation of Attention Sinks in KV Cache Quantization for LLMs. In *Conference on Language Modeling*, 2025. arXiv:2508.04257v1.

Yuhe Sui and Jianing Zhang. The Approximation Rank of Softmax Attention: Sharp Geometric Laws and Robust Interaction Dimension. *arXiv preprint arXiv:2608.28150v1*, 2026.

Pardis Taghavi, Reza Langari, and Gaurav Pandey. Partition the Support, Reconstruct the Residual: Training-Free Sparse Attention for Video Generation and World Models. *arXiv preprint arXiv:2608.18484v1*, 2026.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Danning Ke, Shikuan Hong, Yiwu Yao, and Gongyi Wang. RazorAttention: Efficient KV Cache Compression Through Retrieval Heads. In *International Conference on Learning Representations*, 2025. arXiv:2407.15891v1.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-Aware Sparsity for Efficient Long-Context LLM Inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47901–47911. PMLR, 2024.

Alexander Tian, Aditya Ghai, Sanjit Neelam, Zaal Vasania, and Akshay Mishra. COBS: Cumulant Order Block Sparse Attention. *arXiv preprint arXiv:2607.09052*, 2026.

Georgios Tzachristas, Lei Deng, Ioannis Tzachristas, Gong Zhang, and Renhai Chen. A Mathematical Theory of Top-k Sparse Attention via Total Variation Distance. In *2026 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6. IEEE, 2026. doi: 10.1109/ISIT62367.2026.11653946. arXiv:2512.07647.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast Transformers with Clustered Attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21665–21674. Curran Associates, Inc., 2020.

Bailin Wang, Dan Friedman, Tao Lei, and Chong Wang. SPLA: Block Sparse Plus Linear Attention for Long Context Modeling. *arXiv preprint arXiv:2601.22379v1*, 2026.

Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC, 2nd edition, 2010. doi: 10.1201/ebk1439808184.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. In *International Conference on Learning Representations*, 2024.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads. In *International Conference on Learning Representations*, 2025. arXiv:2410.10819v1.

Peng Xie and Amr Alanwar. Error Certificates for KV-Cache Eviction via Randomized Design. *arXiv preprint arXiv:2607.21475v3*, 2026.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. XAttention: Block Sparse Attention with Antidiagonal Scoring. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69819–69831. PMLR, 2025. arXiv:2503.16428v1.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388v1*, 2025.

Chiwun Yang and Xiaoyu Li. Beyond Sparse Weights: When Is Attention Compressible? *arXiv preprint arXiv:2608.21541v1*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1259.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly. In *International Conference on Learning Representations*, 2025. arXiv:2410.02694v3.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23097. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1126.

Yuhang Zhan, Lisi Chen, and Shuo Shang. ResKV: Reconstructing Omitted Attention Contributions for Fixed-Budget KV Cache Compression. *arXiv preprint arXiv:2607.29591*, 2026.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, Joseph E. Gonzalez, Jianfei Chen, and Jun Zhu. SLA: Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse–Linear Attention. In *International Conference on Learning Representations*, 2026. arXiv:2509.24006.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The Hedgehog & the Porcupine: Expressive Linear Attentions with Softmax Mimicry. In *International Conference on Learning Representations*, 2024. arXiv:2402.04347v1.

Minwei Zhang, Haifeng Sun, Jingyu Wang, Shaolong Li, Wanyi Ning, Qi Qi, Zirui Zhuang, and Jianxin Liao. ClusterAttn: KV Cache Compression under Intrinsic Attention Clustering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14451–14473. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.703.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang "Atlas" 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, pp. 34661–34710. Curran Associates, Inc., 2023. doi: 10.52202/075280-1506.

Lin Zheng, Jianbo Yuan, Chong Wang, and Lingpeng Kong. Efficient Attention via Control Variates. In *International Conference on Learning Representations*, 2023.

Kan Zhu, Tian Tang, Qinyu Xu, Zhan Jin, Yile Gu, Zhichen Zeng, Rohan Kadekodi, Liangyu Zhao, Ang Li, Arvind Krishnamurthy, and Baris Kasikci. Tactic: Adaptive Sparse Attention with Clustering and Distribution Fitting for Long-Context LLMs. In *International Conference on Learning Representations*, 2026a.

Yuxuan Zhu, David H. Yang, Mohammad Mohammadi Amiri, Keerthiram Murugesan, Tejaswini Pedapati, and Pin-Yu Chen. OjaKV: Context-Aware Online Low-Rank KV Cache Compression. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10161–10178. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.findings-acl.494.

Appendix Guide

The main paper follows one chain: useful local directions, the finite-state realization gap, the support carried by a small key set, resource allocation, and full-model scope. The appendices separate the evidence needed for each step rather than adding new headline claims.

- **Protocol and mathematics:** Appendices A and B give inputs, metrics, uncertainty, coordinate conventions and proofs.
- **Direction-to-realization gap:** Appendix C gives the full dimension and kernel-family sweeps, numerical conventions and the scope of pointwise guarantees.
- **The shadow evidence:** Appendices D to F give basis controls, all-head coverage, error covariation, support persistence and attention-sink analysis.
- **Representation and resources:** Appendices G and H give resource-matched comparisons, the complete cost account, timing limitations and remainder ablations.
- **Full-model checks:** Appendices I to K give all development arms, causal tests, longer-context outcomes and fresh-input confirmation.

A Experimental Apparatus and Reproducibility Protocol

Model checkpoints and hardware environment. The core operator study and development end-to-end evaluation use the following open-weight models; the separate fresh-input study is specified in Appendix K:

- **Llama-3.2-3B**: `unsloth/Llama-3.2-3B` (a mirror of Meta Llama-3.2-3B), revision `d4446454`, 28 transformer layers, 24 query heads, 8 key/value heads, hidden dimension 3072, head dimension 128.
- **Qwen3-8B-Base**: `Qwen/Qwen3-8B-Base`, revision `49e3418f`, 36 transformer layers, 32 query heads, 8 key/value heads, hidden dimension 4096, head dimension 128.

Experiments were conducted on NVIDIA H100 (80GB) and H200 GPUs using PyTorch 2.13.0 (CUDA 12.9 build) and Transformers 4.57.6. Models run in bfloat16; operator-level analyses and the gated attention are computed in float32 from bf16 activations.

Dataset splits and cluster resampling. NIAH, FWE and VT are RULER-style synthetic tasks (Hsieh et al., 2024) built by our own generator, not the RULER code, with seed 20260726 and public-domain book text as filler; HotpotQA and 2WikiMQA are, for each model, the first 40 records of the LongBench (Bai et al., 2024) files whose prompt has at least 4096 tokens under that model's tokenizer. Synthetic sequences are 16384 tokens long; LongBench prompts are 4.2K–15.9K tokens, middle-truncated to at most 15872. The frozen inputs record the LongBench record identifiers. For operator-level evaluations, sequences 0 and 1 form the training split, sequences 2 and 3 the calibration split, and sequences 4 through 7 the test split (sequences 0–3 together form the development split). The 20 test problem instances are identical across both architectures. The end-to-end evaluation uses items 0 to 39 of every task, so it contains the operator-level sequences 0 to 7, test items included. Items 0 to 7 of every task are the same problems for the two models. Because the length filter depends on the tokenizer, the 40 2WikiMQA records of each model include one that those of the other model do not, so 79 of the 80 LongBench records occur for both models. The extra Qwen record is its 2WikiMQA item 17, so from that item on the 2WikiMQA items of the two models are offset by one position, and items 17 to 39 hold different records for the two models.

Unless a table explicitly states otherwise, confidence intervals and hypothesis tests employ a cluster bootstrap procedure that resamples problem clusters with replacement 2000 times within each task using fixed random seed `202609102` at the operator level and `20260726` end to end. Paired differences are computed on identical sequences and identical head windows before bootstrap aggregation. At the operator level a cluster is one item (a task and a sequence index), whose windows, layers and heads are averaged before resampling; end to end a cluster is one input. Items are treated as exchangeable within a task, and the operator-level estimand is the equal-weight mean over the five tasks of the expected per-item value for items drawn as ours were, for a fixed model and, at the operator level, fixed captured heads and window positions. The development end-to-end primary

endpoint E4 averages NIAH, FWE, HotpotQA and 2WikiMQA equally; E5 additionally includes VT and is reported only as a separate descriptive endpoint. Long-context and fresh-input endpoints are defined in their own sections. The intervals do not cover variation across tasks, models or heads, and at the operator level, with four items per task in the development and test splits, they are best read as descriptive.

Operator-level heads and frozen geometry. The captures hold two query heads per layer, fixed by index before any analysis as heads 0 and $H_q/2$: heads 0 and 12 of Llama-3.2-3B and heads 0 and 16 of Qwen3-8B-Base. Under grouped-query attention they read key-value heads 0 and 4, so each captured head comes with the keys and values of its own group and the two heads never share keys. The resulting 128 head-layer pairs are 7% of the 1,824 query heads of the two models (672 and 1,152). All frozen quantities, the matrix R of Appendix B and the global basis of Appendix B.2, are fitted separately for each captured head in each layer on the train split (sequences 0 and 1 of each task, two windows each: 20 windows and 640 queries per head), and no test data enter them. In each window the tilted covariance C is formed explicitly in float32 and is never inverted, and the local basis comes from a dense symmetric eigensolver (`torch.linalg.eigh`) in float64. The random basis is the orthogonal factor of the QR decomposition of a $d \times d$ standard Gaussian matrix, drawn once per head with a seed derived from the model, layer and head identifiers. Five further random draws change the random-basis diagnostic error on the test split by at most .0009.

Blocks, budgets, ties and grouped-query attention. At the operator level every window is a complete block of L queries ($L = 32$ except in the block-length sweep of Appendix F) with at least $1{,}568$ history keys (when the input length is not a multiple of 32, the window at the end of the sequence is the last complete block), so no padding or truncation occurs, and a budget fraction f gives $\max(1, \mathrm{round}(fN))$ exact keys. End to end, the first block, which has no history, and a final block shorter than 32 tokens are computed densely, and a block starting at position s_b keeps $\min\left(s_b, \max(1, \mathrm{round}(fs_b))\right)$ exact keys, so a budget that covers the whole history makes the block exact. The top keys are taken from float32 scores with `torch.topk` at the operator level and `torch.sort` end to end; both break exact ties in an unspecified order. End to end the representation runs on each query head separately after the key and value heads are repeated for grouped-query attention: every query head has its own anchor rows, exact set, tilted mean $\bar{k}$ and remainder state, although the heads of a group read the same keys and values, and the key-pair account counts query heads. The 120 synthetic end-to-end inputs (NIAH, FWE, VT) have exactly $16{,}384$ tokens, a multiple of the block length, so their question occupies the last rows of a gated block (from row 18 for 37 of the 40 NIAH inputs and row 17 for the other three, from row 5 for FWE, and from row 12 on Llama-3.2-3B and row 8 on Qwen3-8B-Base for VT). The LongBench prompts have $4{,}194$ to $15{,}872$ tokens; for the 63 (Llama-3.2-3B) and 58 (Qwen3-8B-Base) of the 80 whose length is not a multiple of 32, the final partial block, which holds the last $N \bmod 32$ tokens of the prompt, is computed with exact attention.

Generation and evaluation parameters. For end-to-end evaluations, only the prefill attention is replaced, in all layers and heads; the compressed prefill still writes the full KV cache, and decoding then runs greedily ($\mathrm{temperature} = 0$) with exact attention over it. Generation stops at an end-of-sequence token, at the first newline, or at task-specific maximum token limits: 12 tokens for NIAH, 24 tokens for FWE, 8 tokens for VT, and 32 tokens for HotpotQA and 2WikiMQA. Task scores are computed by the scoring function in our code, with frozen inputs, per-input result files and aggregation scripts retained for reproducibility. It reads the first line of the generation, lowercased with whitespace collapsed; NIAH and VT count as correct when the first integer equals the answer, FWE when the listed words are exactly the three target words, and HotpotQA and 2WikiMQA score the maximum token-level F1 over the reference answers.

End-to-end error. In every layer, the output of the gated attention, cast to bf16, is compared for each query head and row with bf16 SDPA attention computed from the same queries, keys and values of that layer, as $\|\hat{o} - o\|_2 / \max(\|o\|_2, 10^{-9})$. Both sides use the inputs that the gated model itself produced, so this is the error that each layer adds, not the distance from the activations of the exact model. rel_q averages it over the question rows, then over query heads, then over layers, and then over the 200 inputs with equal weight; the error over all rows is averaged in the same way.

A.1 Sensitivity to the 20 test items

Every operator-level comparison averages the windows, layers and heads of an item first, so each model contributes 20 per-item differences, four per task. With so few items per task the percentile cluster bootstrap described above uses the plug-in variance, which is $3/4$ of the unbiased one. Table 2 therefore lists, for the operator-level paired comparisons of selectors, summaries and remainders, the bootstrap interval together with a task-stratified t interval (15 degrees of freedom), the number of items on the side of the estimate, a sign-flip test of the task-weighted mean over all 2^{20} sign patterns, exact under the null that the paired differences are symmetric about zero and not a distribution-free test of a zero mean, and the range of the estimate when one task is left out; Figure 6 shows the per-item differences. The t intervals are 22 to 33% wider than the bootstrap intervals, close to the factor of about 1.26 that the plug-in variance and the t quantile predict. We call a paired difference a difference only when the t interval excludes zero and the sign-flip test gives $p < .05$, and we give the number of items whenever fewer than 15 of the 20 are on the side of the estimate; 15 of 20 is the smallest count at which a two-sided sign test gives $p < .05$ (.041).

The comparisons with the few-direction summary, between the two selectors at the same budget and with the same treatment of the remainder, of the remainder against dropping at the same exact-set size at all six budgets and under all four error measures, of the two centers, and of the zero-order against the exact-exponential remainder have the sign of the estimate in all 20 items on both models, except the centering on Llama at $m = N/128$ and, with four anchor rows, at $N/16$ (19 of 20 each), and no left-out task changes their sign. Two kinds of comparison are weaker. The anchor gate at $m = N/16$ and $N/8$ against per-query selection at $N/128$, both with the zero-order remainder, has the side of the estimate in 14 to 18 items, so the budget at which the anchor gate matches per-query selection lies between $N/16$ and $N/8$ but is not sharper than that. The equal-cost comparisons are the most fragile: 10 of their 32 model-level cells fail the criterion, six of them at $N/128$ to $N/32$ and the others under the MAC count on Qwen at $N/16$ to $N/4$ and with four anchor rows at $N/16$; under that count dropping is better on Qwen at $N/128$ and $N/64$, and FWE items carry much of the advantage of the remainder (Table 28). The pre-registered comparison of the anchor-gated representation with per-query selection that drops the rest mixes two treatments of the remainder, and Section 5 no longer quotes it: per-query selection with drop is better in 16 and 17 of the 20 items, and every exception is an FWE item. All these intervals and tests describe the 20 items we have; they do not cover variation across tasks, models or heads.

B Mathematical Proofs and Analytical Derivations

We fix a single attention head with queries and keys in $\mathbb{R}^d$, values in $\mathbb{R}^{d_v}$ and logit scale $\beta = 1/\sqrt{d}$ ($d = d_v = 128$ in both models; the derivations in this appendix hold for any d and d_v), and a block of L consecutive queries $q_1, \dots, q_L$. The key history consists of $k_1, \dots, k_N$ with corresponding values $v_1, \dots, v_N$, and the keys of the current block are $k_{N+1}, \dots, k_{N+L}$. Query q_i attends to the history and to the legal block keys $\mathcal{B}_i = \{N+1, \dots, N+i\}$; we write P_i for its full causal softmax distribution over $\{1, \dots, N\} \cup \mathcal{B}_i$, which defines the attention output $o_i = \sum_j P_i(j)\, v_j$. The anchor query is denoted q_a, and its raw logits on history keys are $s_a(j) = \beta\, q_a^\top k_j$. The tilted anchor distribution is $p_a = \mathrm{softmax}(s_a)$ evaluated strictly over history keys $1 \le j \le N$. We define the tilted key mean as $\bar{k} = \sum_{j=1}^{N} p_a(j) k_j$ and the centered key deviation as $\tilde{k}_j = k_j - \bar{k}$, which satisfies $\sum_j p_a(j) \tilde{k}_j = 0$ by construction.

Scope and boundary conventions. Everything in this appendix concerns one row of one head and finite sums over finitely many keys with finite logits. The identities and bounds therefore need no probabilistic assumption on queries, keys or values, and every normalizer they divide by is positive because every $p_a(j) > 0$; the one exception is the denominator of the energy share in Equation (6), which is positive unless all history keys coincide. Whether the quantities in the bounds are small is an empirical question that the main text answers by measurement. We assume $N \ge 1$: a block without history occurs only as the first block of a sequence and is computed with exact attention. If the exact set holds the whole history, the remainder is empty, $S_0 = 0$, no ratio S_1/S_0 is formed, and the representation is exact attention; otherwise $S_0 > 0$. The decomposition below never inverts C, so it holds for every rank of C, including $C = 0$; when leading eigenvalues of a covariance coincide, the basis is whichever orthonormal eigenvectors the symmetric eigensolver returns. Appendix A gives the implementation conventions for block boundaries, budgets, ties and grouped-query attention.

Table 2: **Operator-level paired comparisons on the 20 test items.** Estimate (a − b, L2/Vmax unless the row names another error measure, task-weighted cluster mean), 95% cluster-bootstrap and task-stratified t intervals (15 degrees of freedom), the number of items on the side of the estimate, the two-sided sign-flip p over all 2^{20} sign patterns (exact under symmetry about zero; 2^{-19} is its smallest value) and the range of the estimate when one task is left out. The equal-cost comparisons at the other budgets and the centering at the other budgets are in Tables 26 and 28.

Comparison	Model	estimate	bootstrap 95%	t 95%	items	sign-flip p	leave one task out
Selectors and summaries, $m = N/128$ unless stated							
gate + zero-order − local r=8 P_2	Llama	−.0423	[−.0448, −.0402]	[−.0454, −.0393]	20/20	2^{-19}	[−.0442, −.0403]
	Qwen	−.0315	[−.0333, −.0296]	[−.0338, −.0291]	20/20	2^{-19}	[−.0340, −.0305]
drop: anchor's set − per-query set	Llama	+.0138	[+.0131, +.0145]	[+.0129, +.0147]	20/20	2^{-19}	[+.0130, +.0147]
	Qwen	+.0197	[+.0189, +.0203]	[+.0187, +.0206]	20/20	2^{-19}	[+.0186, +.0209]
zero-order: anchor's set − per-query set	Llama	+.0115	[+.0107, +.0122]	[+.0106, +.0124]	20/20	2^{-19}	[+.0106, +.0124]
	Qwen	+.0181	[+.0173, +.0188]	[+.0171, +.0190]	20/20	2^{-19}	[+.0172, +.0192]
zero-order: anchor's set at $N/16$ − per-query set	Llama	+.0018	[+.0013, +.0023]	[+.0012, +.0024]	14/20	.004	[+.0011, +.0026]
	Qwen	+.0035	[+.0026, +.0045]	[+.0023, +.0047]	18/20	9.7×10^{-5}	[+.0029, +.0041]
zero-order: anchor's set at $N/8$ − per-query set	Llama	−.0024	[−.0027, −.0020]	[−.0029, −.0019]	18/20	2.5×10^{-5}	[−.0030, −.0019]
	Qwen	−.0013	[−.0021, −.0004]	[−.0025, −.0002]	14/20	.025	[−.0017, −.0009]
per-query set: zero-order − drop	Llama	−.0082	[−.0087, −.0077]	[−.0089, −.0076]	20/20	2^{-19}	[−.0093, −.0060]
	Qwen	−.0096	[−.0099, −.0094]	[−.0100, −.0093]	20/20	2^{-19}	[−.0107, −.0066]
gate: zero-order − exact-exp. remainder (diagnostic)	Llama	+.0100	[+.0096, +.0104]	[+.0095, +.0105]	20/20	2^{-19}	[+.0090, +.0110]
	Qwen	+.0122	[+.0117, +.0127]	[+.0116, +.0128]	20/20	2^{-19}	[+.0104, +.0134]
The remainder							
gate: zero-order − drop, same m, $N/128$	Llama	−.0105	[−.0110, −.0100]	[−.0111, −.0099]	20/20	2^{-19}	[−.0115, −.0082]
	Qwen	−.0112	[−.0115, −.0109]	[−.0116, −.0108]	20/20	2^{-19}	[−.0124, −.0079]
same, $N/64$	Llama	−.0085	[−.0089, −.0081]	[−.0091, −.0080]	20/20	2^{-19}	[−.0095, −.0065]
	Qwen	−.0082	[−.0084, −.0079]	[−.0085, −.0079]	20/20	2^{-19}	[−.0091, −.0055]
same, $N/32$	Llama	−.0066	[−.0070, −.0063]	[−.0071, −.0062]	20/20	2^{-19}	[−.0074, −.0050]
	Qwen	−.0057	[−.0059, −.0056]	[−.0059, −.0055]	20/20	2^{-19}	[−.0064, −.0036]
same, $N/16$	Llama	−.0048	[−.0051, −.0045]	[−.0052, −.0044]	20/20	2^{-19}	[−.0053, −.0036]
	Qwen	−.0038	[−.0039, −.0037]	[−.0039, −.0037]	20/20	2^{-19}	[−.0043, −.0023]
same, $N/8$	Llama	−.0030	[−.0032, −.0028]	[−.0033, −.0027]	20/20	2^{-19}	[−.0034, −.0022]
	Qwen	−.0023	[−.0024, −.0023]	[−.0024, −.0023]	20/20	2^{-19}	[−.0026, −.0013]
same, $N/4$	Llama	−.0015	[−.0017, −.0014]	[−.0017, −.0014]	20/20	2^{-19}	[−.0017, −.0011]
	Qwen	−.0012	[−.0013, −.0012]	[−.0013, −.0012]	20/20	2^{-19}	[−.0014, −.0006]
same, $N/128$, relative L2 error	Llama	−.1046	[−.1081, −.1011]	[−.1092, −.1000]	20/20	2^{-19}	[−.1123, −.0954]
	Qwen	−.0878	[−.0900, −.0857]	[−.0905, −.0851]	20/20	2^{-19}	[−.0930, −.0758]
same, $N/128$, value-RMS error	Llama	−.0188	[−.0196, −.0180]	[−.0198, −.0177]	20/20	2^{-19}	[−.0207, −.0147]
	Qwen	−.0244	[−.0250, −.0238]	[−.0251, −.0236]	20/20	2^{-19}	[−.0269, −.0171]
same, $N/128$, absolute error	Llama	−.0656	[−.0687, −.0625]	[−.0696, −.0616]	20/20	2^{-19}	[−.0720, −.0540]
	Qwen	−.2382	[−.2493, −.2286]	[−.2509, −.2255]	20/20	2^{-19}	[−.2636, −.1764]
center $\bar{k}_{\mathcal{O}}$ − center $\bar{k}$, $N/128$	Llama	−.0011	[−.0012, −.0010]	[−.0013, −.0009]	19/20	3.8×10^{-6}	[−.0013, −.0008]
	Qwen	−.0025	[−.0028, −.0023]	[−.0028, −.0023]	20/20	2^{-19}	[−.0027, −.0023]
same, $N/4$	Llama	−.0012	[−.0013, −.0011]	[−.0013, −.0010]	20/20	2^{-19}	[−.0013, −.0009]
	Qwen	−.0009	[−.0009, −.0009]	[−.0009, −.0009]	20/20	2^{-19}	[−.0010, −.0007]
zero-order − drop, equal key-pair cost, $N/128$	Llama	−.0028	[−.0033, −.0023]	[−.0034, −.0021]	14/20	.014	[−.0038, −.0007]
	Qwen	+.0011	[+.0008, +.0015]	[+.0006, +.0017]	15/20	.341	[+.0002, +.0036]
zero-order − drop, equal MAC count, $N/128$	Llama	+.0011	[+.0006, +.0017]	[+.0004, +.0019]	14/20	.234	[+.0002, +.0029]
	Qwen	+.0063	[+.0057, +.0069]	[+.0055, +.0070]	16/20	4.4×10^{-5}	[+.0052, +.0082]
Pre-registered, mixes two remainder treatments, not quoted in Section 5							
gate + zero-order − per-query set + drop	Llama	+.0033	[+.0022, +.0043]	[+.0019, +.0046]	16/20	.023	[+.0017, +.0054]
	Qwen	+.0084	[+.0076, +.0092]	[+.0075, +.0094]	17/20	2.5×10^{-5}	[+.0071, +.0107]

Every identity and bound of Appendices B, C and H, the worked examples in them and the boundary conventions of the gated representation (empty remainder, singular C, $V_{\max} = 0$, ties, a final block shorter than 32 tokens, grouped-query attention) are checked by float64 unit tests on synthetic arrays that call the code used for the paper's numbers; the tests are retained with the experimental code.

B.1 The anchor decomposition identity

We expand the raw attention logit $\beta q_i^\top k_j$ around the anchor reference q_a and the tilted mean $\bar{k}$:

$$\beta\, q_i^\top k_j \;=\; \beta\, q_a^\top k_j \;+\; \beta\,(q_i - q_a)^\top \bar{k} \;+\; \beta\,(q_i - q_a)^\top \tilde{k}_j. \tag{3}$$

Let $C = \sum_j p_a(j)\tilde{k}_j\tilde{k}_j^\top$ be the tilted key covariance matrix, and let R be any symmetric positive definite matrix. Defining the query displacement $u_i = \sqrt{\beta}\, R^{-1}(q_i - q_a)$ and the key deviation $w_j = \sqrt{\beta}\, R\,\tilde{k}_j$, their inner product satisfies $u_i^\top w_j = \beta(q_i - q_a)^\top \tilde{k}_j$ for every such R. In all experiments R is fixed per model, layer and query head and is not derived from C: with $\Sigma_\Delta = V\Lambda V^\top$ the second-moment matrix of the query displacements $q_i - q_a$ over the 640 queries of the 20 training windows (sequences 0 and 1 of every task, two windows each), $R = V \max(\Lambda, \lambda_0 I)^{1/2} V^\top$ with the eigenvalue floor $\lambda_0 = \max\left(10^{-6}\,\mathrm{tr}(\Sigma_\Delta)/d,\ 10^{-12}\right)$. The floor is the only regularization and makes R invertible, so R^{-1} is an exact inverse and the identity below holds exactly. With this choice the query displacements are whitened: the training second moment of u_i is βI wherever the floor is inactive. The key deviations are not: the tilted covariance of w_j is βRCR. We therefore call u_i the

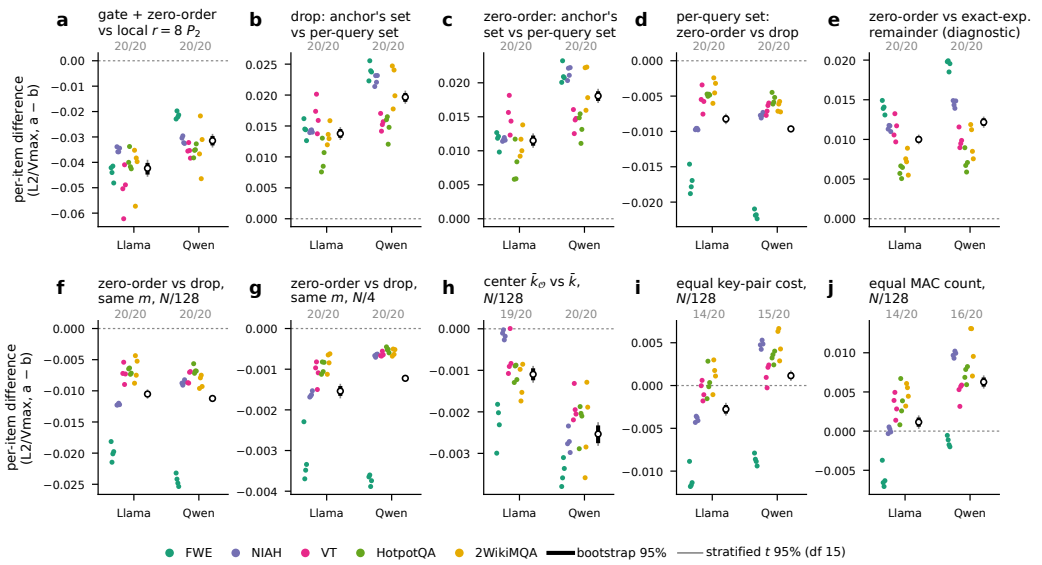

Figure 6: **Per-item differences for paired comparisons of Table 2** (test split, L2/Vmax, a − b; negative favors a). Each dot is one of the 20 items, colored by task; the black bar is the 95% cluster-bootstrap interval around the estimate (open circle), the grey line the task-stratified t interval, and the number above each model the items on the side of the estimate. (a)–(e) selectors and summaries at $m = N/128$; (f)–(g) the remainder against dropping at the same exact-set size; (h) the two centers; (i)–(j) the remainder against dropping at equal cost, $m = N/128$.

whitened query displacement and w_j simply the key deviation. Consequently, we obtain the exact algebraic identity:

$$\beta\, q_i^\top k_j \;=\; s_a(j) \;+\; c_i \;+\; u_i^\top w_j, \qquad c_i := \beta\,(q_i - q_a)^\top \bar{k}, \tag{4}$$

with no remainder or truncated terms. In our float32 code the three terms reproduce the logit up to rounding: over all development windows the largest absolute deviation is 2.0×10^{-4} on Llama and 4.8×10^{-5} on Qwen. This is standard exponential-tilting (change-of-measure) algebra; we use it only to split each logit into a key-only term, a per-query constant and a bilinear interaction.

Dividing the unnormalized weight $\exp(\beta q_i^\top k_j)$ by the history normalizer reveals that every query q_i in the block views an exponential tilt of the anchor distribution:

$$p_i(j) \;=\; \frac{p_a(j)\, e^{u_i^\top w_j}}{M_i}, \qquad M_i := \sum_{j=1}^{N} p_a(j)\, e^{u_i^\top w_j}. \tag{5}$$

Here p_i is the *history-conditional* distribution of query q_i: $p_i(j) = P_i(j)/h_i$ for $j \le N$, where $h_i = \sum_{j=1}^{N} P_i(j)$ is the mass that the full causal softmax places on the history. Since $\mathcal{B}_i$ contains at least the query's own key, $h_i < 1$. Because $\sum_j p_a(j) w_j = 0$, Jensen's inequality guarantees that $\log M_i \ge u_i^\top \sum_j p_a(j) w_j = 0$. The scalar shift c_i is independent of the key index j; it cancels within the softmax over history tokens, but it does not cancel against the causal keys of the current block. Therefore, any compact summary of history must be scaled by the query-dependent factor e^{c_i} to compete on an equal footing with the exact in-block causal keys.

B.2 Bases, restricted covariance and energy share

All bases act in the w coordinates, and a basis U with r orthonormal columns projects the interaction to $t_{ij} = (U^\top u_i)^\top (U^\top w_j) = u_i^\top U U^\top w_j$, which in general differs from the full interaction $u_i^\top w_j$ of Equation (4) unless $r = d$. The exact-exponential diagnostic gives history key j the logit $s_a(j) + c_i + t_{ij}$ and the P_2 summary the logit $s_a(j) + c_i + \log P_2(t_{ij})$, where dense attention has $s_a(j) + c_i + u_i^\top w_j$; all three keep the legal block keys exact and normalize over the history and those keys together. The local basis U_r of a window holds the leading r eigenvectors of

$J = \sum_{j=1}^{N} p_a(j)\, w_j w_j^\top = \beta RCR$, the tilted covariance of the key deviations. Our code diagonalizes $\beta J = \beta^2 RCR$; the eigenvectors and energy shares are unchanged, while the stored eigenvalues and traces are β times those of J. Since the u_i are whitened, these directions maximize the tilted variance of the projected interaction for isotropic query displacements. The global basis holds the leading eigenvectors of $\beta R\bar{C}R$, where $\bar{C}$ averages with equal weight the tilted key covariances of the 20 training windows, each weighted by its own anchor distribution and centered at its own tilted mean. The random basis is the orthogonal factor of a Gaussian matrix seeded by the head identity. There is one global basis and one random basis per model, layer and query head.

The restricted basis of Section 4 keeps the center $\bar{k}$ and the weights p_a of the full history and drops the keys outside $\mathcal{H}$: it holds the leading eigenvectors of $J_{\mathcal{H}} = \sum_{j\in\mathcal{H}} p_a(j)\, w_j w_j^\top$, with w_j centered at the full-history mean $\bar{k}$. The weights are not renormalized on $\mathcal{H}$, which would rescale $J_{\mathcal{H}}$ without changing its eigenvectors, and the center is not recomputed on $\mathcal{H}$. Recentering would change the matrix and not only its scale: with $\alpha = \sum_{j\in\mathcal{H}} p_a(j)$, $\bar{k}_{\mathcal{H}} = \alpha^{-1} \sum_{j\in\mathcal{H}} p_a(j)\, k_j$ and $C_{\mathcal{H}}$ the covariance of the keys in $\mathcal{H}$ under p_a renormalized on $\mathcal{H}$, we have $J_{\mathcal{H}} = \alpha\beta R\,[C_{\mathcal{H}} + (\bar{k}_{\mathcal{H}} - \bar{k})(\bar{k}_{\mathcal{H}} - \bar{k})^\top]\, R$, and centering at $\bar{k}_{\mathcal{H}}$ drops the rank-one term, which can change the eigenvectors. The default restricted-basis results use the full-history center; the self-centered control is identified explicitly. The directional energy share is

$$E_{\mathcal{H}} \;=\; \frac{\operatorname{tr}(U_8^\top J_{\mathcal{H}}\, U_8)}{\operatorname{tr}(U_8^\top J\, U_8)} \;=\; \frac{\sum_{j\in\mathcal{H}} p_a(j)\, \|U_8^\top w_j\|_2^2}{\sum_{j=1}^{N} p_a(j)\, \|U_8^\top w_j\|_2^2}, \tag{6}$$

where U_8 is the full-history local basis at $r = 8$ and $\mathcal{H}$ holds the top $1/128$, $1/64$ or $1/16$ of the history by p_a; Figure 11(a) plots its mean by mass decile over development windows at $m = N/128$, and Table 3 gives its median over test windows at the same budget.

High anchor mass does not by itself decide the leading directions. Take three keys $(x, 0)$, $(0, y)$ and $(0, -y)$ with anchor masses $1 - \varepsilon$, $\varepsilon/2$ and $\varepsilon/2$, which a softmax realizes with an anchor along the first axis. Their tilted key covariance is $\operatorname{diag}(\varepsilon(1-\varepsilon)x^2,\ \varepsilon y^2)$, so with $R = I$ and $y^2 > (1-\varepsilon)x^2$ the leading direction is the second axis, set by two keys that together carry mass ε; with $\varepsilon = 0.01$, $x = 1$ and $y = 20$ the covariance is $\operatorname{diag}(0.0099, 4)$ although the first key holds 99% of the mass. Whether a few high-mass keys account for the leading directions therefore depends on how far the low-mass keys lie from the tilted mean, which is why we measure the energy share directly instead of inferring it from the mass share. In this example $E_{\mathcal{H}} = 0$ at $r = 1$ for $\mathcal{H}$ the first key, while $1 - T = 0.99$, so the mass share gives no lower bound on the energy share.

B.3 Variance captured by the restricted basis

Write $J = J_{\mathcal{H}} + J_{\mathcal{O}}$, where $J_{\mathcal{O}} = \sum_{j\in\mathcal{O}} p_a(j)\, w_j w_j^\top$ runs over the keys outside $\mathcal{H}$ with the same center $\bar{k}$ and weights p_a as $J_{\mathcal{H}}$, so that $J_{\mathcal{O}} \succeq 0$ and $T = \sum_{j\in\mathcal{O}} p_a(j)$ is the anchor tail mass. For a symmetric matrix A let $\lambda_1(A) \geq \lambda_2(A) \geq \dots$ be its eigenvalues, and write $\lambda_k = \lambda_k(J)$. For $r \leq d$ let $B = U_r$ and $B_{\mathcal{H}}$ hold the leading r eigenvectors of J and of $J_{\mathcal{H}}$, and let $E_{\mathcal{H}} = \operatorname{tr}(B^\top J_{\mathcal{H}} B)/\operatorname{tr}(B^\top J B)$, which is Equation (6) at $r = 8$. The captured share $\operatorname{tr}(B_{\mathcal{H}}^\top J B_{\mathcal{H}})/\sum_{k\leq r}\lambda_k$ compares the tilted variance along the restricted basis with the largest tilted variance that any r orthonormal directions can hold.

Proposition 3 (Variance captured by the restricted basis)**.** *For every $r \leq d$ with $\sum_{k\leq r}\lambda_k > 0$,*

$$\begin{aligned} 1 \;\geq\; \frac{\operatorname{tr}(B_{\mathcal{H}}^\top J B_{\mathcal{H}})}{\sum_{k\leq r}\lambda_k} \;&\geq\; \frac{\sum_{k\leq r}\lambda_k(J_{\mathcal{H}})}{\sum_{k\leq r}\lambda_k} \;\geq\; E_{\mathcal{H}} \\ &\geq\; 1 - \frac{\sum_{k\leq r}\lambda_k(J_{\mathcal{O}})}{\sum_{k\leq r}\lambda_k} \;\geq\; 1 - \frac{\min\{r\|J_{\mathcal{O}}\|_2,\ \operatorname{tr} J_{\mathcal{O}}\}}{\sum_{k\leq r}\lambda_k}. \end{aligned}$$

Proof. By the Ky Fan maximum principle (Fan, 1949), the maximum of $\operatorname{tr}(V^\top A V)$ over $d \times r$ matrices V with orthonormal columns is $\sum_{k\leq r}\lambda_k(A)$, attained by leading eigenvectors of A. Applied to J, it gives the first inequality and $\operatorname{tr}(B^\top J B) = \sum_{k\leq r}\lambda_k$. Since $J_{\mathcal{O}} \succeq 0$, $\operatorname{tr}(B_{\mathcal{H}}^\top J B_{\mathcal{H}}) \geq \operatorname{tr}(B_{\mathcal{H}}^\top J_{\mathcal{H}} B_{\mathcal{H}}) = \sum_{k\leq r}\lambda_k(J_{\mathcal{H}})$, the second. Applied to $J_{\mathcal{H}}$, the principle gives $\sum_{k\leq r}\lambda_k(J_{\mathcal{H}}) \geq \operatorname{tr}(B^\top J_{\mathcal{H}} B)$, the third. Since $\operatorname{tr}(B^\top J_{\mathcal{H}} B) = \sum_{k\leq r}\lambda_k - \operatorname{tr}(B^\top J_{\mathcal{O}} B)$, the principle applied to $J_{\mathcal{O}}$

Table 3: **Proposition 3 on the test split** at $m = N/128$ and $r = 8$: median [10th, 90th percentile] over 2,240 (Llama) and 2,880 (Qwen) windows. The first five rows are the chain of Proposition 3, in order; it holds in every window.

Quantity	Llama-3.2-3B	Qwen3-8B-Base
Captured share $\mathrm{tr}(B_{\mathcal{H}}^{\top} J B_{\mathcal{H}}) / \sum_{k\le 8} \lambda_k$	.978 [.936, .997]	.985 [.938, .998]
$\sum_{k\le 8} \lambda_k(J_{\mathcal{H}}) / \sum_{k\le 8} \lambda_k$	.700 [.386, .952]	.739 [.417, .954]
Energy share $E_{\mathcal{H}}$	.688 [.372, .949]	.729 [.404, .952]
$1 - \sum_{k\le 8} \lambda_k(J_{\mathcal{O}}) / \sum_{k\le 8} \lambda_k$	.676 [.361, .944]	.713 [.393, .944]
$1 - \min\{8\|J_{\mathcal{O}}\|_2, \mathrm{tr}\, J_{\mathcal{O}}\} / \sum_{k\le 8} \lambda_k$	.583 [.160, .924]	.643 [.187, .930]
windows where it is positive	97.9%	97.7%
$\|J_{\mathcal{O}}\|_2 / \gamma_8$	47.0 [15.8, 185.3]	44.9 [12.7, 200.0]
windows where the sin-theta bound is below $\sqrt{8}$	0.04%	0.17%
Largest principal angle (degrees)	61.6 [31.1, 84.9]	54.5 [21.0, 83.3]
$\|\sin\Theta\|_F$	1.14 [.69, 1.46]	1.02 [.45, 1.40]

gives the fourth. The last holds because each $\lambda_k(J_{\mathcal{O}}) \le \|J_{\mathcal{O}}\|_2$ and, as $J_{\mathcal{O}} \succeq 0$, $\sum_{k\le r} \lambda_k(J_{\mathcal{O}}) \le \mathrm{tr}\, J_{\mathcal{O}}$. □

By Weyl's inequality the two spectra also stay within the spectral norm of the omitted part: $\lambda_k(J_{\mathcal{H}}) \le \lambda_k \le \lambda_k(J_{\mathcal{H}}) + \|J_{\mathcal{O}}\|_2$ for every k. The proposition needs no eigenvalue gap, and it bounds only the variance that the restricted basis captures; it says nothing about the diagnostic error, whose relation to the captured variance in Section 4 is empirical. It also covers the self-centered variant of Appendix B.2 for nonempty $\mathcal{H}$: with $\alpha = \sum_{j\in\mathcal{H}} p_a(j) > 0$, $\Delta = \sqrt{\beta}\, R\,(\bar{k}_{\mathcal{H}} - \bar{k})$ and $J_{\mathcal{H}}^{\mathrm{self}}$ the restricted matrix centered at $\bar{k}_{\mathcal{H}}$, $J = J_{\mathcal{H}}^{\mathrm{self}} + (\alpha\Delta\Delta^{\top} + J_{\mathcal{O}})$ with a positive semidefinite second term, so the chain holds with $J_{\mathcal{H}}^{\mathrm{self}}$ in place of $J_{\mathcal{H}}$, also in $B_{\mathcal{H}}$ and $E_{\mathcal{H}}$, and $\alpha\Delta\Delta^{\top} + J_{\mathcal{O}}$ in place of $J_{\mathcal{O}}$.

Table 3 evaluates the chain at $r = 8$ and $m = N/128$ on the test split. Every inequality of the chain holds in all 30,720 window-budget pairs of both splits and the three budgets (float64, relative tolerance 10^{-9}), as do the self-centered chain down to its energy share and Weyl's bounds on the 17 leading eigenvalues that we store. With a median captured share of .978 and .985, an energy share of .688 and .729, and a last bound of .583 and .643 that is positive in 97.9% and 97.7% of windows, the chain is informative but not tight. Most of the slack lies in the second inequality (median .273 and .241), which leaves out the variance $\mathrm{tr}(B_{\mathcal{H}}^{\top} J_{\mathcal{O}} B_{\mathcal{H}})$ that the keys outside $\mathcal{H}$ add along the restricted directions; the energy share in Equation (6) is therefore a conservative floor on the captured share.

The principal angles need a gap. A bound on the angles between B and $B_{\mathcal{H}}$, rather than on captured variance, needs an eigenvalue gap; let $r < d$ (for $r = d$ both bases span $\mathbb{R}^d$). The variant of the Davis–Kahan $\sin\Theta$ theorem by Yu, Wang and Samworth (Davis & Kahan, 1970; Yu et al., 2015) gives $\|\sin\Theta\|_F \le 2\min\{\sqrt{r}\,\|J_{\mathcal{O}}\|_2,\ \|J_{\mathcal{O}}\|_F\}/\gamma_r$ with $\gamma_r = \lambda_r - \lambda_{r+1} > 0$. Since $\|\sin\Theta\|_F \le \sqrt{r}$ always, the bound says something only if $\gamma_r > 2\min\{\|J_{\mathcal{O}}\|_2,\ \|J_{\mathcal{O}}\|_F/\sqrt{r}\}$. At $r = 8$ and $m = N/128$ this holds in 0.04% and 0.17% of test windows: the gap γ_8 is a median 15% of λ_8 in both models, and $\|J_{\mathcal{O}}\|_2/\gamma_8$ has median 47 and 45. The measured angles are correspondingly large, with a median largest principal angle of 61.6° and 54.5° and $\|\sin\Theta\|_F$ of 1.14 and 1.02 (Appendix D.2 reports the largest angle for each r and budget). We therefore state the result for captured variance, which needs no gap, and report the angles only as measurements.

Remark 1 (Tail energy beyond rank m)**.** *Let $\mathcal{O} \neq \varnothing$ and let $P_{\mathcal{H}}$ be the orthogonal projector onto the span of $\{w_j : j \in \mathcal{H}\}$, of rank at most m. Then*

$$\sum_{\ell>m} \lambda_\ell \;\le\; \mathrm{tr}\big((I - P_{\mathcal{H}})J\big) \;=\; \sum_{j\in\mathcal{O}} p_a(j)\,\|(I - P_{\mathcal{H}})w_j\|_2^2 \;\le\; \mathrm{tr}\, J_{\mathcal{O}} \;\le\; T B_w^2,$$

where $B_w = \max_{j\in\mathcal{O}}\|w_j\|_2 \le \sqrt{\beta}\,\|R\|_2 \max_{j\in\mathcal{O}}\|k_j - \bar{k}\|_2$.

The first inequality is the Ky Fan principle for the projector $P_{\mathcal{H}}$, whose rank is at most m, together with $J \succeq 0$: $\mathrm{tr}(P_{\mathcal{H}} J) \le \sum_{\ell\le\min(m,d)} \lambda_\ell$. The equality holds because $I - P_{\mathcal{H}}$ is an orthogonal projector and $(I - P_{\mathcal{H}})\, w_j = 0$ for $j \in \mathcal{H}$, and the remaining steps use $\|(I - P_{\mathcal{H}})\, w_j\|_2 \le \|w_j\|_2$ and $w_j = \sqrt{\beta}\, R\,(k_j - \bar{k})$. For $m \ge d$ the left side is zero, so the remark says nothing about the tail. It controls the spectrum beyond rank m, not the leading $r = 8$ directions, for which

Proposition 3 and the measured captured share are the relevant statements; that our code diagonalizes βJ (Appendix B.2) leaves all ratios here unchanged. On the windows with $m < d$ at $m = N/128$ (1,568 of 2,240 and 2,016 of 2,880 test windows), $\operatorname{tr} J_{\mathcal{O}}$ is a median 37% and 33% of $\operatorname{tr} J$, but the tilted variance outside the span of the high-mass deviations is only a median 4.0% (Llama) and 3.1% (Qwen) of $\operatorname{tr} J$, against 1.8% and 1.4% for the optimal rank-m tail, whereas TB_w^2 exceeds $\operatorname{tr} J$ in the median window (median ratio 1.38 in both models; below $\operatorname{tr} J$ in only 36% and 37% of these windows): a few low-mass keys far from the tilted mean, as in the example of Appendix B.2, make the last step loose, so the remark shows which quantities control the tail rather than certifying it.

B.4 Proof of the two-term error decomposition

Let $\mathcal{H} \subset \{1, \dots, N\}$ denote the exact set of history keys, $\mathcal{O} = \{1, \dots, N\} \setminus \mathcal{H}$ the unselected remainder, and $\mathcal{E}_i = \mathcal{H} \cup \mathcal{B}_i$ the set of keys that query q_i evaluates key by key, which includes the legal keys of the current block. We assume $\mathcal{O} \neq \varnothing$, so that $S_0 > 0$ and $\hat{\mu}$ is defined; when $\mathcal{O} = \varnothing$ the two outputs coincide. All masses and means below are taken under the full causal softmax P_i, not under the history-conditional p_i. Let $\pi_i = \sum_{j \in \mathcal{O}} P_i(j)$ be the true mass on $\mathcal{O}$ and $\hat{\pi}_i = e^{c_i} S_0 / \left(e^{c_i} S_0 + \sum_{j \in \mathcal{E}_i} e^{\beta q_i^\top k_j}\right)$ the mass that the representation assigns to it. Let $\mu_i = \sum_{j \in \mathcal{O}} \frac{P_i(j)}{\pi_i} v_j$ be the true conditional mean of the values on the remainder, $\hat{\mu} = S_1 / S_0$ the zero-order anchor mean, and $\nu_i = \sum_{j \in \mathcal{E}_i} \frac{P_i(j)}{1 - \pi_i} v_j$ the conditional mean of the values on $\mathcal{E}_i$.

Because the anchor-gated representation evaluates all keys in $\mathcal{E}_i$ key by key, the relative softmax weights within $\mathcal{E}_i$ are identical between the dense reference and the approximation. Both outputs can therefore be expressed as convex combinations of the identical exact-part mean ν_i with their respective remainder means:

$$o_i \;=\; (1 - \pi_i)\, \nu_i \;+\; \pi_i\, \mu_i, \tag{7}$$

$$\hat{o}_i \;=\; (1 - \hat{\pi}_i)\, \nu_i \;+\; \hat{\pi}_i\, \hat{\mu}. \tag{8}$$

Proof of Proposition 2. Subtracting the two expressions:

$$\hat{o}_i - o_i \;=\; (\pi_i - \hat{\pi}_i)\, \nu_i \;+\; \hat{\pi}_i\, \hat{\mu} \;-\; \pi_i\, \mu_i. \tag{9}$$

Adding and subtracting $\pi_i \hat{\mu}$:

$$\hat{\pi}_i\, \hat{\mu} \;-\; \pi_i\, \mu_i \;=\; \pi_i\, (\hat{\mu} - \mu_i) \;+\; (\hat{\pi}_i - \pi_i)\, \hat{\mu}. \tag{10}$$

Substituting this back and grouping the terms involving $(\hat{\pi}_i - \pi_i)$:

$$\hat{o}_i - o_i \;=\; \pi_i\, (\hat{\mu} - \mu_i) \;+\; (\hat{\pi}_i - \pi_i)\, (\hat{\mu} - \nu_i). \tag{11}$$

This completes the proof. No approximations or dropped terms are introduced. □

The first term, $\pi_i(\hat{\mu} - \mu_i)$, represents the value mean error weighted by the true tail mass. The second term, $(\hat{\pi}_i - \pi_i)(\hat{\mu} - \nu_i)$, represents the tail mass estimation error weighted by the separation between the zero-order remainder mean $\hat{\mu}$ and the exact-set mean ν_i. The two terms are distinct sources of error but not orthogonal: writing them a_i and b_i, $\|\hat{o}_i - o_i\|_2^2 = \|a_i\|_2^2 + \|b_i\|_2^2 + 2\langle a_i, b_i \rangle$, and the cross term can take either sign. Correcting one statistic therefore need not reduce the error of a row. With scalar values $\nu_i = 0$, $\mu_i = 1/2$, $\hat{\mu} = 1$ and masses $\pi_i = 1/2$, $\hat{\pi}_i = 1/4$, the two terms are $+1/4$ and $-1/4$ and $\hat{o}_i = o_i$, whereas supplying the true mass leaves the error $1/4$.

To bound the value mean error, define the anchor base measure on $\mathcal{O}$ as $\hat{q}(j) = p_a(j)/T$, where $T = \sum_{j \in \mathcal{O}} p_a(j)$ is positive because $\mathcal{O} \neq \varnothing$ and every $p_a(j) > 0$. The likelihood ratio between the true conditional distribution and $\hat{q}$ is $L_i(j) = e^{u_i^\top w_j - \delta_i}$, where $\delta_i = \log \sum_{j \in \mathcal{O}} \hat{q}(j) e^{u_i^\top w_j}$. Then:

$$\mu_i - \hat{\mu} \;=\; \mathbb{E}_{\hat{q}}\left[(L_i(j) - 1)\, (v_j - \hat{\mu})\right]. \tag{12}$$

By the Cauchy–Schwarz inequality:

$$\|\mu_i - \hat{\mu}\|_2 \;\leq\; \sqrt{\operatorname{Var}_{\hat{q}}(L_i)}\, \sqrt{\mathbb{E}_{\hat{q}}\left[\|v_j - \hat{\mu}\|_2^2\right]}. \tag{13}$$

The value mean error is thus bounded by the root-mean-square value dispersion of the unselected keys times the standard deviation of the likelihood ratio. Since $\mathbb{E}_{\hat{q}}[L_i] = 1$, the variance in this

bound is the chi-square divergence $\mathrm{Var}_{\hat{q}}(L_i) = \chi^2(P_{i,\mathcal{O}} \,\|\, \hat{q})$ of the true conditional distribution on the remainder, $P_{i,\mathcal{O}}(j) = P_i(j)/\pi_i = \hat{q}(j)\, L_i(j)$, from the anchor base measure. With the log-moment generating function $K_{\mathcal{O}}(u) = \log \mathbb{E}_{\hat{q}}\, e^{u^\top w_j}$ of the remainder, for which $\delta_i = K_{\mathcal{O}}(u_i)$, the divergence is $\chi^2(P_{i,\mathcal{O}} \,\|\, \hat{q}) = e^{K_{\mathcal{O}}(2u_i) - 2K_{\mathcal{O}}(u_i)} - 1$: it grows with the spread of the interaction over the remainder and does not shrink merely because the remainder holds many keys. The constant in Equation (13) cannot be improved without assumptions on the values, since equality holds when $v_j - \hat{\mu} = a\,(L_i(j) - 1)\,\xi$ for a scalar a and a fixed vector ξ. The bound is sufficient, not an estimate: it is loose when $L_i - 1$ and $v_j - \hat{\mu}$ are poorly aligned, and we do not claim that it is tight on our data (Appendix B.6 measures its slack). Appendix B.6 combines it with the mass error of Equation (16) into bounds on the output error.

B.5 The tail amplification identity

Let $T = \sum_{j \notin \mathcal{H}} p_a(j)$ denote the anchor tail mass on the unselected set $\mathcal{O} = \mathcal{H}^c$, which is obtained directly from a single anchor QK row. Defining the conditional log-normalizer on the remainder as $\delta_i = \log \sum_{j \notin \mathcal{H}} p_a(j) e^{u_i^\top w_j} - \log T$ and the net log-amplification factor as $A_i = \delta_i - \log M_i$, the share of the history-conditional distribution p_i that falls on the remainder is exactly

$$\pi_i^{\mathrm{hist}} \;=\; \sum_{j \notin \mathcal{H}} p_i(j) \;=\; \frac{\sum_{j \notin \mathcal{H}} p_a(j) e^{u_i^\top w_j}}{M_i} \;=\; \frac{T\, e^{\delta_i}}{M_i} \;=\; T\, e^{A_i}. \tag{14}$$

This is not the remainder mass π_i of Proposition 2, which is taken under the full causal softmax P_i and therefore competes with the legal block keys. With $Z_a = \sum_{j=1}^{N} e^{s_a(j)}$ and $Z_{\mathcal{B},i} = \sum_{j \in \mathcal{B}_i} e^{\beta q_i^\top k_j}$, the history carries the unnormalized weight $\sum_{j=1}^{N} e^{\beta q_i^\top k_j} = e^{c_i} Z_a M_i$, so

$$\pi_i \;=\; h_i\, \pi_i^{\mathrm{hist}} \;=\; h_i\, T\, e^{A_i}, \qquad h_i \;=\; \frac{e^{c_i} Z_a M_i}{e^{c_i} Z_a M_i + Z_{\mathcal{B},i}}. \tag{15}$$

The two masses would agree only if no block key competed with the history, which never happens because $\mathcal{B}_i$ contains the query's own key. The gap is material: on the development split at $m = N/128$, the median over windows of the mean history-conditional share is .132 (Llama) and .288 (Qwen), against .097 and .206 for the full mass π_i. The correlations with $\log T + A_i = \log \pi_i^{\mathrm{hist}}$ reported in Appendix H use the history-conditional share (it also gives them with the full mass); the oracle-mass counterfactuals use the full mass π_i. The eight features of the auxiliary budget predictor of the earlier single-anchor run in Appendix I include T and the linear estimate of A_i (the mean omitted interaction $\bar{x}_i$ of Appendix B.6), but not the true A_i, π_i or π_i^{hist}.

The anchor-gated representation replaces the true unnormalized remainder weight $\sum_{j \notin \mathcal{H}} e^{\beta q_i^\top k_j} = e^{c_i} S_0\, e^{\delta_i}$ by $e^{c_i} S_0$, that is, it sets δ_i to zero, while every weight on $\mathcal{E}_i$ stays exact. The odds of the remainder under P_i are therefore multiplied by exactly $e^{-\delta_i}$, $\mathrm{logit}(\hat{\pi}_i) - \mathrm{logit}(\pi_i) = -\delta_i$, which yields the exact mass error

$$\hat{\pi}_i - \pi_i \;=\; \pi_i\,(1 - \hat{\pi}_i)\,\left(e^{-\delta_i} - 1\right). \tag{16}$$

The log-odds distortion depends only on δ_i and not on the value vectors. The absolute mass error also depends on π_i, and hence on how much mass $\mathcal{H}$ and the block keys take: two rows with the same T and δ_i can have different mass errors. Since $\mathrm{logit}(\hat{\pi}_i) = \mathrm{logit}(\pi_i) - \delta_i$, the error is largest when $\mathrm{logit}(\pi_i) = \delta_i/2$, where the two logistic values differ by $\tanh(\delta_i/4)$; hence for every row, whatever mass $\mathcal{H}$ and the block keys take,

$$|\hat{\pi}_i - \pi_i| \;\le\; \tanh\!\big(|\delta_i|/4\big) \;\le\; |\delta_i|/4. \tag{17}$$

B.6 Norm bounds for the zero-order remainder

Two consequences of Proposition 2 make the role of each remainder statistic explicit. They use the conditional distribution $P_{i,\mathcal{O}} = \hat{q}\, L_i$ of query i on the remainder, which is the same whether attention is normalized over the history or over the full causal prefix, and the anchor-tilted value dispersion $\sigma_{v,\mathcal{O}} = (\mathbb{E}_{\hat{q}} \|v_j - \hat{\mu}\|_2^2)^{1/2}$ of the remainder, so that Equation (13) reads $\|\mu_i - \hat{\mu}\|_2 \le \sqrt{\chi^2(P_{i,\mathcal{O}} \,\|\, \hat{q})}\; \sigma_{v,\mathcal{O}}$.

Proposition 4 (Norm bound)**.** *For every query i in the block,*

$$\|\hat{o}_i - o_i\|_2 \;\le\; \pi_i \sqrt{\chi^2(P_{i,\mathcal{O}} \,\|\, \hat{q})}\; \sigma_{v,\mathcal{O}} \;+\; |\hat{\pi}_i - \pi_i|\, \|\hat{\mu} - \nu_i\|_2 .$$

Proof. Apply the triangle inequality to Proposition 2 and bound $\|\mu_i - \hat{\mu}\|_2$ by Equation (13). □

The first term measures how far the shape of the tail under query i departs from its shape under the anchor, weighted by how much the tail values vary; the second measures the error in the weight of the tail. Like Equation (13), the bound names the controlling quantities rather than giving a numerical estimate. In the units of our error metric it takes a closed form. The means $\hat{\mu}$ and ν_i are convex combinations of values on the legal prefix of row i, so $\|\hat{\mu} - \nu_i\|_2 \le 2V_{\max}$, with $V_{\max}$ the largest value norm on that prefix (Section 2). Together with Equation (17), whenever $V_{\max} > 0$,

$$\frac{\|\hat{o}_i - o_i\|_2}{V_{\max}} \;\le\; \pi_i \sqrt{\chi^2(P_{i,\mathcal{O}} \,\|\, \hat{q})}\; \frac{\sigma_{v,\mathcal{O}}}{V_{\max}} \;+\; 2\tanh\!\Big(\frac{|\delta_i|}{4}\Big). \tag{18}$$

It uses no distributional assumption: the first term is small when query i tilts the remainder little away from the anchor or the remainder values vary little, and the second when the log-odds distortion δ_i is small.

To see when both terms are small, write the interaction that the zero-order remainder omits as $x_{ij} = u_i^\top w_j = \beta\,(q_i - q_a)^\top (k_j - \bar{k})$ for $j \in \mathcal{O}$ (Equation (4)). Let $\bar{k}_{\mathcal{O}} = \sum_{j\in\mathcal{O}} \hat{q}(j)\, k_j$ be the anchor-tilted key mean of the remainder, $\bar{x}_i = \mathbb{E}_{\hat{q}}\, x_{ij} = \beta\,(q_i - q_a)^\top (\bar{k}_{\mathcal{O}} - \bar{k})$ the mean omitted interaction, and $\varepsilon_i = \max_{j\in\mathcal{O}} |x_{ij} - \bar{x}_i|$ its largest deviation from that mean.

Proposition 5 (Bounded residual interaction)**.** *For every query i in the block:*

(a) $e^{-2\varepsilon_i} \le L_i(j) \le e^{\varepsilon_i}$ *for every* $j \in \mathcal{O}$. *Hence* $\mathrm{TV}(P_{i,\mathcal{O}}, \hat{q}) \le \frac{1}{2}(e^{2\varepsilon_i} - 1)$ *and* $\|\mu_i - \hat{\mu}\|_2 \le (e^{2\varepsilon_i} - 1)\, \sigma_{v,\mathcal{O}}$.

(b) $\delta_i = \bar{x}_i + \rho_i$ *with* $0 \le \rho_i \le \varepsilon_i^2/2$, *and* $|\hat{\pi}_i - \pi_i| \le \pi_i\,(1 - \hat{\pi}_i)\,(e^{|\bar{x}_i| + \varepsilon_i^2/2} - 1)$.

(c) $\|\hat{o}_i - o_i\|_2 \le \pi_i \big[(e^{2\varepsilon_i} - 1)\, \sigma_{v,\mathcal{O}} + (e^{|\bar{x}_i| + \varepsilon_i^2/2} - 1)\, \|\hat{\mu} - \nu_i\|_2\big]$.

Proof. Let $y_j = x_{ij} - \bar{x}_i$, so that $\mathbb{E}_{\hat{q}}\, y_j = 0$ and $|y_j| \le \varepsilon_i$, and let $\rho_i = \log \mathbb{E}_{\hat{q}}\, e^{y_j} = \delta_i - \bar{x}_i$. Jensen's inequality gives $\rho_i \ge 0$, Hoeffding's lemma (Hoeffding, 1963) gives $\rho_i \le \varepsilon_i^2/2$, and $\rho_i \le \varepsilon_i$ because $y_j \le \varepsilon_i$. (a) Since $L_i(j) = e^{y_j - \rho_i}$ and $-2\varepsilon_i \le y_j - \rho_i \le \varepsilon_i$, we have $|L_i(j) - 1| \le e^{2\varepsilon_i} - 1$. The total variation bound follows from $\mathrm{TV} = \frac{1}{2}\mathbb{E}_{\hat{q}}|L_i - 1|$, and the mean bound from $\mu_i - \hat{\mu} = \mathbb{E}_{\hat{q}}[(L_i - 1)(v_j - \hat{\mu})]$ and $\mathbb{E}_{\hat{q}}\|v_j - \hat{\mu}\|_2 \le \sigma_{v,\mathcal{O}}$. (b) Equation (16) gives $|\hat{\pi}_i - \pi_i| = \pi_i(1 - \hat{\pi}_i)\,|e^{-\delta_i} - 1| \le \pi_i(1 - \hat{\pi}_i)(e^{|\delta_i|} - 1)$, and $|\delta_i| \le |\bar{x}_i| + \rho_i$. (c) Insert (a) and (b) into Proposition 2 and use $1 - \hat{\pi}_i \le 1$. □

Centering the remainder. Part (b) shows that the mass error of the zero-order remainder is first order in the mean omitted interaction $\bar{x}_i$, which is the linear term $u_i^\top \bar{w}_{\mathcal{O}}$, with $\bar{w}_{\mathcal{O}} = \mathbb{E}_{\hat{q}}\, w_j$, of the Gaussian estimate in Appendix H, and only second order in its spread. The first-order part exists because the per-query factor e^{c_i} is centered at the full-history mean $\bar{k}$. Centering it at $\bar{k}_{\mathcal{O}}$ instead, that is, giving the remainder the logit $\log S_0 + \beta\,(q_i - q_a)^\top \bar{k}_{\mathcal{O}}$, multiplies the unnormalized remainder weight by $e^{\bar{x}_i}$. The exponent $-\delta_i$ in Equation (16) then becomes $-\rho_i$ with $\rho_i \in [0, \varepsilon_i^2/2]$, so the mass term in (c) drops to at most $\pi_i(1 - e^{-\varepsilon_i^2/2})\, \|\hat{\mu} - \nu_i\|_2$, while the value term is unchanged because $\hat{\mu} = S_1/S_0$ does not depend on the center. That value term has a first-order part of its own: scaling the displacement to λu_i gives $\mu_i(\lambda) - \hat{\mu} = \lambda\, \mathrm{Cov}_{\hat{q}}(v_j, w_j)\, u_i + O(\lambda^2)$, with $\mathrm{Cov}_{\hat{q}}(v_j, w_j) = \mathbb{E}_{\hat{q}}[(v_j - \hat{\mu})(w_j - \bar{w}_{\mathcal{O}})^\top]$, so a zero-order value summary errs at first order in the displacement unless the remainder values are uncorrelated with the key deviations along u_i, whatever the center. The state keeps its size, one d-dimensional mean and $d_v + 1$ numbers, and $\bar{k}_{\mathcal{O}}$ is a sum over $\mathcal{O}$ rather than over the whole history. Because $\rho_i \ge 0$, the recentered remainder never overestimates the tail mass; it underestimates it by the Jensen gap alone. Because ε_i is a worst case over the remainder, the bounds in (a)–(c) are conservative; their role is to show which quantity controls each term.

All remainder results in this paper use the full-history center $\bar{k}$ except the recentered columns of Tables 26 and 27 and the per-item comparison of the two centers in Appendix A.1; in Table 26, centering at $\bar{k}_{\mathcal{O}}$ lowers the test error at every exact-set size m, for both models and both selectors.

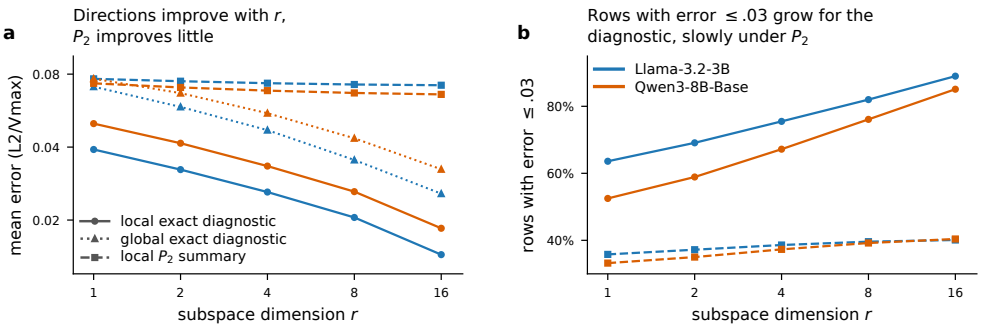

Figure 7: **Subspace dimension sweep across $r \in \{1, 2, 4, 8, 16\}$ on the development split (5120 windows, both models).** (a) Mean L2/Vmax error against subspace dimension r. While the exact-exponential diagnostic (Local Exact) drops monotonically as r increases (reaching .0144 on Llama and .0185 on Qwen), the separable quadratic polynomial (P_2) improves only from .0765 to .0720 on Llama and from .0733 to .0660 on Qwen, barely better than a P_2 summary on the global basis (Table 4). (b) Share of rows with L2/Vmax error $\leq 3\%$. On Llama it grows from 63.6% to 89.0% for the diagnostic, but only from 35.8% to 40.1% under P_2; on Qwen from 52.5% to 85.1% and from 33.2% to 40.4%.

How tight the bounds are. We evaluated the bounds of this appendix in float64 on all 5120 test windows, on the rows after the first of each block, with the full-history center and at $m = N/128$, $N/64$ and $N/16$; none is violated on any row (tolerance 10^{-6}). On the test split the bound of Proposition 4 exceeds the actual error by a median factor of 12.9 on Llama and 13.4 on Qwen at $m = N/128$ (interquartile ranges $[7.5, 23.4]$ and $[8.1, 26.2]$) and of 11.8 and 12.7 at $m = N/16$ ($[7.3, 19.6]$ and $[8.1, 22.4]$). The bound is loose on our data; it identifies the remainder mass as the scale of the error rather than its size. It still orders rows by error (Spearman ρ of .81 and .83 at $m = N/128$), and its ratio to the error grows with $\chi^2(P_{i,\mathcal{O}} \,\|\, \hat{q})$ (Appendix H.3). Equation (13) alone exceeds $\|\mu_i - \hat{\mu}\|_2$ by a median factor of 18.3 and 19.6 at $m = N/128$. Equation (18) is loose on our data as well (median ratio 50.5 and 58.7 at $m = N/128$, 87.1 and 120.9 at $m = N/16$): its mass term $2\tanh(|\delta_i|/4)$ takes the worst case over π_i, so, like Proposition 4, it identifies the scale of the error rather than its size. Equation (17) exceeds $|\hat{\pi}_i - \pi_i|$ by a median factor of 4.8 and 3.4 at $m = N/128$ and 9.2 and 8.5 at $m = N/16$ (over the rows with $|\hat{\pi}_i - \pi_i| > 10^{-6}$) and is attained only where $\mathrm{logit}(\pi_i) = \delta_i/2$. Proposition 5 is non-vacuous, that is, the bound of its part (c) is below $\|o_i\|_2$, on none of the 69,440 Llama rows and on 8 of the 89,280 Qwen rows at $m = N/128$ (14 and 129 rows at $m = N/16$); the median ε_i is 6.9 and 8.0 because ε_i is a worst case over the whole remainder. We therefore read Proposition 5 only as a sufficient condition: it names the quantities that control each term but does not account for the accuracy that the zero-order remainder reaches on our data.

C Direction Quality and Finite-State Kernel Realizations

Object of the comparison. A low-dimensional bilinear interaction need not yield a low-rank matrix after entrywise exponentiation. For example, $x_i = i$ and $y_j = \log z_j$ with distinct positive z_j give a rank-one logit matrix but an exponential matrix with Vandermonde form, which can be full rank. The experiments therefore compare the quality of retained coordinates with the cost and accuracy of concrete kernel realizations; they do not test an algebraic equivalence between these notions.

Section 3 showed that at subspace dimension $r = 8$, local covariance directions carry strong structural signal under an exact exponential, but that a separable positive quadratic polynomial $P_2(t) = 1 + t + t^2/2$ stays far from the diagnostic on the same directions. Here we provide the complete empirical sweep across subspace dimensions $r \in \{1, 2, 4, 8, 16\}$ on the development split (5120 query windows).

Figure 7(a) plots the mean L2/Vmax error as a function of r. For the local diagnostic with an exact exponential, error decreases monotonically with dimension across both architectures: on Llama-3.2-3B, error drops from .0391 at $r = 1$ to .0144 at $r = 16$; on Qwen3-8B-Base, error drops from .0500 at $r = 1$ to .0185 at $r = 16$. The global exact diagnostic also improves steadily, from .0710 to .0257 on Llama and .0766 to .0324 on Qwen.

Table 4: **Subspace dimension sweep on the development split.** Error is mean L2/Vmax; $\leq 3\%$ is the proportion of rows within 3% error. The exact exponential diagnostic captures real low-dimensional structure that improves with r, while the separable P_2 summary improves little with r.

Model	r	Local Exact (Diag.)		Local P_2 Summary		Global Exact	Global P_2
		Mean Err.	$\leq 3\%$	Mean Err.	$\leq 3\%$	Mean Err.	Mean Err.
Llama-3.2-3B	1	.0391	63.6%	.0765	35.8%	.0710	.0817
	2	.0323	69.1%	.0748	37.2%	.0586	.0789
	4	.0261	75.5%	.0734	38.6%	.0470	.0743
	8	.0205	82.0%	.0725	39.6%	.0354	.0726
	16	.0144	89.0%	.0720	40.1%	.0257	.0722
Qwen3-8B-Base	1	.0500	52.5%	.0733	33.2%	.0766	.0833
	2	.0415	58.9%	.0705	35.0%	.0668	.0787
	4	.0334	67.2%	.0684	37.3%	.0552	.0743
	8	.0262	76.1%	.0669	39.2%	.0435	.0711
	16	.0185	85.1%	.0660	40.4%	.0324	.0689

In contrast, the separable quadratic operator (P_2) improves only slightly. On Llama, the local P_2 error is .0765 at $r = 1$ and .0720 at $r = 16$. On Qwen, it moves from .0733 to .0660. At every dimension $r \geq 4$, the local P_2 operator is within about 1% of a global P_2 baseline on Llama and 4–8% below it on Qwen (the global baseline reaches .0722 on Llama and .0689 on Qwen at $r = 16$).

Figure 7(b) displays the fraction of rows achieving error $\leq 3\%$. Under the exact diagnostic, this proportion increases from 63.6% ($r = 1$) to 89.0% ($r = 16$) on Llama, and from 52.5% to 85.1% on Qwen. Under the P_2 realization, the proportion stays between 35.8% and 40.1% on Llama, and between 33.2% and 40.4% on Qwen.

Paired changes on the same rows. The smaller steps on the test split behave like the step from global $r = 1$ to local $r = 8$ in Section 3. On Llama and Qwen respectively, the mean error of P_2 falls by .0043 [.0038, .0047] and .0061 [.0058, .0064] from local $r = 1$ to $r = 8$ (diagnostic: .0178 and .0223), by .0025 [.0011, .0036] and .0055 [.0046, .0062] from global to local directions at $r = 8$ (diagnostic: .0225 and .0219), and by .0009 [.0008, .0011] and .0015 [.0014, .0016] from local $r = 4$ to $r = 8$ (diagnostic: .0053 and .0068). For each of these steps the diagnostic error falls while the error of P_2 rises on 20–34% of the rows after the first of each block, and the two changes have rank correlation .37–.59 across rows. An independent float64 reconstruction from the raw queries, keys and values reproduces the stored errors of Section 3 within 2×10^{-15} in four checked windows, ruling out a numerical explanation there. On the development split, from local $r = 1$ to $r = 16$ (Table 4), the error of P_2 falls by .0045 [.0038, .0051] on Llama and .0074 [.0067, .0080] on Qwen and that of $P_2(t/2)^2$ by .0138 and .0158, against .0246 and .0316 for the diagnostic; the diagnostic error falls while the error of P_2 rises on 39.4% and 28.9% of the rows after the first of each block, and while that of $P_2(t/2)^2$ rises on 20.7% and 16.4%.

A polynomial of higher degree. To test whether a polynomial of higher degree removes this limitation, we evaluated $P_2(t/2)^2 = (1 + t/2 + t^2/8)^2 = 1 + t + t^2/2 + t^3/8 + t^4/64$ at $r = 8$. It is positive and has degree four, but it matches $\exp(t)$ only to second order at $t = 0$, since $e^t - P_2(t/2)^2 = t^3/24 + O(t^4)$. While $P_2(t/2)^2$ expands the number of block states from 45 to 495, each a vector of $d_v + 1$ numbers, it only reduces the mean error on the development split from .0725 to .0488 on Llama and from .0669 to .0496 on Qwen. It remains $2.4\times$ and $1.9\times$ worse than the exact-exponential diagnostic (.0205 and .0262), so raising the degree in this way closes less than half of the gap while the state count grows elevenfold. This polynomial is still bounded below, by $1/4$, so the truncation account of Section 3 applies to it as well. The degree-four Taylor polynomial $T_4(t) = 1 + t + t^2/2 + t^3/6 + t^4/24$, with the same 495 block states and fourth-order accuracy at $t = 0$, reaches .0702 and .0564 on the test split, and the degree-six Taylor polynomial, with 3003 states, reaches .0642 and .0488; neither is more accurate than $P_2(t/2)^2$, which reaches .0469 and .0488 on the same split, so accuracy near $t = 0$ is not what limits these polynomials. The degree-four Taylor polynomial grows as $t^4/24$ for $t \ll 0$ against $t^4/64$, and its excess weight on strongly negative interactions outweighs its gain on positive ones. Higher degrees close more of the gap only at a state count that grows quickly: a polynomial of degree n in t that keeps every monomial of the r projected

coordinates needs $\binom{r+n}{n}$ block states, 3003 and 12,870 for $n = 6$ and $n = 8$ at $r = 8$, which is comparable to the number of history keys (up to 16,352). Appendix C.1 reports these and the other separable realizations we tested.

The diagnostic is not a lower bound. Exponentiating the projected interaction exactly does not minimize the output error over functions of the retained coordinates, because the error of the projection and the error of the kernel can cancel. Take four history keys $k = (2s, \tau)$ with $s, \tau \in \{-1, +1\}$, an anchor $q_a = 0$, so that p_a is uniform, $\bar{k} = 0$ and $C = \mathrm{diag}(4, 1)$, and $R = I$, so that $r = 1$ retains the first coordinate. For a query with $\beta q_i = (1/2, 1)$ the logits are $s + \tau$ and the projected interaction is $t = s$. With scalar values $v = s + a\tau$ and $a = 2/(3 \tanh 1) - 1$, the exact output is $(1 + a) \tanh 1 = 2/3$. The diagnostic weights the keys by e^s and returns $\tanh 1 \approx 0.762$, while P_2, with $P_2(1) = 5/2$ and $P_2(-1) = 1/2$, returns $2/3$ exactly. Adding in-block keys with key 0 and value $2/3$ leaves P_2 exact and the diagnostic in error. On real windows the same happens on 10.2% of test rows at local $r = 8$ (Section 3).

What controls the output error of a kernel approximation. A pointwise guarantee asks for $P(t) \approx e^t$ at every key, but the output error depends on a weaker, weighted quantity. Consider any realization that gives history key j the unnormalized weight $e^{s_a(j)+c_i}\, g_{ij}$ with $g_{ij} \geq 0$ and keeps the legal block keys exact, and let

$$\Gamma_i \;=\; \sum_{j=1}^{N} p_a(j)\, \left| g_{ij} - e^{u_i^\top w_j} \right|$$

be its anchor-weighted kernel error against the dense interaction $u_i^\top w_j$ of Equation (4). A separable summary has $g_{ij} = P(t_{ij})$ and the exact-exponential diagnostic $g_{ij} = e^{t_{ij}}$; an anchor-gated representation has $g_{ij} = e^{u_i^\top w_j}$ on its exact set, so only the remainder keys enter Γ_i, and the zero-order remainder has $g_{ij} = 1$ there.

Proposition 1 (Weighted kernel error, restated from Section 3)**.** *The full causal distribution $\hat{P}_i$ of the realization satisfies $\|\hat{P}_i - P_i\|_1 \leq 2h_i\Gamma_i/M_i$, and $\|\hat{o}_i - o_i\|_2 \leq 2h_i V_{\max}\Gamma_i/M_i$, with h_i and M_i as in Appendix B.*

Proof. In units of $Z_a e^{c_i}$ the two rows have the same block weights, their history weights differ in ℓ_1 norm by Γ_i, and the dense row has total weight M_i/h_i. For nonnegative vectors x and y with positive sums X and Y, $\|x/X - y/Y\|_1 \leq \|x - y\|_1/Y + |X - Y|/Y \leq 2\|x - y\|_1/Y$. The output bound follows from $\hat{o}_i - o_i = \sum_j (\hat{P}_i(j) - P_i(j))\, v_j$. □

The proposition needs no pointwise guarantee: each key enters Γ_i in proportion to its anchor mass, so a large pointwise error on keys of negligible anchor mass is compatible with a small output error. It is an upper bound only; the errors in Table 4 are measured directly.

C.1 Other separable realizations

Candidates. Tables 5 to 8 compare P_2 with other realizations that give every history key a weight from one function of its projected interaction, on the local $r = 8$ coordinates of Section 3, in float64, on all 5120 test windows (both captured heads, every layer, quarter and late windows). History key j of query i receives the unnormalized weight $e^{s_a(j)+c_i}\, g(t_{ij})$, the keys inside the block stay exact, and nothing is truncated. The candidates are the rescaled squares $P_2(t/s)^s$ for $s = 2, 4$; the Taylor polynomials $T_n(t) = \sum_{k \leq n} t^k/k!$ for $n = 4, 6, 8$, which are positive because n is even; P_2 expanded around a per-query point, $e^{s_i} P_2(t - s_i)$, with $s_i = \sum_j p_a(j)\, t_{ij}^2$ or, as an oracle, the mean of t under the diagnostic's attention; Chebyshev interpolants of e^t of degree 4, 6 and 8, and squares of Chebyshev interpolants of $e^{t/2}$ of half that degree, each on a fixed interval $[-B, B]$ or, as an oracle, on the range of t in the window; a per-window least-squares fit of degree 4, 6 or 8 weighted by the diagnostic's attention (oracle); and orthogonal positive random features (Choromanski et al., 2021) for e^t on the eight projected coordinates with $M = 45$, 495 and 3003 features, averaged over five seeds. B is the 99th percentile over rows of the largest $|t|$ on the development split, 16.5 on Llama and 21.9 on Qwen. A polynomial of degree n in the eight coordinates needs $\binom{8+n}{n}$ block states, 45, 495, 3003 and 12,870 for $n = 2, 4, 6, 8$, and random features need M; each state is a vector of

$d_v + 1$ numbers. For quantitative comparison, the closure $g = (e_{P_2} - e)/(e_{P_2} - e_{\text{diag}})$ is the share of the gap between P_2 and the diagnostic that a candidate closes; we fixed before the sweep that a candidate closes the gap if $g \geq .75$ in both models. Figure 8 shows the pointwise error of the positive polynomials and the distribution of t.

Results. The mismatch is concrete: at $t = -10$ the true factor is about 4.5×10^{-5} while $P_2(t) = 41$, and at $t = 10$ it is about 22,026 while the polynomial gives 61. With at most 495 block states, none of these realizations closes three quarters of the gap between P_2 (.0695 and .0653 on Llama and Qwen) and the diagnostic (.0205 and .0268). The best, $P_2(t/2)^2$, reaches .0469 and .0488 and closes .46 and .43 of it. Evaluating the top $N/128$ keys exactly and keeping the same P_2 on the remaining keys reaches .0228 and .0290 and closes .954 [.947, .962] and .944 [.930, .958] of the same gap (the exact-set rows of Tables 6 and 8); this is the evidence behind Section 4: the part of the gap that no tested separable kernel closes at a moderate state count lies on the few high-mass keys, and the zero-order state of Section 5 (.0272 and .0338) already closes .865 and .819 of it. With the same 495 states, T_4 is less accurate than $P_2(t/2)^2$ by .0233 [.0223, .0243] and .0076 [.0067, .0085], and on Llama no more accurate than P_2 (+.0007 [−.0003, +.0019]); T_6, with 3003 states, is less accurate than $P_2(t/2)^2$ by .0173 [.0162, .0184] on Llama and indistinguishable from it on Qwen (−.00003 [−.0010, +.0009]). On the projected coordinates, the anchor-weighted kernel error of Proposition 1 per unit of mass, Γ_i/M_i, restricted to the keys with $t < 0$ averages .37 and .16 for T_4 against .086 and .038 for $P_2(t/2)^2$, and restricted to $t \geq 0$ it is .085 and .18 against .13 and .23. At 495 and at 12,870 states the rescaled square is more accurate than the Taylor polynomial with the same state count in both models. Positive random features are less accurate than P_2 at every size, at .1090 and .1176 with 45 features, .0916 and .1043 with 495 and .0832 and .0977 with 3003, and expanding P_2 around a per-query point raises its error to about .16 for either choice of the point. Raw Chebyshev interpolants give some keys negative weight on 71 to 97% of rows, and even on the range of t in each window their normalizer is not positive on 20 to 39% of rows, so their mean errors exceed .5; the least-squares fits also give negative weights, and their mean errors exceed .35. Squared, positive Chebyshev interpolants avoid this but remain less accurate than P_2, at best .0743 and .0999 (degree eight on the window's range). Only positive polynomials of degree eight come close: $P_2(t/4)^4$ reaches .0268 and .0366 and closes .873 [.864, .882] of the gap on Llama and .746 [.734, .757] on Qwen, with an interval crossing three quarters, using 12,870 block states, and T_8 with the same state count reaches .0549 and .0421.

Detailed kernel table conventions. The following tables retain every tested realization. Errors and intervals are L2/$V_{\max}$ cluster means on the test split, including anchor rows. A state stores $d_v + 1$ numbers. The closure g uses the P_2 and exact-exponential references of the same model. ‡ marks an oracle requiring all projected interactions or diagnostic attention; these are not deployable competitors. Fixed Chebyshev intervals are $[-B, B]$, with $B = 16.5/21.9$ (Llama/Qwen); "range" uses the observed window range. Signed fits can assign negative key weights. Invalid denotes a nonpositive normalizer or nonfinite output error; means retain finite errors even on invalid rows and exclude nonfinite ones. † marks full-dimensional random features, whose target is dense attention rather than the projected diagnostic. Sign-split and gated rows are not separable. The latter use $N/128$ exact keys and the stated remainder treatment.

Sign split. To see which departure of P_2 costs more, we keep P_2 on one side of $t = 0$ and restore e^t on the other; these counterfactuals are oracles and are not separable. Let g_- be the closure when e^t is restored on $t < 0$ and g_+ when it is restored on $t \geq 0$; the two shares need not sum to one. For P_2, $g_- = .601$ [.579, .622] and $g_+ = .414$ [.393, .435] on Llama, and $g_- = .319$ [.298, .339] and $g_+ = .703$ [.687, .720] on Qwen. We had fixed, before the test sweep but after a pilot on 320 windows, that a side dominates if its share is at least .60 and the other's at most .40 in both models; neither side does, so both departures matter, and which matters more depends on the model. $P_2(t/2)^2$ splits its own gap to the diagnostic the same way (.585 and .416 on Llama, .212 and .797 on Qwen; the g column of Tables 6 and 8 measures these rows against the gap of P_2 instead). The split is in line with where t falls (Figure 8(b)): on average over rows 1 to 31, 2.2% and 1.2% of the anchor mass lies at $t < -5$, where P_2 exceeds e^t more than a thousandfold, and 6.1% and 15.6% of the dense attention lies at $t > 5$, where P_2 falls short of e^t by a factor of eight or more. Inside the anchor-gated representation at $m = N/128$, where a remainder on the same directions covers only the keys outside the exact set, restoring e^t on $t \geq 0$ closes .83 and .90 of the gap between a P_2 remainder and an exact-exponential remainder, and restoring it on $t < 0$ closes .12 and .08.

Table 5: Kernel realizations on local rank-eight coordinates: Llama-3.2-3B, positive candidates and random features. Mean error and 95% cluster-bootstrap interval, percentage of rows below 3% error, gap closure and invalid-row percentage.

Realization	States	Mean [95% CI]	$\leq 3\%$	g	Invalid (%)
Reference					
exact e^t (diagnostic)	–	.0205 [.0197, .0215]	82.0	–	–
$P_2(t)$	45	.0695 [.0669, .0723]	41.2	–	–
Positive polynomials and shifted expansions					
$P_2(t/2)^2$	495	.0469 [.0447, .0492]	60.9	.462	–
$P_2(t/4)^4$	12,870	.0268 [.0254, .0282]	75.5	.873	–
T_4	495	.0702 [.0674, .0732]	51.6	−.014	–
T_6	3,003	.0642 [.0614, .0670]	60.0	.109	–
T_8	12,870	.0549 [.0525, .0574]	66.5	.298	–
shifted P_2, variance center	45	.1621 [.1572, .1669]	29.6	−1.893	–
shifted P_2, oracle center‡	45	.1604 [.1538, .1670]	32.2	−1.857	–
squared Cheb., $[-B, B]$, degree 4	495	.2091 [.2061, .2120]	4.0	−2.852	–
squared Cheb., $[-B, B]$, degree 6	3,003	.1208 [.1186, .1231]	27.1	−1.048	–
squared Cheb., $[-B, B]$, degree 8	12,870	.2114 [.2088, .2140]	4.0	−2.898	–
squared Cheb., range‡, degree 4	495	.1094 [.1059, .1128]	32.9	−.815	–
squared Cheb., range‡, degree 6	3,003	.0969 [.0937, .1003]	36.6	−.560	–
squared Cheb., range‡, degree 8	12,870	.0743 [.0712, .0778]	50.8	−.098	–
Positive random features (mean of five seeds)					
$r = 8$ coordinates, $M = 45$	45	.1090 [.1052, .1124]	34.0	−.806	–
$r = 8$ coordinates, $M = 495$	495	.0916 [.0882, .0948]	39.2	−.452	–
$r = 8$ coordinates, $M = 3003$	3,003	.0832 [.0803, .0858]	42.2	−.279	–
full dimension†, $M = 45$	45	.2460 [.2409, .2508]	7.8	–	–
full dimension†, $M = 495$	495	.2199 [.2143, .2251]	11.1	–	–
full dimension†, $M = 3003$	3,003	.2007 [.1946, .2062]	13.9	–	–

Table 6: Kernel realizations on local rank-eight coordinates: Llama-3.2-3B, signed fits and diagnostic interventions. Mean error and 95% cluster-bootstrap interval, percentage of rows below 3% error, gap closure and invalid-row percentage.

Realization	States	Mean [95% CI]	$\leq 3\%$	g	Invalid (%)
Signed fits					
Cheb., $[-B, B]$, degree 4	495	2.13 [1.84, 2.48]	3.1	−42.2	95.6
Cheb., $[-B, B]$, degree 6	3,003	3.59 [2.79, 4.53]	3.2	−71.8	5.0
Cheb., $[-B, B]$, degree 8	12,870	3.71 [3.35, 4.06]	3.5	−74.3	86.9
Cheb., range‡, degree 4	495	.7896 [.6279, .9941]	20.6	−14.7	39.2
Cheb., range‡, degree 6	3,003	.6060 [.4767, .7570]	35.4	−11.0	26.3
Cheb., range‡, degree 8	12,870	.5346 [.3435, .8066]	46.5	−9.500	20.4
weighted LS‡, degree 4	495	.3971 [.3224, .4829]	32.2	−6.692	4.9
weighted LS‡, degree 6	3,003	2.38 [.4024, 6.19]	44.4	−47.2	7.2
weighted LS‡, degree 8	12,870	.3598 [.3002, .4380]	56.5	−5.930	6.7
Sign-split interventions (not separable)					
P_2 with e^t on $t < 0$	–	.0401 [.0377, .0426]	66.3	.601	–
P_2 with e^t on $t \geq 0$	–	.0492 [.0479, .0506]	52.1	.414	–
$P_2(t/2)^2$ with e^t on $t < 0$	–	.0315 [.0297, .0334]	72.3	.777	–
$P_2(t/2)^2$ with e^t on $t \geq 0$	–	.0359 [.0346, .0372]	69.9	.686	–
Exact set plus remainder (not separable)					
zero-order remainder	–	.0272 [.0261, .0282]	75.1	.865	–
P_2 remainder on $r = 8$	–	.0228 [.0218, .0238]	79.8	.954	–
exact-e^t remainder on $r = 8$	–	.0172 [.0164, .0180]	86.0	1.069	–
P_2 remainder, e^t on $t < 0$	–	.0221 [.0211, .0231]	80.8	.968	–
P_2 remainder, e^t on $t \geq 0$	–	.0181 [.0173, .0190]	84.8	1.050	–

Table 7: Kernel realizations on local rank-eight coordinates: Qwen3-8B-Base, positive candidates and random features. Mean error and 95% cluster-bootstrap interval, percentage of rows below 3% error, gap closure and invalid-row percentage.

Realization	States	Mean [95% CI]	$\leq 3\%$	g	Invalid (%)
Reference					
exact e^t (diagnostic)	–	.0268 [.0259, .0278]	74.9	–	–
$P_2(t)$	45	.0653 [.0625, .0679]	40.9	–	–
Positive polynomials and shifted expansions					
$P_2(t/2)^2$	495	.0488 [.0466, .0510]	56.0	.427	–
$P_2(t/4)^4$	12,870	.0366 [.0352, .0380]	66.2	.746	–
T_4	495	.0564 [.0536, .0592]	51.0	.230	–
T_6	3,003	.0488 [.0461, .0513]	59.2	.428	–
T_8	12,870	.0421 [.0398, .0442]	65.1	.602	–
shifted P_2, variance center	45	.1609 [.1546, .1671]	25.6	-2.488	–
shifted P_2, oracle center‡	45	.1631 [.1553, .1706]	27.1	-2.544	–
squared Cheb., $[-B, B]$, degree 4	495	.1740 [.1715, .1765]	6.4	-2.828	–
squared Cheb., $[-B, B]$, degree 6	3,003	.1512 [.1479, .1548]	20.4	-2.235	–
squared Cheb., $[-B, B]$, degree 8	12,870	.1763 [.1737, .1788]	6.2	-2.888	–
squared Cheb., range‡, degree 4	495	.1253 [.1231, .1275]	27.9	-1.562	–
squared Cheb., range‡, degree 6	3,003	.1181 [.1158, .1204]	31.3	-1.375	–
squared Cheb., range‡, degree 8	12,870	.0999 [.0969, .1029]	40.4	$-.902$	–
Positive random features (mean of five seeds)					
$r = 8$ coordinates, $M = 45$	45	.1176 [.1146, .1203]	22.7	-1.361	–
$r = 8$ coordinates, $M = 495$	495	.1043 [.1016, .1069]	26.9	-1.015	–
$r = 8$ coordinates, $M = 3003$	3,003	.0977 [.0954, .1000]	29.5	$-.844$	–
full dimension†, $M = 45$	45	.1910 [.1890, .1931]	6.5	–	–
full dimension†, $M = 495$	495	.1765 [.1743, .1789]	8.7	–	–
full dimension†, $M = 3003$	3,003	.1660 [.1635, .1686]	10.6	–	–

Table 8: Kernel realizations on local rank-eight coordinates: Qwen3-8B-Base, signed fits and diagnostic interventions. Mean error and 95% cluster-bootstrap interval, percentage of rows below 3% error, gap closure and invalid-row percentage.

Realization	States	Mean [95% CI]	$\leq 3\%$	g	Invalid (%)
Signed fits					
Cheb., $[-B, B]$, degree 4	495	4.54 [2.78, 7.89]	3.1	-116.5	94.6
Cheb., $[-B, B]$, degree 6	3,003	5.30 [4.30, 6.49]	3.2	-136.1	8.8
Cheb., $[-B, B]$, degree 8	12,870	4.44 [4.09, 4.79]	3.5	-113.9	81.1
Cheb., range‡, degree 4	495	1.02 [.6369, 1.51]	18.7	-24.7	36.3
Cheb., range‡, degree 6	3,003	.7878 [.5928, 1.02]	28.0	-18.8	33.1
Cheb., range‡, degree 8	12,870	.6541 [.5734, .7484]	35.4	-15.3	24.4
weighted LS‡, degree 4	495	.4226 [.3481, .5140]	22.8	-9.294	6.2
weighted LS‡, degree 6	3,003	.6139 [.5215, .7256]	33.0	-14.3	10.8
weighted LS‡, degree 8	12,870	.5784 [.4308, .7735]	43.1	-13.3	10.5
Sign-split interventions (not separable)					
P_2 with e^t on $t < 0$	–	.0530 [.0508, .0553]	55.0	.319	–
P_2 with e^t on $t \geq 0$	–	.0382 [.0367, .0398]	59.6	.703	–
$P_2(t/2)^2$ with e^t on $t < 0$	–	.0442 [.0424, .0460]	61.0	.548	–
$P_2(t/2)^2$ with e^t on $t \geq 0$	–	.0313 [.0299, .0326]	70.2	.884	–
Exact set plus remainder (not separable)					
zero-order remainder	–	.0338 [.0326, .0350]	67.8	.819	–
P_2 remainder on $r = 8$	–	.0290 [.0278, .0301]	74.1	.944	–
exact-e^t remainder on $r = 8$	–	.0216 [.0206, .0226]	81.5	1.135	–
P_2 remainder, e^t on $t < 0$	–	.0284 [.0272, .0295]	74.7	.960	–
P_2 remainder, e^t on $t \geq 0$	–	.0223 [.0213, .0233]	80.8	1.117	–

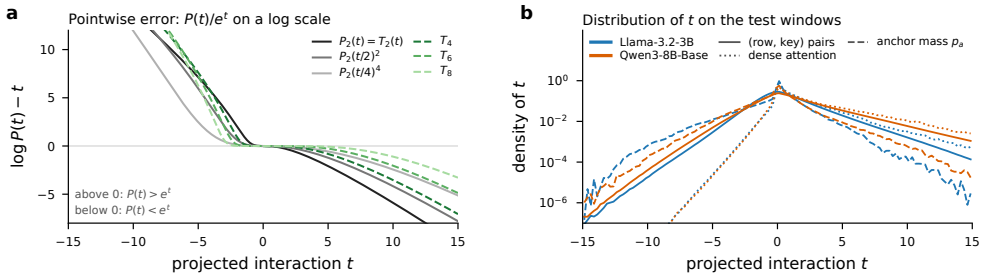

Figure 8: **Pointwise error of the positive polynomials, and where t falls.** (a) $\log P(t) - t = \log(P(t)/e^t)$: above zero a polynomial overweights a key relative to e^t, below zero it underweights it; $T_2 = P_2$, and T_n is the Taylor polynomial of degree n. (b) Distribution of the projected interaction t on the local $r = 8$ coordinates of the test windows, over rows 1 to 31 and all history keys (on the anchor row $t = 0$ for every key), weighted by count, by dense attention and by anchor mass p_a; window-equal average of per-window histograms with bins of width 0.25.

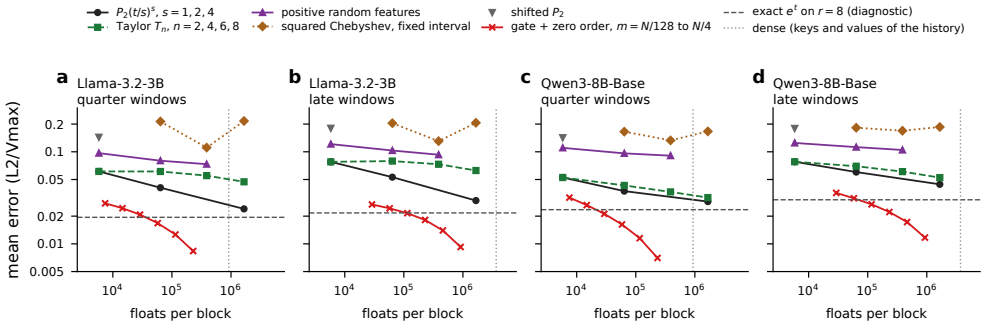

Figure 9: **Error against the floats kept per block, test split, windows a quarter into the sequence and final windows.** Separable realizations on the local $r = 8$ coordinates keep their states times $d_v + 1$ floats per block, not counting the eight directions: the $P_2(t/s)^s$ and Taylor lines both start at $P_2 = T_2$, random features have $M = 45$, 495 and 3003, and the squared Chebyshev interpolants use the fixed interval, degree 4, 6 and 8. The zero-order gate of Section 5 at $m = N/128, N/64, \ldots, N/4$ keeps $m(d + d_v) + d + d_v + 1$ floats, the keys and values of its exact set and the remainder state, averaged over the windows. Dashed: the exact-exponential diagnostic on the same directions, which reads every key. Dotted: $N(d + d_v)$, the keys and values of the whole history, averaged over the windows. Cluster means as in Table 5.

How tight the weighted bound is. On the test windows, Proposition 1 holds on every row for every positive realization, both against dense attention and, with t_{ij} in place of $u_i^\top w_j$, against the diagnostic. On rows 1 to 31, for P_2 against dense attention, the bound exceeds the error by a median factor of 11.6 on Llama and 14.9 on Qwen, and it orders rows by error with Spearman ρ of .88 and .71. The bound thus ranks rows by difficulty but overstates the error by about an order of magnitude, so we use it only qualitatively.

Quarter and late windows. Figure 9 plots the error against the floats that each realization keeps per block, separately for the windows a quarter into the sequence and the final windows, together with the zero-order gate of Section 5 at $m = N/128$ to $N/4$ on the same windows. In each of the four panels, every separable realization with more than 45 block states is less accurate than the gate at some budget that keeps fewer floats per block, and the 45-state realizations, which keep fewer floats than any gate budget, are less accurate than the gate at every budget. In the quarter windows the degree-eight polynomials keep more numbers per block (1.66 million) than the keys and values of the whole history (0.92 and 0.93 million on average). The rescaled squares $P_2(t/s)^s$, the Taylor polynomials and the random features are less accurate on the final windows, whose history is about four times longer, than on the quarter windows in both models (P_2: .0612 to .0778 and .0526 to .0779), while the error of the gate at $m = N/128$ changes by less than .005 (.0276 to .0268 and .0318 to .0358).

Table 9: **Projected-only against full-logit correction on the top $N/128$ keys** (test split; mean L2/Vmax against dense attention, 95% cluster-bootstrap intervals; closure against the diagnostic, g_{E} against arm E). The letters follow the arms requested in the second-round review.

				Llama-3.2-3B		Qwen3-8B-Base	
Arm	on $\mathcal{H}$	elsewhere		error [95%]	closure [95%] (g_{E})	error [95%]	closure [95%] (g_{E})
A	$P_2(t)$	$P_2(t)$		.0695 [.0669, .0723]	—	.0653 [.0625, .0679]	—
B	e^t	$P_2(t)$	new	.0259 [.0248, .0272]	.890 [.885, .894] (.832)	.0339 [.0329, .0350]	.815 [.803, .825] (.717)
C	$e^{u^\top w}$	$P_2(t)$	Section 4	.0228 [.0218, .0238]	.954 [.947, .962] (.892)	.0290 [.0278, .0301]	.944 [.930, .958] (.832)
D	e^t	e^t	diagnostic	.0205 [.0197, .0215]	—	.0268 [.0259, .0278]	—
E	$e^{u^\top w}$	e^t		.0172 [.0164, .0180]	1.069 [1.064, 1.075]	.0216 [.0206, .0226]	1.135 [1.124, 1.148]

C.2 Controls for the exact-set intervention

The intervention of Section 4 gives the top $N/128$ keys their full logits $e^{u_i^\top w_j}$, while the diagnostic uses the projected $e^{t_{ij}}$, so on those keys it removes the polynomial error and the projection error at once; and attention sinks, positions 0 to 3, carry much of the set's anchor mass. The controls below separate these effects. They were run on the same 5,120 test windows with copies of the analyzers of Appendix C.1 and Appendix D.2, with the rules frozen before a pilot; the copies reproduce every archived per-window error of Tables 6 and 8 and of Appendix D.2 bit for bit, and the aggregate closure of the full-logit arm, .9536 and .9441, to the last digit. Every arm keeps P_2 on the keys it does not make exact and the legal in-block keys exact; closure is $(E[P_2] - E[\text{arm}])/(E[P_2] - E[\text{diag}])$, a ratio of cluster means with jointly resampled 95% cluster-bootstrap intervals, and $\mathcal{H}$ is the top $N/128$ set of Section 4, S the positions 0 to 3, and $\mathcal{H}'$ the top $N/128$ keys among positions ≥ 4.

What the intervention changes. Let D_i be the full causal distribution of the diagnostic, $o_i^D = \sum_j D_i(j)\, v_j$ its output, and write the unnormalized weights of a nonnegative realization as $w_{ij}^D(1 + \epsilon_{ij})$ with $\epsilon_{ij} = 0$ on the legal block keys; for P_2, $\epsilon_{ij} = P_2(t_{ij})\, e^{-t_{ij}} - 1$. With $\bar{\epsilon}_i = \sum_j D_i(j)\, \epsilon_{ij}$,

$$o_i' - o_i^D \;=\; \frac{\sum_j D_i(j)\, \epsilon_{ij}\, (v_j - o_i^D)}{1 + \bar{\epsilon}_i},$$

because $D_i'(j) = D_i(j)(1 + \epsilon_{ij})/(1 + \bar{\epsilon}_i)$. The error relative to the diagnostic is a sum over keys of diagnostic probability, relative kernel error and the value's deviation from the diagnostic output, divided by the changed normalizer. The projected-only correction sets $\epsilon_{ij} = 0$ on $\mathcal{H}$ and leaves the sum over the remainder; the full-logit correction also replaces w_{ij}^D by $e^{u_i^\top w_j}$ there, which is why the two arms differ. These controls, like Section 4, use the 20 test items of the two development models; they have not been repeated on fresh traces or on a third architecture.

Projected-only correction. Table 9 lists the five arms of the comparison. Replacing only $P_2(t)$ by e^t on $\mathcal{H}$ (arm B) closes .890 [.885, .894] and .815 [.803, .825] of the gap on Llama and Qwen and reproduces .933 [.927, .938] and .863 [.851, .874] of the reduction of the full-logit arm C; the paired difference $E[\mathrm{B}] - E[\mathrm{C}] = +.0031$ [+.0028, +.0035] and +.0050 [+.0047, +.0053] is the part that the omitted directions on $\mathcal{H}$ contribute. With arm E (full logits on $\mathcal{H}$, e^t elsewhere) as the endpoint, the closures are .832 and .717 for B and .892 and .832 for C. In the terms of Proposition 1, of P_2's projected weighted kernel error Γ_i/M_i (mean over rows 1 to 31, cluster mean .355 and .414), .284 and .288 lies on $\mathcal{H}$; arm B removes exactly that part.

Sinks exact in every arm. Table 10 makes positions 0 to 3 exact with full logits in every comparator. Alone they close .316 [.294, .340] and .418 [.396, .446] of the gap, and $\mathcal{H} \setminus S$ alone closes .660 [.639, .681] and .651 [.629, .673]. With S exact in both arms, adding the rest of $\mathcal{H}$ lowers the error by .0314 [.0306, .0322] and .0205 [.0196, .0213], which is .641 and .533 of the gap: .673 [.650, .696] and .565 [.537, .587] of the closure of Section 4 survives the control, and the sinks account for .332 and .443 of the gate's reduction. Against an endpoint that also keeps S exact (full logits on S, e^t elsewhere), $\mathcal{H} \cup S$ closes .922 [.912, .933] and .891 [.870, .913] of the remaining gap. With e^t instead of full logits on the exact sets the numbers are .306 and .401 for S alone, survival .658 and .510, and sink-controlled closure .843 and .694. In the median window $\mathcal{H}$ contains one of the four sink positions (85% and 77% of windows contain exactly one) and its anchor mass is .924 and .893, of which .793 and .314 lies on S.

Table 10: **Sink controls and the sink-excluded basis** ($m = N/128$, test split). Left: exact sets with full logits, P_2 elsewhere; closure against the diagnostic, and against an endpoint with S exact for the last row. Right: the exact-exponential diagnostic of bases built from different key sets (full-history center; the recentered variant recomputes the center from positions ≥ 4 and answers a different question); capture is the median share of the variance of the optimal rank-eight basis.

	Llama-3.2-3B		Qwen3-8B-Base			Llama-3.2-3B		Qwen3-8B-Base	
Exact set	error	closure	error	closure	Basis from	error	capture	error	capture
S (positions 0–3)	.0540	.316	.0492	.418	full history $U_8(J)$	.0205	1	.0268	1
$\mathcal{H} \setminus S$	.0372	.660	.0403	.651	frozen global	.0430	—	.0487	—
$\mathcal{H}$	.0228	.954	.0290	.944	$\mathcal{H}$ (restricted, Appendix D.2)	.0223	.978	.0289	.985
$\mathcal{H} \cup S$	.0226	.958	.0287	.951	$\mathcal{H}'$ (no positions 0–3)	.0273	.969	.0321	.980
$\mathcal{H} \cup S$, S-exact endpoint		.922		.891	$\mathcal{H}'$, recentered	.0455	.827	.0389	.950
$\mathcal{H}'$	.0371	.662	.0401	.654	all positions ≥ 4	.0226	1.000	.0287	1.000

Sink-excluded basis. The right half of Table 10 rebuilds the restricted basis from $\mathcal{H}'$, whose median anchor mass is .157 and .503 against .924 and .893 for $\mathcal{H}$. Its diagnostic error, .0273 [.0260, .0285] and .0321 [.0309, .0334], exceeds the full-history basis by $+.0067$ [$+.0061$, $+.0074$] and $+.0053$ [$+.0048$, $+.0058$] and the $\mathcal{H}$ basis by $+.0050$ and $+.0032$, and beats the global basis by $-.0158$ [$-.0181$, $-.0136$] and $-.0166$ [$-.0185$, $-.0147$]; its median capture is .969 and .980. Made exact with full logits and P_2 elsewhere, $\mathcal{H}'$ closes .662 [.641, .683] and .654 [.633, .676] of the gap (.612 and .547 with e^t). At $m = N/64$ the $\mathcal{H}'$ basis captures .982 and .989 and errs .0260 and .0310, and the exact set closes .703 and .750. Recomputing the center from the sink-excluded history raises the error to .0455 and .0389 and lowers the capture to .827 and .950, because the center itself carries the sink mass; the basis from every key at positions ≥ 4 with the full-history center errs .0226 and .0287.

Which keys are made exact. With $m = N/128$ keys exact (full logits) and P_2 elsewhere, the closure is .014 and .026 for uniform sets, .051 and .122 for the next-ranked keys (ranks $m+1$ to $2m$), .900 and .869 for sets drawn by anchor mass without replacement, .954 and .944 for the top-m set, .971 and .983 for mass-matched sets, which need a median $2.7m$ and $2.2m$ keys to reach the mass of $\mathcal{H}$, and 1.083 and 1.200 for a per-row oracle that picks the m keys with the largest $p_a(j)\,|P_2(t_{ij}) - e^{t_{ij}}|$ (three seeds averaged within a window for the random sets; every paired difference from the top-m arm excludes zero; with e^t on the exact set the six closures are .009, .028, .854, .890, .893 and .947 on Llama and .015, .061, .770, .815, .826 and .922 on Qwen). The row-mean anchor mass of the oracle set has median .707 and .567.

Capture and closure across windows. The per-window closure of the full-logit arm has median .988 and .977 (10th percentiles .446 and .413). Over the windows with a positive gap, it has Spearman $\rho = .273$ [.230, .315] and .234 [.199, .272] with the capture $\mathcal{C}_{\mathcal{H}}$ of Section 4 (.208 and .172 after controlling for the anchor mass of $\mathcal{H}$), .217 and .162 with that mass, and .205 and .128 with the part of P_2's projected weighted kernel error that lies on $\mathcal{H}$; for the projected-only arm the three correlations are .407 and .445, .420 and .423, and .472 and .535 (task-stratified cluster-bootstrap intervals exclude zero throughout). Geometry sufficiency and gap closure are thus related at the window level, but weakly, largely because the closure is high in most windows; this analysis is descriptive and was added after the rules above were frozen.

C.3 What a pointwise guarantee costs

Proposition 1 needs only the anchor-weighted kernel error; Remark 2 states what the stronger pointwise guarantee costs, which a Taylor summary such as P_2 implicitly targets near $t = 0$. By a pointwise guarantee we mean a bound on the relative kernel error $|g\,e^{-t} - 1|$ that holds for every t in an interval, whichever queries and keys produce it. A separable realization on the local coordinates gives a query with coordinates $x = U_r^\top u_i$ and a history key with coordinates $y = U_r^\top w_j$, both in $\mathbb{R}^r$, the kernel value $\phi(x)^\top \psi(y)$ for maps $\phi, \psi\colon \mathbb{R}^r \to \mathbb{R}^D$, so that $t_{ij} = x^\top y$, and it keeps D block states of $d_v + 1$ numbers each.

Remark 2 (Cost of a pointwise guarantee)**.** *Let $0 < \varepsilon < 1$. (a) If $\phi(x)^\top \psi(y) = p(x^\top y)$ for all x and y in two nonempty open sets, with one polynomial $p(t) = \sum_{k=0}^{n} c_k t^k$ of degree n for all queries, then $D \geq \sum_{k:\, c_k \neq 0} \binom{r+k-1}{k} \geq \binom{r+n-1}{n}$. At $r = 8$ the sum is 45 for P_2 and 495 for $P_2(t/2)^2$ and T_4, the state counts of Table 5. The last bound still holds if the polynomial changes with the query, as long as ψ is fixed and the polynomial has degree n for every x in the open*

set. (b) If $|p(t)e^{-t} - 1| \le \varepsilon$ *for every* t *in an interval of width* $2W > 0$*, then* $(1-\varepsilon)e^{W} \le (1+\varepsilon)\cosh(n \operatorname{arccosh} 3)$*, that is,* $n \ge \operatorname{arccosh}\big(\max\{1, \frac{1-\varepsilon}{1+\varepsilon} e^{W}\}\big) / \operatorname{arccosh} 3$*. (c) Positive random features with* D *i.i.d. Gaussian directions (Choromanski et al., 2021) estimate* $e^{x^\top y}$ *without bias and with relative variance* $(e^{\|x+y\|^2} - 1)/D$*. Since* $\|x+y\|^2 \ge 4x^\top y$*, a relative root-mean-square error of at most* ε *on a key with* $x^\top y \ge \tau > 0$ *needs* $D \ge (e^{4\tau} - 1)/\varepsilon^2$*. (d) Conversely, if the weights* g_{ij} *of a realization as in Proposition 1 satisfy* $|g_{ij} - e^{t_{ij}}| \le \varepsilon e^{t_{ij}}$ *at every history key, then* $\Gamma_i \le \varepsilon M_i$ *with* t_{ij} *in place of* $u_i^\top w_j$ *in* Γ_i*,* M_i *and* h_i*, and Proposition 1 places the realization within* $2\varepsilon h_i$ *of the diagnostic in* ℓ_1 *and within* $2\varepsilon h_i V_{\max}$ *in output.*

Justification. (a) For fixed y, the function $x \mapsto p(x^\top y)$ lies in the span of the D coordinate functions of ϕ, so these functions, over y in the open set, span a space S of dimension at most D. Replacing y by λy for $n+1$ distinct λ near 1 and inverting a Vandermonde matrix puts every $c_k(x^\top y)^k$ in S. For $c_k \ne 0$ these powers span all homogeneous polynomials of degree k in x, a space of dimension $\binom{r+k-1}{k}$, because the coefficient of x^α in $(x^\top y)^k$ is a positive multiple of y^α and distinct monomials are linearly independent on an open set; homogeneous parts of different degrees are linearly independent, and a polynomial that vanishes on an open set is zero. If the polynomial changes with the query, fix x instead: the functions $y \mapsto p_x(x^\top y)$ lie in the span of the coordinates of ψ, and their parts of degree n, $c_n(x)(x^\top y)^n$ with $c_n(x) \ne 0$, already span the homogeneous polynomials of degree n in y, and taking the part of degree n is linear. (b) Relative error does not change when t is shifted by c and p is replaced by $e^{-c}p(\cdot + c)$, so let the interval be $[-W, W]$. On $[-W, 0]$, $0 < p(t) \le (1+\varepsilon)e^{t} \le 1 + \varepsilon$. The affine map that sends $[-W, 0]$ to $[-1, 1]$ sends W to 3, and the Chebyshev extremal property (Rivlin, 1990) states that a polynomial of degree n bounded by 1 in absolute value on $[-1, 1]$ is at most $C_n(3) = \cosh(n \operatorname{arccosh} 3)$ at 3, where C_n is the Chebyshev polynomial of degree n; hence $(1-\varepsilon)e^{W} \le p(H) \le (1+\varepsilon)\cosh(n \operatorname{arccosh} 3)$. (c) With features $e^{\omega^\top x - \|x\|^2/2}$ and $\omega \sim \mathcal{N}(0, I)$, one pair of features has mean $e^{x^\top y}$ and second moment $e^{2\|x+y\|^2 - \|x\|^2 - \|y\|^2} = e^{2x^\top y + \|x+y\|^2}$, and averaging D independent pairs divides the variance by D; the inequality is $\|x+y\|^2 - 4x^\top y = \|x-y\|^2 \ge 0$. (d) $\Gamma_i = \sum_j p_a(j)\,|g_{ij} - e^{t_{ij}}| \le \varepsilon \sum_j p_a(j)\, e^{t_{ij}} = \varepsilon M_i$. Parts (a) and (b) together bound the coordinates of any exact polynomial realization that holds the guarantee on an interval; the smallest degree for a given width follows from the minimax relative error (Table 11).

At the observed ranges. On the test windows (rows 1 to 31, local $r = 8$), the interaction t of a row spans an interval of median width 8.4 on Llama and 9.4 on Qwen, and that of a window, over all its rows and keys, a median width of 14.3 and 16.5. Features fixed before the queries do not know which values of t the queries of a block will produce, so we ask the guarantee to hold on the whole interval they span; only its width matters, because relative error does not change under a shift of t. One polynomial that keeps $|p(t)e^{-t} - 1| \le .1$ on the whole range of t of the median window needs degree at least 9 and 10, and the range of the median row alone needs 6 in both models (Table 11; part (b) alone gives 5 for the median window and 3 for the median row, in both models). By part (a), an exact separable realization of such a polynomial needs at least 11,440 and 19,448 coordinates for the median window. A polynomial fitted only to the values of t that occur could make do with a lower degree; the numbers concern the uniform guarantee. Counting only the multiply-adds that accumulate and read a dense state, a block state of D coordinates costs $ND(d_v + 1)$ to build over N history keys, more than the $LN(d + d_v)$ of exact attention for the $L = 32$ queries of the block once $D > L(d + d_v)/(d_v + 1) \approx 63.5$; a state shared by the whole prefix costs about $2D(d_v + 1)$ per token to update and read, against $N(d + d_v)$ for exact attention, so it pays off only while $D < N(d + d_v)/(2(d_v + 1)) \approx .99N$, for histories of 1,792 to 16,352 keys on our windows. For positive random features, part (c) applies to the keys with $t > 5$, which hold 6.1% and 15.6% of the dense attention on average (Appendix C.1): a relative root-mean-square error of 10% on each of them needs $D \ge 100\,(e^{20} - 1) \approx 4.9 \times 10^{10}$ i.i.d. features.

Degree and weighted error. The Taylor polynomials we ran hold such a guarantee on few rows: T_8 keeps 10% on the whole range of only 19.5% and 19.2% of rows, and $T_2 = P_2$ on fewer than 1% (Table 11). The median of their weighted kernel error still falls with the degree: against the diagnostic, the median over rows of the bound of Proposition 1 falls from .336 and .360 for P_2 to .131 and .147 for T_4, .033 and .040 for T_6 and .006 and .008 for T_8, while the 90th percentile stays between .57 and 1.47. From T_4 on, the median is below the .2 that a 10% guarantee would give through part (d) with $h_i \le 1$; this compares medians and gives these polynomials no such guarantee. It is possible

Table 11: **What a pointwise guarantee costs, and the weighted error of the Taylor polynomials we ran.** Coordinates: the lower bound $\binom{7+n}{n}$ of Remark 2(a) at $r = 8$ for an exact realization of a polynomial of degree n in t. Widest interval: an upper estimate of the largest width of an interval on which some polynomial of degree n keeps $|p(t)e^{-t} - 1| \le .1$ for every t, from Remark 2(b) and from the minimax error on 1001 Chebyshev points in 70-digit arithmetic; a polynomial that meets the bound on the whole interval, checked at its endpoints and at all its stationary points, exists for a width smaller by at most .0038. Relative error does not change under a shift of t, so only the width matters. Share: rows 1 to 31 and windows of the test split (Llama/Qwen) whose range of t is at most that wide, an upper estimate of the share on which some polynomial of degree n could hold the guarantee on the whole range, and rows on whose whole range the Taylor polynomial T_n holds it. Last column: median over rows 1 to 31 of the bound $2h_i\Gamma_i/M_i$ of Proposition 1 for T_n against the diagnostic (t_{ij} in place of $u_i^\top w_j$), with the 10th and 90th percentiles, from the runs of Appendix C.1 ($T_2 = P_2$).

		Widest interval		Share (%) where $\le 10\%$ can or does hold			
n	Coordinates	(b)	minimax	rows, best p	windows, best p	rows, T_n	$2h_i\Gamma_i/M_i$ of T_n: Llama; Qwen
2	36	6.1	2.76	3.9/3.2	0.0/0.0	0.5/0.8	.336 [.023, 1.364]; .360 [.040, 1.387]
3	120	9.6	4.40	14.6/11.7	0.0/0.0	1.5/1.6	–
4	330	13.1	6.12	30.0/24.7	0.6/1.3	3.4/3.3	.131 [.002, 1.471]; .147 [.004, 1.264]
5	792	16.6	7.88	45.7/38.4	5.1/6.3	6.3/6.1	–
6	1,716	20.2	9.68	59.8/51.7	15.7/11.8	10.1/9.9	.033 [< .001, 1.260]; .040 [< .001, .957]
7	3,432	23.7	11.50	71.5/63.6	27.5/19.2	14.6/14.4	–
8	6,435	27.2	13.34	80.7/73.7	42.6/29.7	19.5/19.2	.006 [< .001, .749]; .008 [< .001, .565]
9	11,440	30.7	15.20	87.6/81.7	57.2/40.3	25.0/24.3	–
10	19,448	34.3	17.06	92.1/87.7	69.9/53.4	30.6/29.7	–
12	50,388	41.3	20.82	97.4/94.9	85.4/73.4	42.2/40.7	–
16	245,157	55.4	28.42	99.9/99.2	98.7/93.1	63.8/61.6	–

because $\Gamma_i/M_i = \sum_j \left(p_a(j)\, e^{t_{ij}}/M_i\right) |g_{ij}e^{-t_{ij}} - 1|$ averages the relative error with the diagnostic's attention as weights, whereas a pointwise guarantee bounds its largest value. The error against dense attention also carries the omitted directions and is reported in Tables 5 and 7.

Scope. The remark covers exact realizations of a polynomial in t on open sets of queries and keys, guarantees that hold uniformly on an interval, and the variance of i.i.d. positive random features; it does not cover orthogonal random features, which we ran in Tables 5 and 7, other positive features, or realizations that only approximate a polynomial. Features built for the N keys of one history need not obey part (a): its argument without the scaling step gives D at least the dimension of the span of the functions $x \mapsto p(x^\top y_j)$, which is at most N, and a state with a coordinate per key is no compression. Once the $L = 32$ queries of a block are known, its $L \times N$ matrix of kernel values has rank at most 32, so features computed from those queries can reproduce it with 32 coordinates, at the cost of evaluating the kernel at every key, as exact attention does. Finally, a small output error does not require a pointwise guarantee (Proposition 1): the representation of Section 5 computes the keys of highest anchor mass exactly and asks for no guarantee on the other keys, and with a P_2 remainder on those keys, which holds no pointwise guarantee either, its mean L2/Vmax error against dense attention at $m = N/128$ is .0228 and .0290 (Tables 6 and 8).

D Geometry Supported by a Small Key Set

D.1 Full 64-layer profile

Figure 10 presents the profile of all 64 layers ordered by network depth, from the first layer to the last, and Table 12 lists representative layers.

Across both architectures, the two errors rise and fall together layer by layer (Spearman $\rho = .94$ over the 64 layer medians), while ESS alone does not rank the layers:

- **Llama**: ESS peaks in the middle of the network (19.7 at layer 12), and so do both errors (.0642 and .0895); sharply concentrated layers such as layer 2 (ESS = 1.2) and layer 16 (ESS = 1.1) have errors between .002 and .004.
- **Qwen**: the shallowest layers are the most diffuse (ESS between 29 and 129 in layers 0–4, 129.2 at layer 2), yet their errors are moderate (.0207 and .0283 at layer 2); both errors peak at layer 20 (ESS = 14.6; .0560 and .0493).

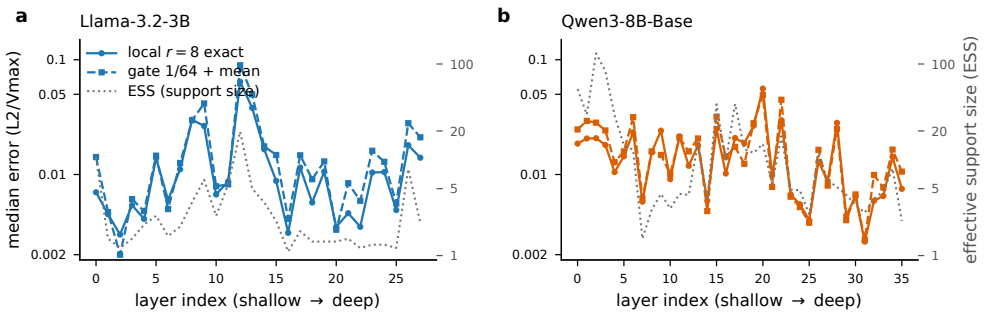

Figure 10: **Architecture-wide attention geometry across network depth (all 64 layers across both models).** Left axis: Median L2/Vmax error for the local $r = 8$ exact diagnostic (solid) and the anchor-gated representation with $m = N/64$ and a zero-order remainder (dashed). Right axis: Effective support size (ESS, dotted grey). (a) Llama-3.2-3B (28 layers). (b) Qwen3-8B-Base (36 layers). In both models the two errors rise and fall together across depth. Llama is most diffuse in its middle layers (ESS = 19.7 at layer 12); Qwen is most diffuse in its first few layers (ESS up to 129), where its errors are nonetheless moderate, and its errors peak at layer 20.

Table 12: **Representative layer statistics across network depth.** Median L2/Vmax errors for the local $r = 8$ exact diagnostic, the local $r = 8$ P_2 summary, and the anchor-gated representation at $m = N/64$ with zero-order remainder, alongside the effective support size (ESS).

Model	Layer	ESS	Local $r = 8$ (Exact)	Local $r = 8$ (P_2)	Gate $1/64$ + Rem.	Gate $1/64$ (Drop)
Llama-3.2-3B	0	10.4	.0070	.0178	.0142	.0258
	2	1.2	.0030	.0768	.0020	.0039
	9	6.2	.0265	.0620	.0416	.0543
	12	19.7	.0642	.1097	.0895	.1055
	16	1.1	.0031	.0133	.0041	.0076
	27	2.2	.0140	.0480	.0211	.0253
Qwen3-8B-Base	0	54.7	.0186	.0280	.0247	.0337
	2	129.2	.0207	.0305	.0283	.0525
	15	38.0	.0249	.0389	.0319	.0398
	20	14.6	.0560	.1261	.0493	.0593
	31	2.8	.0026	.0316	.0027	.0035
	35	2.3	.0075	.0234	.0106	.0273

The few-direction diagnostic and the anchor-gated representation therefore have high and low error on the same layers, even where ESS is a poor summary of the tilted distribution. As in Section 4, where holding window statistics and the layer fixed leaves a partial rank correlation of .81 in both models, this is a correlation, and what the two errors share beyond those statistics is not identified.

D.2 Restricted basis: definitions, null control and subspace distance

Definitions. The restricted basis $U_8^{\mathcal{H}}$ of Section 4 holds the leading eight eigenvectors of $J_{\mathcal{H}}$ (Appendix B.2), and $J_{\mathcal{O}} = J - J_{\mathcal{H}}$ is the positive semidefinite part over the keys outside $\mathcal{H}$. For a key set S, the mass share is $M_S = \sum_{j \in S} p_a(j)$ and the energy share E_S is Equation (6) with S in place of $\mathcal{H}$; for $S = \mathcal{H}$, $M = 1 - T$. Let $\lambda_1 \geq \lambda_2 \geq \dots$ be the eigenvalues of J and $\gamma_8 = \lambda_8 - \lambda_9$. The cosines of the principal angles $\theta_1 \leq \dots \leq \theta_8$ between the column spaces of U_8 and $U_8^{\mathcal{H}}$ are the singular values of $U_8^\top U_8^{\mathcal{H}}$; $\|\sin\Theta\|_2 = \sin\theta_8 = \|U_8 U_8^\top - U_8^{\mathcal{H}} U_8^{\mathcal{H}\top}\|_2$ and $\|\sin\Theta\|_F = (\sum_k \sin^2\theta_k)^{1/2} \leq \sqrt{8}$. The variance captured by $U_8^{\mathcal{H}}$ is $\mathrm{tr}(U_8^{\mathcal{H}\top} J\, U_8^{\mathcal{H}}) / \sum_{i \leq 8} \lambda_i$, which is at most one and equals one exactly when $U_8^{\mathcal{H}}$ spans a leading eight-dimensional eigenspace of J. Every control basis is built like the restricted basis from a different key set S, as the leading eigenvectors of $\sum_{j \in S} p_a(j)\, w_j w_j^\top$ with the full-history center $\bar{k}$ and without renormalization. The sets are all keys outside $\mathcal{H}$; m keys drawn uniformly without replacement; the keys ranked $m+1$ to $2m$ by p_a; m keys drawn without replacement with probability proportional to p_a; the m keys with the largest energy $p_a(j)\|U_8^\top w_j\|_2^2$, which maximize E_S over all sets of m keys for the fixed U_8 (leverage); the m keys with the largest $\|U_8^\top w_j\|_2$, a ranking that ignores the mass (geometry); and keys drawn with probability proportional to p_a without replacement until their mass reaches M (mass-matched). Random sets use three seeds, averaged within each window. The self-centered

variant recenters $J_{\mathcal{H}}$ at $\bar{k}_{\mathcal{H}}$, which drops the rank-one term of Appendix B.2. Spectral quantities are computed in float64 from the same inputs as the diagnostic errors, which are cluster means with 95% cluster-bootstrap intervals as in Appendix A; geometric quantities are medians over windows. Statements in Section 4 are decided on the test split at $m = N/128$; we fixed this budget, this split and every threshold below before running the analysis, and for every control the development split gives paired differences of the same sign.

Energy against mass. Across development windows the energy share rises with the mass share (Spearman ρ of .47 and .74 at $m = N/128$) but stays below it in all but the lowest-mass decile (Figure 11(a)); at $m = N/16$ it stays below the mass share in every decile. Equality of the two shares is not a null model for weighted covariance. The cluster mean of $E - M$ over development windows at $m = N/128$ is $-.168$ $[-.177, -.159]$ and $-.105$ $[-.110, -.101]$ on Llama and Qwen; with four sequences per task on this split, such intervals are descriptive.

Shuffled reference. An energy share equal to the mass share is not what the weighting alone produces. For the shuffled reference we permute the anchor weights p_a among the history keys of a window (three seeds), recompute $\bar{k}$, J and U_8, take the top m keys under the permuted weights and compute their energy share; their mass share equals the observed one exactly. On the test split at $m = N/128$, the cluster mean of E/M is .776 [.766, .785] on Llama and .968 [.965, .970] on Qwen under the shuffle, both below one, against .816 and .886 observed. The paired difference, observed minus shuffled, is $+.040$ $[+.036, +.043]$ on Llama and $-.081$ $[-.085, -.078]$ on Qwen, with the same signs and intervals excluding zero on the development split ($+.045$ and $-.079$) and at $m = N/64$ and $N/16$. At $m = N/128$ on the test split both exceed in size the margin of .03 that we fixed in advance for reading a difference as more than the weighting itself; at $m = N/64$ and $N/16$ the Llama difference falls below it ($+.029$ and $+.011$). On Llama, therefore, the high-mass keys carry more direction energy per unit of mass than randomly placed weights, although their energy share stays below their mass share; on Qwen they lie closer to the tilted mean than the same weights placed on randomly chosen keys. In neither model does the cluster mean of E/M reach one.

Subspace distance. Figure 11(b)–(d) repeats the comparison of Section 4 for $r \in \{1, 2, 4, 8, 16\}$ at the three budgets. The leading direction of the two bases nearly coincides at every budget (median angle at most 5.4° at $r = 1$), while the largest principal angle grows with r: at $m = N/128$ its median rises from 4.8° and 5.4° at $r = 1$ to 61.6° and 54.5° at $r = 8$ and 77.7° and 74.5° at $r = 16$. The restricted basis nonetheless captures a median of at least 97.8% of the top-r variance at every r and budget, and its error exceeds that of the full-history basis at the same r by $+.0003$ to $+.0027$. A larger exact set brings the subspaces closer: at $m = N/16$ and $r = 8$ the largest angle is below 30° in 67% and 76% of test windows. At $m = N/128$ and $r = 8$, the largest angle between $U_8^{\mathcal{H}}$ and the leading twelve full-history directions, which measures how far the restricted subspace leaves a slightly larger full-history one, has median 25.7° and 20.1° and is below 30° in 63% and 76% of test windows. At $r = 16$ and $m = N/128$, windows with $m < 16$ are excluded (168 of 2,240 and 216 of 2,880).

Attention distributions. The bases differ in the distributions they induce. Between the restricted and full-history exact-exponential diagnostics, the median history-conditional total variation over non-anchor test rows is .022 on Llama and .053 on Qwen (90th percentiles .118 and .181), compared with .073 and .165 between the full-history diagnostic and dense attention. Diagnostic output error, captured variance, principal angles and distributional distance thus measure different properties; the shadow interpretation concerns useful local geometry, not equality of the complete distributions.

Equal-m controls and centering. Table 13 lists all controls. The sets that carry most of the anchor mass (leverage, drawn in proportion to p_a, mass-matched) stay within .001 of the restricted basis. The equal-m sets that carry little mass (uniform, next, geometry) fall between the global basis and the anchor-only baseline, and the basis from all keys outside $\mathcal{H}$, which carries a median 8% and 11% of the mass but 31% and 27% of the energy, has an error only slightly below that of the global basis. Recentering at $\bar{k}_{\mathcal{H}}$ raises the error by $+.0023$ and $+.0019$ at $m = N/128$ and by $+.0005$ and $+.0004$ at $m = N/16$; at $m = N/128$ the error of the self-centered basis exceeds that of the full-history basis by $+.0040$ in both models. The full-history center, which one anchor row provides, thus accounts for part of the accuracy of the restricted basis, but the self-centered basis keeps most of the lead over the global basis (.0184 and .0179 of .0208 and .0198).

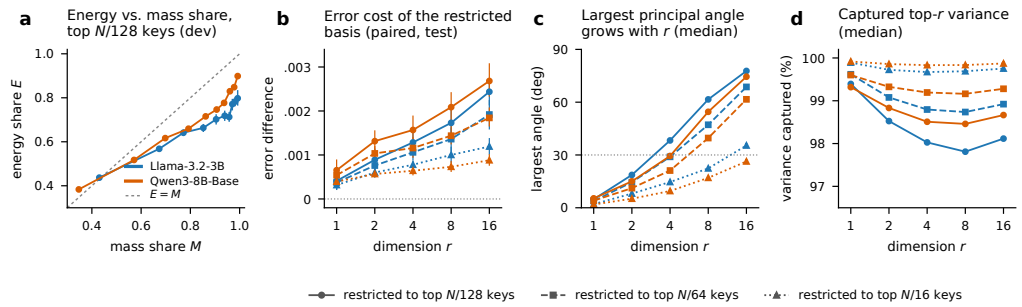

Figure 11: **Restricted basis: energy against mass, and dependence on the dimension r.** (a) Mean energy share against mean mass share of the top $N/128$ keys in each mass decile of the development windows (224 and 288 windows per decile), with 95% cluster-bootstrap intervals for the energy share; the dashed line marks equal shares. (b)–(d) Restricted basis from the top $N/128$ (solid), $N/64$ (dashed) and $N/16$ (dotted) keys against the full-history local basis of the same dimension r, test split: (b) paired difference in the exact-exponential diagnostic error with 95% cluster-bootstrap intervals; (c) median largest principal angle, dotted line at 30°; (d) median share of the top-r variance of J that the restricted basis captures. At $r = 16$ and $m = N/128$, windows with $m < 16$ are excluded.

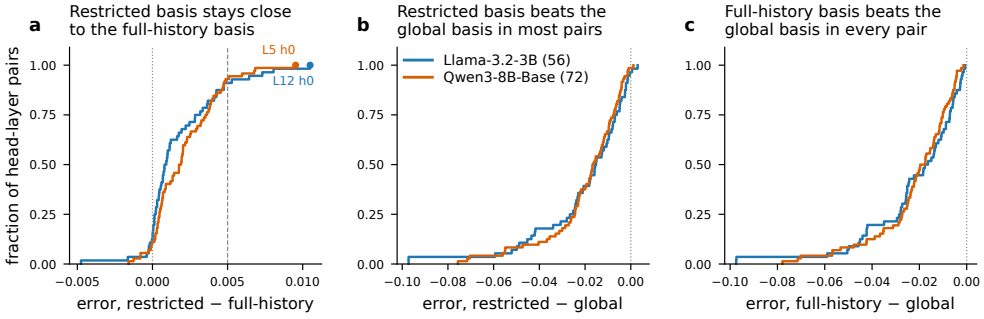

Figure 12: **Restricted, full-history and global bases per head-layer pair.** Empirical distribution functions over the 56 (Llama-3.2-3B) and 72 (Qwen3-8B-Base) captured head-layer pairs of the per-pair test error difference (L2/$V_{\max}$, windows averaged within cluster, clusters and tasks equally weighted as in Appendix A): (a) restricted basis from the top $N/128$ keys minus the full-history local basis, dashed line at .005, markers at the largest gap per model; (b) restricted minus global basis; (c) full-history minus global basis.

Head-layer pairs. Figure 12 shows the per-pair differences over the 128 fixed head-layer pairs, and Table 14 lists the pairs where the restricted basis loses most. The pairs with the largest gaps are the same on both splits (Spearman ρ of .94 and .92 between the development and test gaps at $m = N/128$; 9 of the 10 test pairs beyond .005 also exceed it on the development split). Head 0 and head $H_q/2$ do not differ systematically: head 0 has the larger gap in 14 of 28 and 16 of 36 layers. The 128 pairs are two fixed query heads per layer, 7% of the query heads of the two models (Appendix A). Appendix D.3 repeats the comparison on every query head.

D.3 All query heads

Protocol. The operator-level results use two query heads per layer, fixed by index (Appendix A). To test whether they are representative, we repeated the capture with every query head of every layer selected, 672 head-layer pairs on Llama-3.2-3B and 1,152 on Qwen3-8B-Base, on the same inputs and windows. For each pair the matrix R and the global basis were refitted on the train split, and the pair was analysed on its 40 test windows, and on its 40 development windows as a check, with the arithmetic of the fixed heads batched over heads. In this run the query, key and value tensors of heads 0 and $H_q/2$ were bit-identical to the archived captures for all 40 sequences and all layers of both models, the archived analysis code applied to them reproduced every archived window value exactly, and the batched analyser matched the archived window values to within 0.3% (values of at

Table 13: **Restricted basis and equal-m controls** (test split, $m = N/128$, $r = 8$). Every basis is built like the restricted basis from the key set in the first column (center $\bar{k}$, weights p_a, no renormalization) except the self-centered, global and anchor-only rows. Error: exact-exponential diagnostic, cluster mean; Δ: paired difference from the restricted basis with 95% cluster-bootstrap interval; M, E: median mass and energy share of the set; angle: median largest principal angle to the full-history basis U_8, in degrees. Size is the median number of keys over m. Leverage ranks keys by their energy $p_a(j)\|U_8^\top w_j\|_2^2$; mass-matched sets are drawn in proportion to p_a until they reach the mass M of $\mathcal{H}$ (Appendix D.2). Random sets are averaged over three seeds within each window.

Key set	Size	Error	Δ	M	E	Angle
Llama-3.2-3B						
Full history	–	.0205	$-.0017$ $[-.0020, -.0014]$	1	1	0
Top m by p_a (restricted)	1	.0223	–	.924	.688	61.6
Restricted, centered at $\bar{k}_{\mathcal{H}}$	1	.0246	$+.0023$ $[+.0021, +.0025]$	.924	–	64.0
Top m by energy (leverage)	1	.0221	$-.0002$ $[-.0003, -.0001]$	.917	.702	59.7
m drawn $\propto p_a$	1	.0225	$+.0003$ $[+.0001, +.0004]$	.896	.572	71.1
Mass-matched, drawn $\propto p_a$	2.7	.0218	$-.0005$ $[-.0006, -.0004]$	.924	.692	56.3
All keys outside $\mathcal{H}$	127	.0416	$+.0193$ $[+.0175, +.0211]$	.076	.312	64.5
Ranks $m{+}1$ to $2m$ by p_a	1	.0476	$+.0254$ $[+.0231, +.0275]$	.016	.058	77.7
m drawn uniformly	1	.0511	$+.0289$ $[+.0268, +.0308]$	.001	.004	82.6
Top m by $\|U_8^\top w_j\|_2$	1	.0502	$+.0279$ $[+.0259, +.0298]$	.000	.005	84.4
Global basis	–	.0430	$+.0208$ $[+.0184, +.0233]$	–	–	–
Anchor only ($r = 0$)	–	.0903	–	–	–	–
Qwen3-8B-Base						
Full history	–	.0268	$-.0021$ $[-.0024, -.0018]$	1	1	0
Top m by p_a (restricted)	1	.0289	–	.893	.729	54.5
Restricted, centered at $\bar{k}_{\mathcal{H}}$	1	.0308	$+.0019$ $[+.0017, +.0020]$	.893	–	58.8
Top m by energy (leverage)	1	.0285	$-.0004$ $[-.0005, -.0003]$	.885	.741	52.3
m drawn $\propto p_a$	1	.0295	$+.0006$ $[+.0005, +.0008]$	.847	.626	66.4
Mass-matched, drawn $\propto p_a$	2.2	.0284	$-.0005$ $[-.0007, -.0004]$	.895	.733	49.9
All keys outside $\mathcal{H}$	127	.0465	$+.0176$ $[+.0160, +.0192]$	.107	.271	71.1
Ranks $m{+}1$ to $2m$ by p_a	1	.0527	$+.0238$ $[+.0222, +.0255]$	.030	.061	80.0
m drawn uniformly	1	.0549	$+.0260$ $[+.0247, +.0276]$	.002	.004	83.6
Top m by $\|U_8^\top w_j\|_2$	1	.0564	$+.0275$ $[+.0259, +.0289]$	.000	.001	85.2
Global basis	–	.0487	$+.0198$ $[+.0177, +.0221]$	–	–	–
Anchor only ($r = 0$)	–	.0899	–	–	–	–

least 10^{-3}; TF32 disabled); the refitted global bases of the two fixed heads lie within 10^{-5} degrees of the archived ones. The shares of the fixed heads below therefore equal those of Figure 12.

Comparison with the fixed heads. We compare the distribution of restricted-minus-full and restricted-minus-global diagnostic errors over all head-layer pairs with the two fixed heads per layer. Write Δ_{full} and Δ_{glob} for restricted-minus-full and restricted-minus-global diagnostic error. The same training/test split, $m = N/128$ support budget, $r = 8$ dimension, weighting and error metric are retained. The full-head sweep broadens coverage of heads, not the number of independent input instances. We report both the typical behavior and the fractions outside the one-sided criterion $\Delta_{\text{full}} \leq .003$ (not an absolute-difference tolerance).

Result. The all-head distributions closely track those of the fixed heads (Table 15): over all pairs the medians of the two differences lie within .0018 of those of the fixed heads, and the median energy and mass shares within .031. The restricted-minus-full diagnostic error is at most .003 in 74% and 72% of pairs (75% and 69% for the fixed heads) and beats the global basis in 97% and 96% (96% and 99%); 73% and 71% of pairs meet both conditions. On the development split the shares with restricted-minus-full error at most .003 are 69% and 67% (68% and 67% for the fixed heads) and 92% and 94% ahead of the global basis (93% and 99%), and the per-pair gaps of the two splits have Spearman ρ of .91 and .88. Figure 13(a) and (b) show the distributions over pairs, and (d) and (e) every pair layer by layer with the fixed heads marked.

Where the restricted basis loses. The restricted basis loses more than .003 in 26% and 28% of pairs and more than .005 in 98 and 174 pairs, and the pairs that miss either condition have a median

Table 14: **Head-layer pairs where the restricted basis loses most** (test split, $m = N/128$, $r = 8$; per-pair error averaged as in Appendix A). Top: the pairs whose restricted-basis error exceeds the full-history error by more than .005. Bottom: the pairs whose restricted basis is worse than the global basis. ESS is the median over the pair's 40 test windows; kv is the key-value head the query head reads.

Model	Layer	Head	kv	ESS	Restricted − full	Restricted − global
Restricted − full-history $>$ .005						
Llama-3.2-3B	12	0	0	22.6	+.0105	−.0150
Llama-3.2-3B	12	12	4	14.4	+.0081	−.0340
Llama-3.2-3B	13	12	4	5.8	+.0073	−.0178
Llama-3.2-3B	9	0	0	15.0	+.0064	−.0159
Llama-3.2-3B	23	0	0	3.4	+.0053	−.0233
Qwen3-8B-Base	5	0	0	26.5	+.0095	−.0127
Qwen3-8B-Base	20	16	4	19.2	+.0068	−.0355
Qwen3-8B-Base	17	16	4	42.0	+.0068	−.0243
Qwen3-8B-Base	15	16	4	20.9	+.0060	−.0329
Qwen3-8B-Base	22	0	0	11.9	+.0051	−.0300
Restricted worse than global						
Llama-3.2-3B	0	12	4	48.3	+.0047	+.0030
Llama-3.2-3B	25	12	4	1.2	+.0011	+.0005
Qwen3-8B-Base	15	0	0	121.0	+.0026	+.0012

ESS of 5.9 and 15.1, against 2.1 and 5.1 for the others. Across strata of ESS (Figure 13(c)), the share of pairs that meet both conditions falls from .92 and .82 at ESS < 2 to .24 and .64 at $8 \leq$ ESS < 16, while the median diagnostic error rises from .0073 and .0115 to .0461 and .0293 and the median gate error at $m = N/64$ from .0081 and .0106 to .0591 and .0335. The exceptions are therefore heads with diffuse anchor distributions, where both the few-direction and the few-key approximations are least accurate. Spearman ρ between the gap and $\log$ ESS is .59 and .36, but relative to the full-history error the gap is nearly unrelated to ESS (.06 and .11). At least half of the heads miss a condition in layers 8–10 and 12–13 of Llama-3.2-3B and 1–2, 17 and 20–21 of Qwen3-8B-Base. The restricted basis is worse than the global basis in 21 and 49 pairs; 11 of the Llama pairs lie in layer 0, and the two largest gaps, $+.048$ and $+.041$, belong to heads 11 and 9 of layer 15 of Qwen3-8B-Base. These pairs bound the scope of Section 4: the restricted basis tracks the full-history basis where the anchor distribution is concentrated, which is the case in most heads but not in all.

Energy, error pairing, subspace and captured mass. Over all pairs, the per-pair window median of the energy share of the top $N/128$ keys has median 69% and 73% and that of the mass share 88% and 89% (Table 15 averages the windows of a pair instead and gives .675 and .713, and .852 and .849), and the energy share stays below the mass share in 98% of pairs in both models, as in Appendix D.2. The Spearman correlation between the gate error at $m = N/64$ and the local $r = 8$ diagnostic error, .895 and .861 for the fixed heads (Table 16), is .905 [.899, .911] and .888 [.880, .895] over the 26,880 and 46,080 test windows of all pairs, and .961 and .954 between the per-pair errors. As for the fixed heads, the two bases agree in error rather than in subspace: the largest principal angle between the restricted and the full-history basis is below 30° in 9% and 17% of test windows, against 9% and 19% for the fixed heads. Finally, the ratio between the history mass that the anchor's top $N/128$ keys capture for a query and the mass that the query's own top $N/128$ keys hold, taken as a median over the queries of a window and then over windows, is at least 0.8 in 91% and 78% of pairs, against 96% and 81% for the fixed heads, but in only 13% and 47% once the first four tokens are excluded; those tokens hold a median .70 and .42 of the history mass, and more than half of it in 78% and 37% of pairs, so the weight of the attention sink on Llama-3.2-3B (Appendix F) is not specific to the fixed heads.

What the controls do and do not isolate. The mass-matched random sets use 2.690 and 2.217 times as many keys as the top-m set and retain 69.3% and 76.1% of its members. Their similar diagnostic performance therefore does not establish invariance to key identity at fixed cardinality. Equal-m sampling proportional to p_a is a closer control, but it too overlaps the high-mass set. Conversely, the full-tail basis captures .931 and .919 of the optimal rank-eight variance yet has error .0416 and .0465, well above the restricted basis. High captured variance alone is not a sufficient

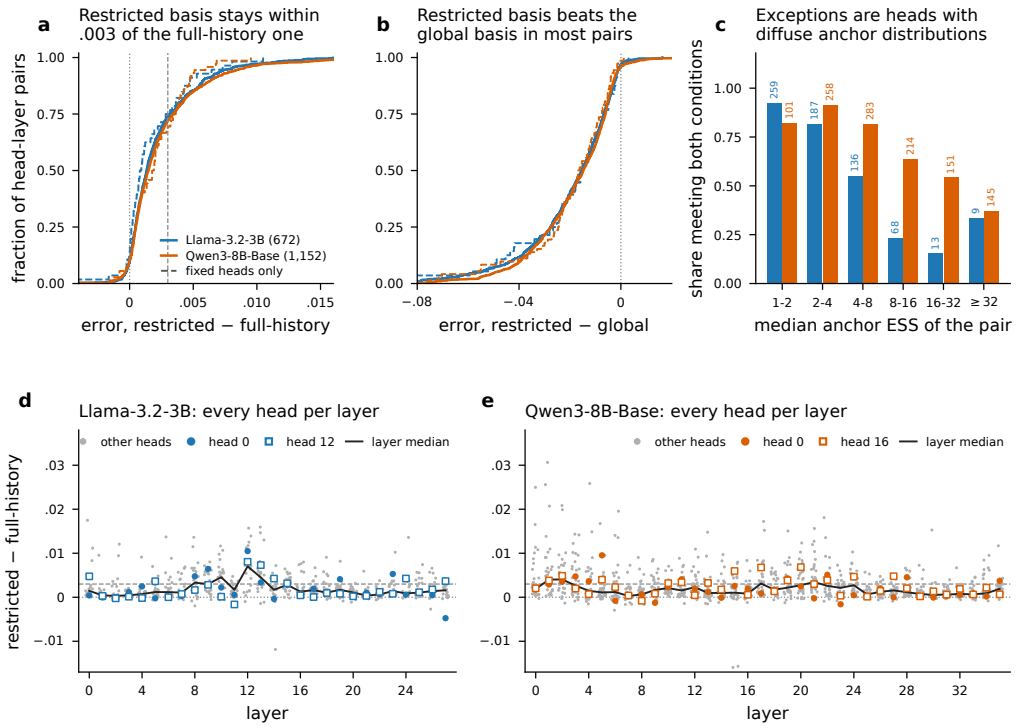

Figure 13: **All 1,824 head-layer pairs against the two fixed heads** (test split, $m = N/128$, $r = 8$; per-pair errors averaged as in Appendix A). (a) Empirical distribution function of the diagnostic error of the restricted basis minus that of the full-history basis, over all pairs (solid) and over heads 0 and $H_q/2$ of every layer in the same run (dashed); dashed line at .003. (b) The same for the restricted minus the global basis. 3 and 14 pairs lie outside the axis in (a) and 7 and 6 in (b). (c) Share of pairs that meet both conditions ($\Delta_{\text{full}} \leq .003$ and $\Delta_{\text{glob}} < 0$) by the median anchor ESS of the pair over its 40 test windows; the numbers give the pairs in each stratum. (d), (e) Restricted minus full-history basis for every head of every layer (grey, jittered), the layer median (black) and the two fixed heads (colored).

predictor of output error. The restricted construction also uses the full-history center and the training-fitted matrix R, and its diagnostic reads every history key. It identifies a geometric support, not an online method that constructs or evaluates the diagnostic by reading only $N/128$ keys.

E Covariation Beyond Measured Concentration and Difficulty

We pair the local rank-eight exact-exponential diagnostic error with the one-anchor zero-order gate at $m = N/64$, on the same 2,240 Llama and 2,880 Qwen test windows. Window count is not the number of independent samples: uncertainty resamples the 20 task–item clusters, preserving all windows, layers and heads within each cluster. Concentration covariates C are log effective support size, entropy and log anchor tail mass; difficulty covariates D describe mean query displacement, remainder-value dispersion and value scale. Partial Spearman correlation is the Pearson correlation between residuals after regressing the ranked errors on ranked continuous covariates and the stated fixed-effect indicators. This controls the measured linear rank associations, not arbitrary nonlinear confounding.

In Table 16, the residual correlation is .808 and .809 after concentration, difficulty and layer controls, and .718 and .758 after additionally controlling layer–head identity. Removing the shared $V_{\max}$ denominator leaves .815 and .799 after the same controls. Four reference pairings on the same windows (gate versus anchor-only, global-basis or local-P_2 error, and diagnostic versus per-query top-m drop error) have unadjusted correlations ranging from .286 to .709, below the target pairing. Thus neither the measured concentration summaries nor the common divisor readily explains the observed relationship.

Table 15: **All query heads against the two fixed heads** (test split, $m = N/128$, $r = 8$; per-pair values averaged as in Appendix A). All: the 672 (Llama-3.2-3B) and 1,152 (Qwen3-8B-Base) head-layer pairs; fixed: heads 0 and $H_q/2$ of every layer in the same run. Shares and medians are over pairs, with 95% cluster-bootstrap intervals (one resample of the 20 test items applied to every pair). Δ_{full}: restricted minus full-history basis; Δ_{glob}: restricted minus global basis; E, M: energy and mass shares of the top $N/128$ keys. The criterion column gives the thresholds fixed before the run (CI: the interval of the fixed heads; the row of M applies the margin set for E); each criterion holds in both models. The last row is the Spearman correlation between the gate error at $m = N/64$ and the diagnostic error over all test windows.

Statistic	All pairs	Fixed heads	Criterion
Llama-3.2-3B (672 and 56 pairs)			
Share with $\Delta_{\text{full}} \leq .003$	.738 [.692, .775]	.750 [.696, .804]	all $\geq$ fixed $-.10$
Share with $\Delta_{\text{glob}} < 0$	.969 [.951, .970]	.964 [.929, .964]	all $\geq$ fixed $-.05$
Median Δ_{full}	$+.0013$ [$+.0011, +.0015$]	$+.0009$ [$+.0007, +.0011$]	in CI or $\pm.001$
Median Δ_{glob}	$-.0155$ [$-.0172, -.0134$]	$-.0154$ [$-.0181, -.0124$]	in CI or $\pm.001$
Median E	.675 [.662, .686]	.665 [.646, .679]	in CI or $\pm.05$
Median M	.852 [.843, .860]	.883 [.875, .894]	in CI or $\pm.05$
Spearman, gate vs. diagnostic	.905 [.899, .911]	.895	–
Qwen3-8B-Base (1,152 and 72 pairs)			
Share with $\Delta_{\text{full}} \leq .003$	.722 [.693, .740]	.694 [.639, .764]	all $\geq$ fixed $-.10$
Share with $\Delta_{\text{glob}} < 0$	.957 [.947, .960]	.986 [.958, 1.000]	all $\geq$ fixed $-.05$
Median Δ_{full}	$+.0015$ [$+.0013, +.0016$]	$+.0018$ [$+.0013, +.0021$]	in CI or $\pm.001$
Median Δ_{glob}	$-.0147$ [$-.0161, -.0127$]	$-.0165$ [$-.0180, -.0134$]	in CI or $\pm.001$
Median E	.713 [.704, .720]	.691 [.679, .705]	in CI or $\pm.05$
Median M	.849 [.843, .855]	.851 [.842, .866]	in CI or $\pm.05$
Spearman, gate vs. diagnostic	.888 [.880, .895]	.861	–

Table 16: Error covariation on the test split. Correlations and 95% item-cluster bootstrap intervals; all comparisons use paired windows.

Adjustment or normalization	Llama-3.2-3B	Qwen3-8B-Base
Unadjusted	0.895 [0.887, 0.903]	0.861 [0.850, 0.871]
Concentration covariates C	0.831 [0.816, 0.843]	0.848 [0.837, 0.860]
Difficulty covariates D	0.915 [0.909, 0.920]	0.866 [0.853, 0.877]
Layer indicators	0.821 [0.805, 0.835]	0.811 [0.797, 0.824]
$C + D+$ layer indicators	0.808 [0.794, 0.821]	0.809 [0.793, 0.824]
$C + D+$ layer–head indicators	0.718 [0.696, 0.737]	0.758 [0.736, 0.778]
Unnormalized errors, $C + D+$ layer	0.815 [0.801, 0.826]	0.799 [0.787, 0.811]
Value-RMS normalization, $C + D+$ layer	0.805 [0.790, 0.817]	0.819 [0.804, 0.834]
Output-norm normalization, $C + D+$ layer	0.754 [0.738, 0.769]	0.782 [0.768, 0.796]

Both approximations still share the anchor, center, queries, values and evaluation reference. The result supports a structural association; it neither identifies a unique cause nor says that concentration alone controls both errors. In particular, it does not imply that sparse and low-dimensional terms cannot complement one another: Appendix G.1 directly tests that question under distinct resource constraints.

F Reuse Across Queries: Captured Mass and Support Drift

F.1 Error by distance from the anchor

Key selection for a block of L queries is guided by a single anchor query q_a, the first query of the block, and the operational question is whether the selection becomes stale as queries advance further from it. Every block length starts at the same query, so row i of a block, counted from the anchor as row 0, has the same anchor, exact set, in-block keys and error in every block that contains it. Averages over the quartiles of different block lengths therefore compare different rows, and we report the rows of the $L = 256$ blocks directly. Table 17 gives row 0 on its own and rows 1–31, 32–63, 64–127 and 128–255 of the blocks at the quarter position on both splits (1120 blocks for Llama-3.2-3B and 1440 for Qwen3-8B-Base on each split, from 20 sequences, four per task). Intervals are over sequences,

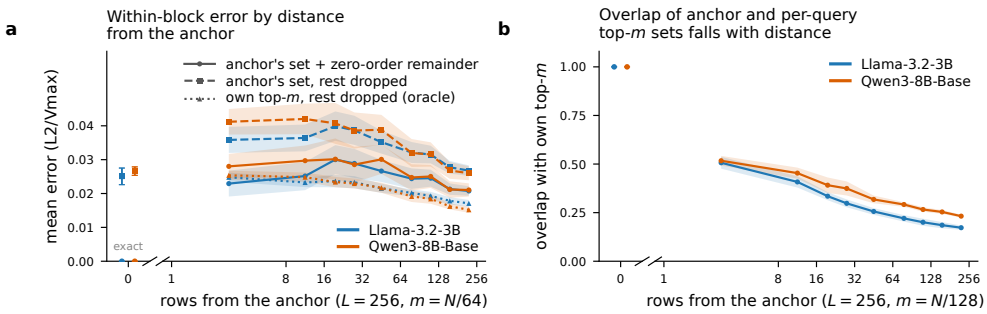

Figure 14: **Within-block error and key overlap by distance from the anchor (development split, 2560 blocks of $L = 256$).** Every block starts at the quarter position of its sequence. Row 0 is the anchor row, the first query of the block, and is shown on its own; the other points are means over rows 1–7, 8–15, 16–23, 24–31, 32–63, 64–95, 96–127, 128–191 and 192–255 with the five task families weighted equally, and bands and bars are 95% intervals over the 20 sequences (the union of the cluster-bootstrap and stratified t intervals). (a) Mean L2/Vmax error at $m = N/64$ with the anchor's exact set and the zero-order remainder (solid) or with the unselected keys dropped (dashed), and with each query's own top-m keys and the rest dropped (dotted; an oracle that needs the full scores of every query). On row 0 the zero-order remainder is exact by construction, because for the anchor query the remainder state reproduces the mass and mean of the unselected keys, and the anchor's set is that row's own top-m set, so the two drop errors coincide. With the unselected keys dropped, the error of the anchor's set is 1.44 times that of per-query selection on rows 1 to 7, 1.57 times on rows 8 to 15 and 1.57 times on rows 192 to 255 on Llama (1.62, 1.69 and 1.71 on Qwen). Both drop errors fall over the block, per-query selection by 31% (Llama) and 40% (Qwen) between rows 1 to 7 and rows 192 to 255, so the later decline is not specific to the anchor's set. Later rows place more attention on keys inside the block, which are computed exactly, so the decline does not by itself show that the anchor's keys remain the right ones; Table 17 separates this effect. (b) Mean overlap fraction $|\mathcal{H}_{\text{anchor}} \cap \mathcal{H}_i|/m$ between the anchor's top-m keys and each query's own top-m keys at $m = N/128$. It is 1 on row 0 by definition and falls from .51 (Llama) and .52 (Qwen) on rows 1 to 7 to .17 and .23 on rows 192 to 255, so the anchor's set is not a proxy for per-query selection.

with the task families weighted equally, and are the union of a cluster bootstrap and a stratified t interval with 15 degrees of freedom; Figure 14 plots the development rows in finer segments.

The error of the block-first anchor with the zero-order remainder falls from rows 1–31 to rows 128–255 on both splits: on the test split from .0252 to .0196 (difference $-.0056$ $[-.0077, -.0034]$) on Llama-3.2-3B and from .0272 to .0202 ($-.0070$ $[-.0106, -.0034]$) on Qwen3-8B-Base, on the development split from .0269 to .0210 and from .0291 to .0211. Over the same rows the in-block mass $1 - h_i$ grows from .16 to .25 and from .23 to .39 (development: .16 to .25 and .22 to .39). Keys inside the block are computed exactly, so a growing in-block mass lowers the full-output error whatever the exact set is. The history-conditional error $e_i^{\text{hist}} = \|\hat{o}_i^{\text{hist}} - o_i^{\text{hist}}\|_2 / V_{\max}$ removes this effect: both outputs are normalized over the history keys only, the exact one with the softmax over the history and the approximate one with the exact keys of $\mathcal{H}$ and the remainder state scaled by the per-query tilt factor. It rises from rows 1–31 to rows 128–255, by $+.0034$ $[+.0008, +.0061]$ and $+.0089$ $[+.0042, +.0136]$ on the test split and by $+.0053$ $[+.0013, +.0094]$ and $+.0084$ $[+.0032, +.0136]$ on the development split. This rise is the less robust finding of this appendix: at $m = N/128$ the test intervals are $[-.0016, +.0034]$ and $[+.0001, +.0099]$ (development $[-.0018, +.0070]$ and $[-.0003, +.0105]$), and by task it comes from the synthetic tasks, while on 2WikiMQA in both models and on HotpotQA in Qwen3-8B-Base the history-conditional error falls ($-.0050$; $-.0019$ and $-.0030$).

A fixed-row sweep separates the distance from the anchor from the position in the block. On rows 225 to 255 of the test blocks at the quarter position we place the anchor at row 224, 192, 128 or 0, or use the query just before the block, at $m = N/64$ with the zero-order remainder. The error is .0138, .0153, .0153, .0176 and .0186 on Llama-3.2-3B and .0146, .0161, .0168, .0179 and .0192 on Qwen3-8B-Base. Moving the anchor from row 224 to row 0 adds .0039 $[.0027, .0050]$ and .0033 $[.0022, .0044]$; the increase is monotone in the distance on Qwen3-8B-Base, while on Llama-3.2-3B anchors 32 and 96 rows away give the same error. On rows 33 to 63, moving the anchor from row 32 to row 0 raises the error from .0209 to .0252 and from .0243 to .0269 (differences .0043 $[.0027, .0059]$

Table 17: **Within-block error by distance from the anchor.** Rows of the $L = 256$ blocks at the quarter position, counted from the anchor (row 0, the block's first query), with the block-first anchor at $m = N/64$. Means with the five task families weighted equally and 95% intervals over the 20 sequences of each split (the union of the cluster-bootstrap and stratified t intervals). The drop error ratio divides the error of the anchor's set by that of each query's own top-m keys, an oracle that needs every query's full scores, with the unselected keys dropped in both. The history-conditional error e^{hist} normalizes the exact and the approximate output over the history keys only; $1 - h_i$ is the attention mass on keys inside the block. On row 0 the zero-order remainder is exact and the two sets coincide.

Quantity	Row 0	Rows 1–31	Rows 32–63	Rows 64–127	Rows 128–255
Llama-3.2-3B, test split					
error, zero-order	0	.0252 [.0226,.0278]	.0252 [.0230,.0273]	.0238 [.0215,.0261]	.0196 [.0184,.0208]
drop error ratio	1	1.56 [1.50,1.62]	1.59 [1.53,1.65]	1.59 [1.53,1.65]	1.53 [1.49,1.57]
e^{hist}, zero-order	0	.0350 [.0312,.0388]	.0406 [.0366,.0446]	.0411 [.0369,.0452]	.0384 [.0354,.0414]
in-block mass $1 - h_i$	.045 [.040,.051]	.164 [.160,.169]	.199 [.193,.205]	.219 [.212,.226]	.246 [.238,.253]
Llama-3.2-3B, development split					
error, zero-order	0	.0269 [.0235,.0302]	.0266 [.0234,.0299]	.0245 [.0221,.0268]	.0210 [.0196,.0224]
drop error ratio	1	1.59 [1.48,1.70]	1.63 [1.54,1.72]	1.61 [1.54,1.68]	1.56 [1.53,1.59]
e^{hist}, zero-order	0	.0361 [.0315,.0408]	.0423 [.0371,.0475]	.0424 [.0382,.0467]	.0415 [.0383,.0446]
in-block mass $1 - h_i$	.053 [.047,.059]	.161 [.155,.166]	.202 [.195,.209]	.223 [.217,.230]	.249 [.244,.253]
Qwen3-8B-Base, test split					
error, zero-order	0	.0272 [.0241,.0304]	.0270 [.0238,.0302]	.0243 [.0228,.0258]	.0202 [.0184,.0220]
drop error ratio	1	1.65 [1.59,1.71]	1.70 [1.63,1.77]	1.67 [1.62,1.72]	1.66 [1.59,1.73]
e^{hist}, zero-order	0	.0379 [.0337,.0420]	.0449 [.0402,.0497]	.0477 [.0442,.0513]	.0468 [.0429,.0506]
in-block mass $1 - h_i$	.049 [.043,.055]	.226 [.218,.234]	.294 [.280,.308]	.347 [.341,.354]	.388 [.376,.400]
Qwen3-8B-Base, development split					
error, zero-order	0	.0291 [.0254,.0329]	.0301 [.0258,.0344]	.0249 [.0214,.0284]	.0211 [.0194,.0228]
drop error ratio	1	1.69 [1.59,1.78]	1.79 [1.69,1.89]	1.69 [1.62,1.76]	1.68 [1.64,1.72]
e^{hist}, zero-order	0	.0403 [.0351,.0455]	.0499 [.0439,.0559]	.0478 [.0431,.0524]	.0487 [.0452,.0522]
in-block mass $1 - h_i$	.047 [.042,.053]	.220 [.210,.231]	.293 [.279,.307]	.345 [.330,.361]	.389 [.381,.398]

and .0027 [.0010, .0043]). The comparison of rows 224 and 0 gives .0024 [.0019, .0028] and .0090 [.0084, .0097] in the block before the question, and raises the error from .0218 to .0300 and from .0272 to .0358 in the final 256 rows. Within 256 rows the anchor's set therefore goes stale at every position we measured.

Each query's own top-m keys, an oracle that needs the full scores of every query, are available only with the unselected keys dropped, so we compare them with the anchor's set with the unselected keys dropped as well: the drop error ratio of Table 17 divides the two segment means, and the zero-order error of the anchor's set is never divided by a drop error. Before computing its intervals we fixed how to read it: as rising if the interval on rows 128–255 lies above that on rows 1–31, and as holding otherwise. It holds in both models and on both splits. On the test split it is 1.56 [1.50, 1.62] on rows 1–31 and 1.53 [1.49, 1.57] on rows 128–255 on Llama-3.2-3B, and 1.65 [1.59, 1.71] and 1.66 [1.59, 1.73] on Qwen3-8B-Base; on the development split it is 1.59 and 1.56, and 1.69 and 1.68, with rows 1–15 lower on Llama-3.2-3B (1.51) and the segments from row 16 on between 1.56 and 1.67 and between 1.68 and 1.79. With both outputs normalized over the history, the drop error ratio moves on the test split from 1.71 [1.63, 1.80] to 1.79 [1.73, 1.86] and from 1.79 [1.72, 1.86] to 1.89 [1.81, 1.97], again with overlapping intervals.

F.2 Position of the block in the sequence

Blocks at the quarter position contain neither the question nor the end of the prompt. The test sweep therefore adds, for blocks of 32 and 256 queries, the final block of each sequence, which on the synthetic tasks contains the start of the question (question tokens make up .10 of rows 128–255 of the final 256-row block, averaged over the five tasks), and, on the three synthetic tasks, the block that ends 37 to 50 tokens before the question starts. Figure 15(d) shows the error of 32-query blocks by quarter at $m = N/64$. At the quarter position the error is nearly flat on the test split (the ratio of the third to the first quarter is 1.09 [.99, 1.21] and 1.09 [.98, 1.22]), whereas on the development split it rises on Llama-3.2-3B, from .0201 to .0300 (1.50 [1.32, 1.70]), and by a factor of 1.23 [1.03, 1.44] on Qwen3-8B-Base (cluster-bootstrap interval; the stratified t interval, [.97, 1.48], includes 1), so

the two splits disagree at this position. Before the question the ratio is 1.31 [1.26, 1.36] and 1.16 [1.13, 1.20], and in the final block it is largest (1.63 and 1.78). The first quarter contains the anchor row, whose zero-order error is 0, so every ratio is $8/7$ times the ratio of the third quarter to rows 1–7; that ratio is .95 and .95 at the quarter position (1.31 and 1.08 on the development split), 1.15 [1.10, 1.19] and 1.02 [.99, 1.05] before the question and 1.42 and 1.56 in the final block, so before the question the error rises on Llama-3.2-3B but hardly on Qwen3-8B-Base. Over 256 rows the final block again stands out, while before the question the full-output error and the drop error ratio rise only on Qwen3-8B-Base. In the final block the drop error ratio grows from 1.58 [1.48, 1.68] on rows 1–31 to 1.84 [1.79, 1.88] on rows 128–255 on Llama-3.2-3B and from 1.90 [1.83, 1.98] to 2.30 [2.22, 2.38] on Qwen3-8B-Base, and even the full-output error rises, from .0208 to .0231 and from .0237 to .0276; before the question the ratio holds on Llama-3.2-3B (1.49 and 1.47) and grows on Qwen3-8B-Base (1.66 to 1.86). A selector that keeps the m keys with the largest per-key maximum over four rows of the block (rows 0, $\lfloor L/3 \rfloor$, $\lfloor 2L/3 \rfloor$ and $L-1$, with the remainder still tilted by the block's first query) lowers the error of the last three quarters of the final 32-query block by 41% [37, 45] and 42% [40, 45] at the same m, and over the final 256 rows its error does not rise (.0209 on rows 1–31 and .0197 on rows 128–255 on Llama-3.2-3B, .0232 and .0224 on Qwen3-8B-Base). It looks ahead within the block, like the four-anchor selector of Section 6; the end-to-end runs exercise every block, including the question onset, where single-anchor selection can lose accuracy (Table 1). The intervals of the block before the question rest on 12 sequences of three tasks, and the NIAH and VT items share their filler text there, so they are narrower than the number of sequences suggests. The NIAH and VT items also largely share their filler at the quarter position and in the final block (8-gram overlap between .83 and 1), so the synthetic part of the other intervals in this appendix also rests on fewer distinct texts than sequences.

F.3 Captured history mass, the attention sink and persistence

The overlap of two top-m sets (next subsection) does not measure how much attention the anchor's set captures. The direct measure is the share of each query's history mass inside the anchor's set, $1 - \pi_i^{\text{hist}}$ with π_i^{hist} the history-conditional remainder share of Equation (14). Before the sweep we fixed three conditions for calling the set block-shared, all at $m = N/128$ on the 32-query blocks at the quarter and final positions of the test split: a median captured share of at least .8; a mean captured share in the last quarter of the block at least .95 times that in the first; and a p_a-weighted persistence of at least .9 for the anchor's 32 highest-weight keys outside positions 0 to 3, where the persistence of a key is the share of the block's rows whose own top-$2m$ set holds it. Llama-3.2-3B meets the first two (.89; .79 against .80) but not the third (.64), and Qwen3-8B-Base meets none (.77; .66 against .73; .76). We therefore describe the set as capturing most, but not all, of each query's history mass. The captured share falls most over the first rows after the anchor (Figure 15(a, b)), and 13% and 27% of queries keep less than half of their history mass inside the set.

The first four positions of the sequence carry much of the captured mass. They hold a median .83 (Llama-3.2-3B) and .36 (Qwen3-8B-Base) of a query's history mass, and .97 and .55 of the mass that the anchor's set captures. Excluding them, the anchor's set captures a median .02 and .26 of the history mass, against .07 and .39 for each query's own top-m keys; the median ratio of the two is .57 and .83, against .98 and .94 with the first four positions included. With those positions included, the anchor's 32 highest-weight keys persist in a p_a-weighted .90 and .84 of the rows, and its single highest-weight key in .99 and .91; excluding them, the figures are .64 and .76, and .71 and .82. Unweighted (Figure 15(c)), 33% and 42% of these keys stay in the own top-$2m$ set of at least three quarters of the rows and 25% and 17% in fewer than a quarter. On Llama-3.2-3B the mass that one query's set holds for the rest of the block is therefore mostly an attention sink, and the other keys of the set serve the later rows much less well than their own top keys would.

F.4 Overlap with per-query selection

To test whether the anchor's keys coincide with what each query would choose, we measured the overlap fraction between the anchor's top-m keys and the top-m keys each query would have chosen independently under full dense scoring:

$$\text{Overlap}_i = \frac{|\mathcal{H}_{\text{anchor}} \cap \mathcal{H}_i|}{m}. \tag{19}$$

At $m = N/128$ the overlap is 1 on row 0 by definition and falls from .38 on rows 1–31 to .19 on rows 128–255 on Llama-3.2-3B and from .45 to .25 on Qwen3-8B-Base (test split; development .38 to .18

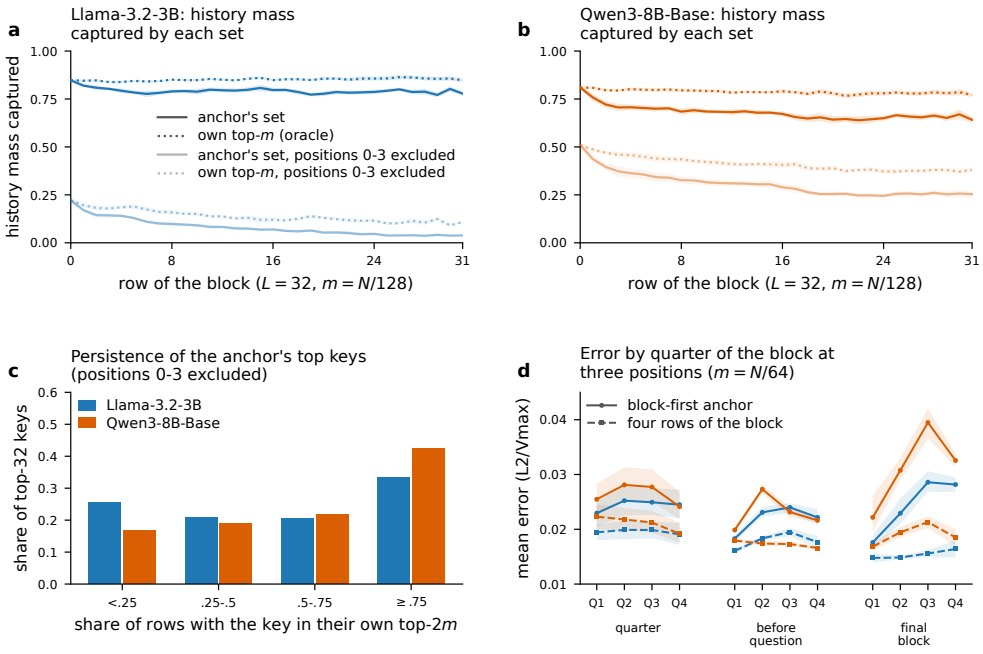

Figure 15: **Captured history mass, persistence of the anchor's keys, and within-block error at three positions (test split).** (a, b) Share of each query's history mass that falls inside the block-first anchor's top-m set (solid) and inside the query's own top-m set (dotted; an oracle that needs every query's full scores), by row of the 32-query blocks at the quarter and final positions, $m = N/128$; the light lines count only the history keys outside the first four positions of the sequence. Means with the five task families weighted equally; bands are 95% intervals over the 20 test sequences (the union of the cluster-bootstrap and stratified t intervals). On row 0 the two sets coincide. (c) Distribution of the persistence of the anchor's 32 highest-weight keys outside positions 0 to 3, the share of the block's rows whose own top-$2m$ set holds the key (same blocks and m). (d) Mean error with the zero-order remainder at $m = N/64$ by quarter of 32-query blocks at the quarter position, in the block that ends 37 to 50 tokens before the question (synthetic tasks only) and in the final block, for the block-first anchor (solid) and for the exact set chosen by the per-key maximum over rows 0, 10, 21 and 31 of the block (dashed); bands are cluster-bootstrap 95% intervals.

and .43 to .24); Figure 14(b) shows it row by row. The anchor's set is therefore not a proxy for each query's own selection, and a low overlap does not by itself raise the error, since keys with similar values can substitute for one another. On the development split, averaged over the 32 queries of a window, the median share of each query's history attention that falls outside the anchor's top $N/128$ keys is .132 on Llama and .288 on Qwen, against an anchor tail mass of .073 and .113 (Appendix H); the zero-order remainder carries this mass, and Appendix H analyzes the mass error that its drift causes. A strictly causal anchor, the query just before the block, leaves about the same mass outside its set as the block-first query in every row segment of Figure 14 after row 0: on the development split, at $m = N/128$ and $N/64$, the differences of the segment means are at most .0042 (Llama) and .0072 (Qwen) of the attention mass.

G Resource-Matched Comparisons and Cost Accounting

G.1 Low-dimensional remainder versus a larger exact set

For each test window we compare one- and four-anchor selection at seven exact-set budgets, $\{N/128, N/16, N/4, 8, 16, 32, 64\}$, and dimensions $r \in \{1, 2, 4, 8\}$. A positive quadratic term is applied only to keys outside the exact set, using either the leading directions of the selected keys with the full-history center (H), frozen training-global directions (G), or full-history local directions (L, a diagnostic reference). Its constant feature replaces the zero-order state rather than being stored twice. With $M_r = 1 + r + r(r+1)/2$, the extra feature storage is $(M_r - 1)(d_v + 1)$, plus rd for a per-block H or L basis; the frozen G basis is stored once per head and is not charged to the block.

Table 18: Representative low-dimensional remainder configurations. Storage is the count of stored numbers per block at $N = 16{,}352$, including exact KV and all per-block state. MACs are relative to dense attention. Errors are test-split L2/$V_{\max}$ means; H, G and L are defined in the text.

Anchors, m	Remainder	Storage	MAC/dense	Llama	Qwen
1, $N/128$	drop	32,768	0.024	0.0377	0.0450
1, $N/128$	zero-order	33,025	0.056	0.0272	0.0338
1, $N/128$	P_2, r=1, H	33,411	0.110	0.0272	0.0332
1, $N/128$	P_2, r=2, H	33,926	0.173	0.0260	0.0322
1, $N/128$	P_2, r=4, H	35,343	0.345	0.0248	0.0310
1, $N/128$	P_2, r=8, H	39,725	0.876	0.0237	0.0300
1, $N/128$	P_2, r=1, G	33,283	0.103	0.0258	0.0323
1, $N/128$	P_2, r=8, G	38,701	0.868	0.0231	0.0297
1, $N/128$	P_2, r=8, L	39,725	1.875	0.0228	0.0290
4, $N/4$	zero-order	1,046,785	0.341	0.0048	0.0042
4, $N/4$	P_2, r=8, H	1,053,485	1.207	0.0040	0.0037
4, $N/4$	P_2, r=1, G	1,047,043	0.376	0.0045	0.0041
4, $N/4$	P_2, r=8, G	1,052,461	0.955	0.0039	0.0037
none	P_2, r=8, L	6,957	1.873	0.0695	0.0653
none	P_2, r=1, G	515	0.095	0.0806	0.0810
none	P_2, r=8, G	5,933	0.867	0.0720	0.0707

The comparison arm is a zero-order gate with the same selector and with its exact-set size solved separately for each window to match either stored numbers or counted MACs. The MAC account includes constructing the selected-set covariance for H, projecting all remainder keys, accumulating all feature states and contracting them for every query. H and L eigendecomposition adds a separately reported $O(d^3)$ term; adding $3d^3$ does not change the comparison's direction. Budgets are rounded to key counts and capped at the full history. The comparison is not a latency measurement and does not charge top-m selection or memory movement.

At the same $m = N/128$ (Table 18), H at $r = 8$ reduces error from .0272/.0338 to .0237/.0300 (Llama/Qwen), while the full-history basis gives .0228/.0290. At equal storage, the low-dimensional remainder is better in 78 of the 112 H/G configurations in both models (paired intervals below zero). At equal counted MACs it is worse in all 112. This remains true for G, whose basis has no per-block construction cost. The result therefore cannot be attributed solely to covariance construction: evaluating features for every remainder key is itself expensive. It establishes a resource-dependent complementarity, not the uselessness of the low-dimensional term.

A standalone P_2 realization on all history keys is also compared with the zero-order gate at the same resources, over both L/G bases, four dimensions and two gate selectors. Every one of these 16 comparisons favors the gate at equal storage and at equal MACs on both models. At $r = 8$, the local standalone summary stores 6,957 numbers; the same storage buys 26 exact keys plus the zero-order state, with errors .0293/.0382 against .0695/.0653 for the standalone summary. These are comparisons among the tested implementations, not a lower bound against every separable kernel.

G.2 Arithmetic and storage conventions

We count the multiply-accumulates (MACs) of the matrix products per block of $L = 32$ queries over N historical keys with key and value dimension $d = d_v = 128$. Exponentials, normalizations, the top-m selection and data movement are not counted; the paragraph after the list returns to them, Appendix G.3 counts them for one fused design, and Appendix G.4 reports measured times. The counts below, Table 21 and the cost of every fixed-budget arm in Tables 1 and 29 come from one script of the supplementary code (`cost_ledger.py`), which reproduces the stored cost of each of these runs exactly.

1. **Dense Attention Reference**: Each query computes N QK dot-products and N AV dot-products, plus in-block causal interactions. For a block of 32 queries, the MACs are:

$$\mathrm{MACs}_{\mathrm{dense}} \;=\; 32 \cdot 2d\,N \;+\; \frac{32 \times 33}{2} \cdot 2d \;=\; 64d\,N \;+\; 1056d. \tag{20}$$

   Amortized per query, this is $2d\,N = 256N$ MACs to leading order in N, plus $33d$ for the causal block.

Table 19: Resource-matched differences for Llama-3.2-3B. Each entry is the candidate's error minus that of the zero-order gate at the matched budget, with a paired 95% cluster-bootstrap interval. Negative favors the candidate; for the standalone rows (none) the gate uses one anchor row. †: matching saturates at all history keys in every window, so the comparison gate is exact.

Anchors, m	Remainder	Equal storage	Equal MACs
1, $N/128$	P_2, r=1, H	$+0.0002$ [$+0.0001, +0.0003$]	$+0.0098$ [$+0.0093, +0.0104$]
1, $N/128$	P_2, r=2, H	-0.0008 [$-0.0009, -0.0007$]	$+0.0128$ [$+0.0121, +0.0135$]
1, $N/128$	P_2, r=4, H	-0.0016 [$-0.0018, -0.0015$]	$+0.0173$ [$+0.0164, +0.0181$]
1, $N/128$	P_2, r=8, H	-0.0015 [$-0.0017, -0.0013$]	$+0.0228$ [$+0.0218, +0.0238$]
1, $N/128$	P_2, r=1, G	-0.0012 [$-0.0013, -0.0011$]	$+0.0077$ [$+0.0072, +0.0082$]
1, $N/128$	P_2, r=8, G	-0.0023 [$-0.0025, -0.0020$]	$+0.0222$ [$+0.0212, +0.0233$]
1, $N/128$	P_2, r=8, L	-0.0023 [$-0.0025, -0.0021$]	$+0.0228$ [$+0.0218, +0.0238$]†
4, $N/4$	P_2, r=8, H	-0.0007 [$-0.0008, -0.0007$]	$+0.0040$ [$+0.0038, +0.0042$]†
4, $N/4$	P_2, r=1, G	-0.0003 [$-0.0003, -0.0003$]	$+0.0003$ [$+0.0002, +0.0003$]
4, $N/4$	P_2, r=8, G	-0.0008 [$-0.0009, -0.0008$]	$+0.0037$ [$+0.0035, +0.0039$]
none	P_2, r=8, L	$+0.0402$ [$+0.0381, +0.0427$]	$+0.0695$ [$+0.0669, +0.0723$]†
none	P_2, r=1, G	$+0.0369$ [$+0.0349, +0.0390$]	$+0.0617$ [$+0.0585, +0.0648$]
none	P_2, r=8, G	$+0.0421$ [$+0.0396, +0.0449$]	$+0.0711$ [$+0.0680, +0.0744$]

Table 20: Resource-matched differences for Qwen3-8B-Base. Each entry is the candidate's error minus that of the zero-order gate at the matched budget, with a paired 95% cluster-bootstrap interval. Negative favors the candidate; for the standalone rows (none) the gate uses one anchor row. †: matching saturates at all history keys in every window, so the comparison gate is exact.

Anchors, m	Remainder	Equal storage	Equal MACs
1, $N/128$	P_2, r=1, H	-0.0002 [$-0.0003, -0.0002$]	$+0.0141$ [$+0.0134, +0.0147$]
1, $N/128$	P_2, r=2, H	-0.0008 [$-0.0009, -0.0007$]	$+0.0180$ [$+0.0172, +0.0187$]
1, $N/128$	P_2, r=4, H	-0.0013 [$-0.0014, -0.0012$]	$+0.0230$ [$+0.0222, +0.0240$]
1, $N/128$	P_2, r=8, H	-0.0003 [$-0.0006, -0.0001$]	$+0.0288$ [$+0.0277, +0.0299$]
1, $N/128$	P_2, r=1, G	-0.0013 [$-0.0013, -0.0012$]	$+0.0123$ [$+0.0117, +0.0129$]
1, $N/128$	P_2, r=8, G	-0.0010 [$-0.0012, -0.0007$]	$+0.0285$ [$+0.0274, +0.0297$]
1, $N/128$	P_2, r=8, L	-0.0013 [$-0.0015, -0.0011$]	$+0.0290$ [$+0.0278, +0.0301$]†
4, $N/4$	P_2, r=8, H	-0.0005 [$-0.0005, -0.0004$]	$+0.0037$ [$+0.0035, +0.0039$]†
4, $N/4$	P_2, r=1, G	-0.0002 [$-0.0002, -0.0001$]	$+0.0004$ [$+0.0004, +0.0004$]
4, $N/4$	P_2, r=8, G	-0.0005 [$-0.0005, -0.0005$]	$+0.0035$ [$+0.0033, +0.0037$]
none	P_2, r=8, L	$+0.0271$ [$+0.0251, +0.0292$]	$+0.0653$ [$+0.0625, +0.0679$]†
none	P_2, r=1, G	$+0.0116$ [$+0.0105, +0.0126$]	$+0.0600$ [$+0.0581, +0.0619$]
none	P_2, r=8, G	$+0.0314$ [$+0.0292, +0.0337$]	$+0.0695$ [$+0.0667, +0.0721$]

2. **Few-Direction Separable Summary (P_2 with r directions)**: The summary keeps $M_r = 1 + r + r(r+1)/2$ block states, $M_8 = 45$ for $r = 8$. For each block it projects the N history keys onto the r directions ($rd\,N$ MACs) and accumulates the M_r features of every key, weighted by $e^{s_a(j)}$, against its value ($M_r(d_v+1)\,N$ MACs); each query then contracts its M_r features against the states ($M_r(d_v+1)$ MACs). The local basis also needs the tilted covariance of the history, $d(d+1)N/2$ MACs per block if it is formed explicitly, which alone is as large as the $64dN$ MACs of dense attention for a block of 32 queries. Forming $J = \beta RCR$ and its leading eigenvectors adds $O(d^3)$ work per block that does not depend on N; our code forms C explicitly and uses a dense symmetric eigensolver. The projection cost $rd\,N$ above presupposes these directions and does not include their construction. The tilt e^{s_a}, the center $\bar{k}$ and the local basis all change with the anchor, so the states are rebuilt for every block: this construction costs $O(N)$ per block, like the anchor-gated representation, and it is not the running-state linear attention of Katharopoulos et al. (2020). We use it as the separable reference of Section 3, not as an efficient operator.
3. **Anchor-Gated Representation (n_a anchor rows, exact budget m)**: For a block of 32 queries with exact-set budget m:
   - **Anchor QK scoring**: n_a anchor rows evaluate QK against N keys: $n_a \cdot d\,N$ MACs (with no AV operations).
   - **Tilted mean $\bar{k}$**: One anchor row's softmax weights compute the d-dimensional mean: $d\,N$ MACs.

- **Zero-order summary states**: Summing scalar weights S_0 and values S_1 over unselected keys: $(d_v + 1)(N - m)$ MACs.
- **Exact set evaluation**: 32 queries compute full QK and AV against m keys: $32 \cdot 2d\, m = 64d\, m$ MACs.
- **Per-query tilt**: Each query forms $\beta (q_i - q_a)^\top \bar{k}$: $32d$ MACs.
- **Remainder in each query's softmax**: Each query adds the remainder value S_1/S_0, weighted by $e^{\log S_0 + c_i}$, to its numerator and that weight to its normalizer: $32\,(d_v + 1)$ MACs.
- **Causal in-block keys**: Exact computation on the causal triangle: $\frac{32 \times 33}{2} \cdot 2d = 1056d$ MACs.

MACs of the listed products for the block:

$$\text{MACs}_{\text{ours}} \;=\; \big((n_a{+}1)\,d{+}d_v{+}1\big)\,N \;+\; (64d{-}d_v{-}1)\,m \;+\; 1088d \;+\; 32\,(d_v{+}1). \quad (21)$$

For $d = d_v = 128$ and the four anchor rows used end to end ($n_a = 4$), the coefficient on N is $5(128) + 129 = 769$ MACs per block, about $24N$ per query, and each exact key costs $(8192 - 129)/32 \approx 252$ MACs per query, about $63N$ at $m = N/4$. The amortized total is about $87N$ MACs per query, 34% of dense attention ($256N$). With one anchor row the coefficient on N is about $12N$ per query and the total about $75N$, 29% of dense.

4. **Key-pair account used in the main text**: The cost columns of Tables 1 and 29 use the coarser key-pair account defined below, which our code evaluates for every block of every run. It charges each anchor row as a full query row ($2d$ MACs per key; for the block-first row the AV slot covers $\bar{k}$) and the summary as $(d_v + 1)/(2d)$ of a key pair for each of the N history keys. It differs from Equation (21) by charging d more MACs per key for each anchor row beyond the first, $4(n_a - 1)N$ per query, by summing the summary over all N keys, and by leaving out the per-query tilt and the remainder term of each query's softmax, $32(d + d_v + 1)$ MACs per block that do not grow with N; one query costs about $(8n_a{+}4)N{+}256m$ dense-equivalent MACs, at 256 per key pair. For large N and $m = N/4$, a gated block costs $(n_a + 129/256 + 8)/32$ of a dense block, .297 with one anchor row and .391 with four. The costs .300 and .393 in Table 1 are this account summed over all blocks of an input, with the first block and any final partial block computed densely, and averaged over the 200 inputs of 4.2K to 16K tokens; shorter inputs give slightly larger ratios (Table 21).

What neither account includes. Neither count includes selecting the top m keys, a partial sort or radix selection over N scores per block and head, the exponentials and normalizations of the anchor rows and of the block softmax, or memory traffic, and neither reuses the anchor rows' scores on the exact set. Whatever m is, every block reads the keys of the whole history for its anchor rows and $\bar{k}$ and the values of the whole history for S_1 and the exact set, and it gathers the exact set from irregular addresses. The arithmetic saving becomes a wall-clock saving only if a fused kernel shares these reads across blocks and keeps the selection cheap. We have not built such a kernel. Our implementation is an unfused PyTorch reference written to measure error; it is slower than dense attention (Appendix G.4), and its run times say nothing about attainable speed. Finally, with m a fixed fraction of N, as in the main fixed-fraction comparisons, the cost per query grows linearly in N and prefill attention remains quadratic in sequence length: the representation lowers the constant, not the order (Table 21). The same holds at a fixed m: the terms of Equation (21) that scale with N are paid once per block, so over a prompt of n tokens they sum to about $n^2/(2L)$ times $(n_a + 1)d + d_v + 1$ MACs, a fraction $\big((n_a + 1)d + d_v + 1\big)/(64d)$ of dense attention, .047 with one anchor row and .094 with four. Table 21 follows one block as its history grows; it is not a scaling law for a whole prefill. Attention is also only part of prefill: from the model configurations, it accounts for .33 of the multiply-accumulates of a 16K-token prefill in Llama-3.2-3B and .26 in Qwen3-8B-Base (the query, key, value, output and MLP projections counted, embeddings and the output head not), so at the key-pair fraction of .39 that Table 21 gives at 16K, the key-pair account bounds the saving on the whole prefill by about 20% and 16% (22% and 17% with the MAC count of Equation (21) summed over the same prefill, a fraction of .34).

G.3 A TRAFFIC MODEL FOR A FUSED KERNEL

The two accounts above count arithmetic only. To see what a fused kernel would have to read and write, we model one design, which we have not built: a pass over the history, a selection, and a

Table 21: **Arithmetic cost across lengths N.** Mega-MACs per block of $L = 32$ queries with N history keys ($d = d_v = 128$, $m = N/4$), from Equation (21) for one and four anchor rows, and their ratio to dense attention. The last column gives the key-pair account defined here for a whole prefill of N tokens, computed as our code does. Neither count includes the top-m selection or memory traffic.

N	Dense	1 anchor row	4 anchor rows	MAC ratio (1 / 4 rows)	Key-pair ratio (1 / 4 rows)
4,096	33.7 M	10.0 M	11.5 M	.296 / .343	.303 / .396
16,384	134.4 M	39.5 M	45.8 M	.294 / .341	.298 / .392
32,768	268.6 M	78.8 M	91.4 M	.293 / .340	.298 / .391
65,536	537.0 M	157.5 M	182.6 M	.293 / .340	.297 / .391
131,072	1073.9 M	314.8 M	365.1 M	.293 / .340	.297 / .391

pass over the selected keys. Proposition 6 counts its multiply-accumulates (with the conventions of Equation (21)), exponentials, logical memory traffic (each stored element read or written once per pass) and temporary memory. The supplementary script `cost_propE.py` evaluates these counts; a per-operation simulation of the design (`propE_bruteforce.py`) reproduces every count, and on small random inputs its output agrees with that of our implementation up to fp32 rounding.

Proposition 6 (Cost of a two-pass gated block)**.** *Take one query head and a block of L queries whose history has N keys, with n_a anchor rows (the block-first query among them), an exact set of $m \le N$ keys and the zero-order remainder. Keys, values, queries and outputs take B bytes per element; anchor scores, block states and indices take 4 bytes. Pass 1 reads the keys and values of the history once: the anchor rows score every key with an online logsumexp, the block-first row also accumulates $\bar{k}$ and the states (S_0, S_1) over all N keys, and the $n_a N$ scores are written out. The selection combines the anchor rows into $\max_a p_a(j)$ in one pass when $n_a > 1$ and finds the top m keys with a radix or histogram select that reads the N scores P times ($2 \le P \le 4$), the last pass writing the m indices. Pass 2 gathers the m selected keys and values, reads the block's own keys and values, and evaluates the softmax of each query over the exact set, its causal keys and the remainder key. The block-first query starts its running maximum at the pass-1 maximum, so its sums over the exact set, whose logits are the anchor logits, are on the scale of (S_0, S_1), and the remainder states are the full-history states minus these sums. Write $[n_a{>}1]$ for 1 if $n_a > 1$ and 0 otherwise. Per block:*

- *(i)* MACs*: $\big((n_a+1)d + d_v + 1\big)N + L(d+d_v)m + \frac{L(L+1)}{2}(d+d_v) + L(d+d_v+1)$. At $L = 32$ and $d = d_v$ this is Equation (21) plus $(d_v+1)m$, because the states are summed over all N keys rather than over the $N - m$ unselected ones; the subtraction takes $d_v + 1$ subtractions and no multiply-accumulates. Reading the anchor rows' exact-set logits from the stored scores instead of recomputing them saves $n_a dm$ MACs and adds $4n_a m$ bytes. On the key-pair account defined here, at $L = 32$ and $d = d_v$, the block costs $\big(\mathrm{MACs}_{\mathrm{ours}} + (n_a - 1)dN + (d_v+1)m - 32(d+d_v+1)\big)/(2d)$ key pairs.*
- *(ii)* Exponentials in softmax numerators*: $n_a N + L(m+1) + L(L+1)/2$, against $LN + L(L+1)/2$ for dense attention; both normalize L output rows.*
- *(iii)* HBM traffic*: $B(d+d_v)(N+m+L)$ bytes of keys and values, $B\big(d(L+n_a) + d_v L\big)$ of queries and outputs, $4\big(n_a + [n_a{>}1](n_a+1) + P\big)N + 8m$ of scores and indices, and $8(d+d_v+2) + 8[n_a{>}1]\,n_a$ of block states. The writes among them are the Bd_vL bytes of the output and the temporary memory of (iv); the rest are reads.*
- *(iv)* Temporary memory *per block in flight, live from pass 1 to pass 2: $4\big((n_a + [n_a{>}1])N + m + d + d_v + 2 + [n_a{>}1]\,n_a\big)$ bytes of scores, combined scores, indices and states (on-chip accumulators and histograms not included).*
- *(v)* Selection*: $P + [n_a{>}1]$ passes, which read $4(P + [n_a{>}1]\,n_a)N + 4[n_a{>}1]\,n_a$ bytes and write $4[n_a{>}1]N + 4m$. A radix select with 11-bit digits (three histogram passes of 2,048 counters and one output pass) has $P = 4$; $P = 2$ assumes that the boundary bucket of the first pass fits on chip.*
- *(vi)* Grouped-query attention*: if the g query heads that share a key-value head are processed together, the keys and values of pass 1 and of the block, $B(d+d_v)(N+L)$ bytes, are read once for the g heads. Each head selects its own exact set, as in our implementation, and computes only on its own m keys, so pass 2 gathers the union of the g sets, which has between m and $\min(N, gm)$ rows. Per query head the gathered bytes therefore lie in the*

interval $\big[B(d+d_v)m/g,\ B(d+d_v)\min(N,gm)/g\big]$*; we have not measured where in this interval the heads fall.*

(vii) Sharing across blocks*: if G consecutive blocks share one pass-1 stream, the history is read once for the G blocks, up to the history of the last one; if they also share one selection, the design is the one above with block size GL.*

Derivation. Each item is a direct count. Pass 1 spends $n_a d$ MACs and n_a exponentials per history key on the anchor rows and $d + d_v + 1$ MACs per key on $\bar{k}$, S_1 and S_0. Pass 2 spends $L(d + d_v)$ MACs and L exponentials per exact key, $d + d_v$ MACs and one exponential per causal pair, and per query d MACs on the tilt and $d_v + 1$ MACs and one exponential on the remainder key. With $L = 32$ and $d = d_v$ this is Equation (21) with $(d_v + 1)(N - m)$ replaced by $(d_v + 1)N$, which gives (i); the key-pair form is $2d$ times the per-block account defined in this appendix minus Equation (21). The block-first query's logits on the exact set never exceed the pass-1 maximum, so starting its running maximum there needs no rescaling. The traffic counts one read or write of each stored element per pass; the temporary memory is everything pass 1 and the selection write, kept until pass 2 of the block. Online-softmax rescalings, the selection's comparisons and the per-block logarithm and reciprocal are not counted, as for dense attention. In fp32 the subtraction loses the remainder when its mass is below about 10^{-7} of the anchor's total mass, as in our implementation, which forms the remainder the same way. □

What the model implies on an H100. We bound the time of a kernel by the largest of three terms: its FLOPs (2× MACs) at the dense bf16 tensor-core peak of an H100 SXM5 (989 TFLOPS), its HBM bytes at 3.35 TB/s, and its exponentials at 3.9×10^{12} per second; the ridge point is 295 FLOP per byte, and the exponential term binds nowhere in the grid below. Ratios compare these bounds, that is, the times of two kernels that both attain the roofline. The reference is a FlashAttention-style dense kernel (Dao et al., 2022) with 128-query tiles that reads the keys and values of the history once per tile: at $d = d_v = 128$ and $B = 2$ a query costs $512N$ FLOPs and $4N$ bytes, 128 FLOP/B. Both bounds are optimistic, ours more so, because the anchor rows and the 32-row exact sets are too thin for the tensor cores unless they are batched across blocks or heads. With $L = 32$, $m = N/4$ and one pass per block, a query of the gated block costs $152N$ FLOPs with one anchor row and $176N$ with four, against $20.4N$ to $21.7N$ bytes: 7.4 to 8.2 FLOP/B, far below the ridge. It moves 5.1 to 5.4 times the bytes of the dense kernel per query, and its bound is 5.0 to 5.3 times the dense kernel's over a whole 16K-token prefill (5.1 to 5.4 at 128K). Sharing the history stream across blocks does not close the gap: with a selection per block, the gathered exact set alone reads $B(d + d_v)m/L = 4N$ bytes per query at $m = N/4$, as many as the dense kernel, so the byte ratio falls only to 1.11 to 1.42 as $G \to \infty$. Sharing one selection among G blocks lifts the intensity above the ridge only for LG of 1,500 to 1,600 queries per selection. Grouped-query attention helps the dense kernel at least as much as ours, and more unless the heads' exact sets coincide: at $g = 3$ (Llama-3.2-3B) and $g = 4$ (Qwen3-8B-Base), with the same packing for both kernels, the byte ratio at 16K is 5.1 to 6.3 if the g sets coincide and 7.0 to 9.1 if they are disjoint. Of the two cache assumptions, only keys and values that stay in L2 can put the gated block's bound below the dense one at the operating point of Section 6. Its HBM traffic is then one read of the keys and values plus the queries, outputs, scores, indices and states, and at 16K its bound is 0.33 to 0.47 of the dense kernel's with one anchor row and 0.91 to 1.05 with four, whose score traffic then dominates; it would need 4.7 to 15 times the L2 bandwidth that the dense kernel needs, and the keys and values of one head at 128K tokens (64 MiB) exceed the 50 MB L2. Across the grid of the supplementary script (N up to 128K, $L \in \{32, 128\}$, $G \in \{1, 4, 16\}$, $m \in \{N/4, N/8, 1024\}$), the per-block selection of 32 queries used in this paper falls below the dense bound without L2 residency only when 16 blocks share each pass over the history and $m \le N/8$. The exceptions need $g = 4$ packed heads, one anchor row and at least 64K tokens: $m = N/4$ with 16 blocks per pass, and then only if the four heads' sets nearly coincide (a union of at most $1.04m$ rows, with $P = 2$), and $m = 1024$ with 4 blocks per pass, which tolerates partial overlap (a union of up to $2.6m$ rows at 64K, with $P = 2$). Under this model, with $L = 32$ and no reuse of the streamed history across blocks, the gated prefill moves more bytes per query than FlashAttention and is memory-bound, whatever m is (still 4.09 times the bytes as $m/N \to 0$); a speedup requires serving each pass over the history to more queries (a larger L, or one selection shared by G blocks), and with a selection per block also a smaller m.

G.4 Measured times of the unfused implementation

Component times. We timed an unfused PyTorch implementation of one gated layer on one H100 (80 GB HBM3), with bf16 random inputs, batch 1, the query and key-value head counts of each model, CUDA events and medians of 20 runs, in 240 configurations: N from 8K to 128K, $L \in \{32, 128\}$, $m \in \{N/4, N/8, 1024, 4096\}$ and three selectors (Table 22 lists the operating point of Section 6). With exact per-block budgets it computes the gated layer of Section 5: its output differs from our fp32 reference by a relative Frobenius norm of at most 2.8×10^{-3}, against 2.0×10^{-3} between bf16 and fp32 dense attention on the same inputs. The timed plans with $m = N/4$ or $N/8$ give every block of a group of consecutive blocks the largest budget in the group, 2 to 6% more exact keys in total. It does not realize the arithmetic saving. At $N = 16{,}384$, $L = 32$ and $m = N/4$, selecting the top m keys alone takes 1.4 (Llama-3.2-3B) and 0.96 (Qwen3-8B-Base) times the time of the fastest dense kernel we timed (FlashAttention-3 or cuDNN), the selection with four anchor rows 1.8 and 1.3 times, gathering the exact sets 7.4 and 6.4 times, and the whole layer with four anchor rows 22 and 18 times; at 128K tokens the top-m selection alone takes 0.6 of the dense time, the gather 6.4 to 6.5 times and the whole layer 14 to 15 times. By the count of Proposition 6 (we could not read hardware counters), at $m = N/4$ the gathered exact sets hold about as many bytes as dense attention reads when each key-value head gathers one union of the sets of its query heads and these sets coincide (0.96 to 0.97 times at 16K), and 2.9 to 3.8 times as many when each query head gathers its own set, as here. Anchor rows, gather, remainder and exact-set attention each take more than three times their bound from Proposition 6 and the H100 roofline of Appendix G.3 (with $P = 4$ and one gathered set per query head) in every configuration; the selection comes closest, at 1.8 to 2.5 times its bound with four anchor rows and $L = 32$ from 32K tokens on. The progressive selector of Section 6 selects and gathers once per sub-block: at 16K its selection takes 3.7 times and its gather 4 times as long as with four anchor rows, and its whole layer takes 45 to 66 times the dense time from 16K to 128K tokens.

Where the layer ran faster than dense attention. In two of the 240 configurations the unfused layer took less time than the fastest dense kernel: at 128K tokens with blocks of $L = 128$ queries, one anchor row and a fixed $m = 1024$, it took 0.82 (Llama-3.2-3B) and 0.79 (Qwen3-8B-Base) of the dense time. We have not measured the operator error with 128-query blocks; with 32-query blocks, $m = 1024$ leaves a relative L2 error of .175 on the 128K filler windows of Llama-3.2-3B (Appendix G.6), far above the target of .05 used there.

The end-to-end times. The end-to-end runs of Section 6 use a different implementation: an fp32 reference of the gated layer, which in every layer also computes a bf16 reference output for the per-layer error (Table 23). At $N = 16{,}384$ with four anchor rows and $m = N/4$, one layer of this reference takes 332 and 442 ms, 150 and 131 times the fastest dense kernel; over the 28 and 36 layers this is 9.3 and 15.9 s, most of the median prefill times of 9.78 and 17.0 s that these runs recorded on the H100, against 0.291 and 0.630 s for exact attention. The recorded times therefore describe the reference implementation, not the representation, and they are not times to first token.

Share of prefill. Dense attention takes about 24% (Llama-3.2-3B) and 19% (Qwen3-8B-Base) of the exact prefill time at 16K tokens on the H100, estimated as the time of one SDPA layer with a causal mask on random inputs, times the number of layers, over the recorded median prefill time. Approximating attention alone could therefore shorten this prefill by at most that share, and a kernel whose time followed the key-pair fraction of .39 by at most 14% and 11%. These are shares of time; the shares of .33 and .26 above count multiply-accumulates.

G.5 Budget scaling with history length

The experiments of this paper fix m as a fraction of N. Here we ask how large m has to be to hold the operator error at a fixed level as the history grows, within the lengths our captures cover. We take windows of 32 queries whose history has $N \in \{1024, 2048, 4096, 8192, 12288, 16320\}$ keys on the captured heads of Section 2, in all eight sequences of each task. On the synthetic tasks these windows end before the question and contain only filler; LongBench items contribute a window only if its history is at least 64 tokens shorter than the item; the final block of each synthetic sequence, which holds the question, is treated separately below. In each window we evaluate the gate with the zero-order remainder in float64, with the block-first anchor or with the four anchor rows of Section 6, on a grid of exact budgets below N that combines the powers of two from 16 to 4096 with the fractions $N/128$ to $N/4$. We average the error over windows as in Section 2 and define $m_\varepsilon(N)$ as

Table 22: **Measured times of one gated layer in our unfused implementation**, in milliseconds, at $L = 32$ and $m = N/4$: one H100 (80 GB HBM3), batch 1, bf16 random inputs with the query and key-value head counts of each model, CUDA events, medians of 20 runs after 2 warm-up runs. Dense is the fastest dense kernel with a causal mask that we timed. The components are the anchor rows (scores and softmax), the selection (combining the anchor rows and the top-m selection), the gather of the exact keys and values, the remainder states, and the attention over the exact set, the causal block and the remainder key; the last column is the whole layer, including the rows before the first gated block, divided by the dense time. † Looks ahead within the block; the progressive selector is strictly causal. These are times of an unfused reference written to measure error, not of a deployable kernel.

Model	Selector	N	Dense	Anchor	Select	Gather	Remainder	Exact attn.	Layer / dense
Llama-3.2-3B	1 anchor	16K	2.2	4.3	3.2	16.3	6.5	10.0	18.3
	4 anchors†	16K	2.2	11.3	4.0	16.3	6.4	10.0	21.8
	progressive	16K	2.2	11.4	14.7	65.2	18.3	34.9	65.5
	1 anchor	128K	153.6	150.7	91.8	1002.8	191.5	331.9	11.5
	4 anchors†	128K	153.6	581.1	136.0	1002.9	195.9	332.4	14.6
	progressive	128K	153.6	581.2	488.8	4011.3	666.5	1280.3	45.8
Qwen3-8B-Base	1 anchor	16K	3.4	5.0	3.3	21.7	7.5	12.3	14.8
	4 anchors†	16K	3.4	14.4	4.3	21.7	7.4	12.2	17.8
	progressive	16K	3.4	14.4	15.9	86.7	21.7	42.0	53.7
	1 anchor	128K	206.9	199.2	121.6	1330.7	253.9	440.3	11.3
	4 anchors†	128K	206.9	770.8	180.3	1330.7	259.7	441.2	14.4
	progressive	128K	206.9	770.8	648.0	5322.5	883.3	1698.5	45.1

Table 23: **Prefill times recorded with the end-to-end run** of Table 29: medians in seconds over the synthetic inputs of 16,384 tokens on each GPU type (39 to 81 inputs per arm and GPU), and the peak allocated GPU memory in GiB, model weights included. Every gated arm runs our fp32 reference implementation of the gated layer and also computes a bf16 reference output in every layer for the per-layer error; the exact arm computes neither. The times describe that reference implementation, not the representation, and are not times to first token.

		Llama-3.2-3B			Qwen3-8B-Base		
Arm	m	H100	H200	GiB	H100	H200	GiB
Exact attention	–	0.29	0.28	8.9	0.63	0.62	19.2
1 anchor	$N/4$	9.2	8.5	10.5	16.0	14.7	20.7
4 anchors	$N/4$	9.8	9.0	11.0	17.0	15.6	21.2
4 anchors	$N/8$	5.6	5.1	11.0	9.8	8.9	21.2
4 anchors, drop	$N/8$	5.5	5.0	11.0	9.7	8.7	21.2
4 anchors	$N/32$	3.2	2.8	11.1	5.6	5.0	21.2
4 anchors, tail τ=.05	adaptive	27.4	26.2	11.0	44.0	37.6	21.2
4 anchors, tail τ=.1	adaptive	14.4	12.8	11.0	26.3	21.5	21.2

the smallest budget at which this average reaches ε, interpolating $\log m$ linearly in the logarithm of the error between grid points. A length at which the smallest grid budget already meets the target, or the largest misses it, is censored and left out of the least-squares fit of $\log m_\varepsilon(N) = a + \alpha \log N$; the 95% interval of α comes from the cluster bootstrap of Section 2, with the whole fit repeated in every replicate. Before the sweep we fixed the targets, a relative L2 error of .05 (close to the question-row error of the four-anchor runs of Section 6) and an L2/Vmax error of .03, with relative L2 deciding when the two disagree, and the reading of the interval: inside $[0.8, 1.2]$ it means growth in proportion to N, below 0.8 sublinear growth, and an interval that crosses 0.8 or 1.2 leaves the question open. Everything in this subsection is limited to histories of at most 16K keys. Appendix G.6 extends the measurement to longer histories on synthetic inputs.

With a relative L2 target of .05, $m_\varepsilon(N)$ grows about in proportion to N on Llama-3.2-3B: the slope is 0.98 [0.93, 1.03] with the block-first anchor and 0.89 [0.86, 0.92] with four anchor rows (Figure 16(a)). On Qwen3-8B-Base the slopes are 0.85 [0.80, 0.90] and 0.78 [0.76, 0.81]; both intervals cross the 0.8 boundary, so four octaves are too short to call this growth sublinear, although neither interval contains 1. With one anchor row, even 4096 keys miss the target at $N = 12288$ and 16320 on Llama-3.2-3B and at 16320 on Qwen3-8B-Base, and these lengths are censored. With four anchor rows, the budget

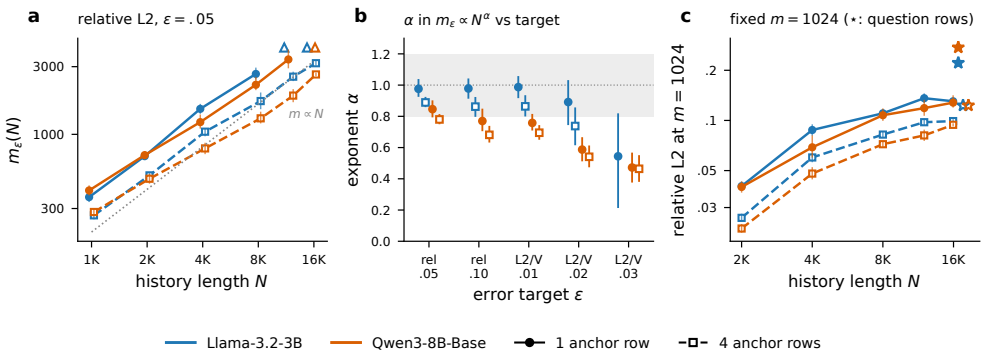

Figure 16: **Exact budget for a fixed operator error against history length** (zero-order remainder; one anchor row: filled circles, solid lines; four anchor rows: open squares, dashed lines). (a) $m_\varepsilon(N)$ at a relative L2 error of .05 on the five tasks, with 95% cluster-bootstrap intervals; a triangle marks a length at which the largest grid budget misses the target; the dotted line has slope 1. (b) Fitted exponent α of $m_\varepsilon \propto N^\alpha$ against the error target, with 95% intervals; the grey band $[0.8, 1.2]$ is the range read as growth in proportion to N, fixed before the sweep; a missing point is not estimable because the smallest grid budget already meets the target at too many lengths. (c) Relative L2 error at a fixed $m = 1024$ on the filler windows of the synthetic tasks, and on the question rows of the final block ($N = 16352$, stars).

falls only from .26 to .19 of the history on Llama-3.2-3B and from .28 to .16 on Qwen3-8B-Base between $N = 1024$ and 16320. Up to 16K keys, the budget that holds the operator error at a fixed level therefore grows about in proportion to N on one model and nearly so on the other, and the representation lowers the constant of prefill attention and not its order. A slower-growing budget would not change the order either, because every block reads the whole history for its anchor rows and its remainder state (the fixed-m share above).

The exponent depends on how tight the target is (Figure 16(b)). At the L2/Vmax level of .03 used in Section 5, 23 to 104 keys suffice at every length on Qwen3-8B-Base with the block-first anchor, and the budget grows only as $N^{0.47}$ [0.38, 0.56] with one anchor row and $N^{0.46}$ [0.39, 0.55] with four. On Llama-3.2-3B the smallest grid budget, 8 to 16 keys, already meets this level at 3 of the 6 lengths with one anchor row and at 5 with four, so the exponent is not estimable with four rows. Tightening the target to an L2/Vmax error of .01 raises the slopes to 0.99 and 0.86 on Llama-3.2-3B and to 0.76 and 0.70 on Qwen3-8B-Base. We decide on relative L2, as fixed before the sweep: at the $\|o\|/V_{\max}$ of .13 to .14 (Llama-3.2-3B) and .16 to .17 (Qwen3-8B-Base) in these windows, an L2/Vmax error of .03 is a relative error of about .2, far looser than the question-row error of the four-anchor runs end to end. The ratio $\|o\|/V_{\max}$ changes by less than 8% across the six lengths, so the two metrics disagree because of the target, not because the denominator drifts with N.

The five-task set changes with N: LongBench items contribute fewer windows at the longest histories and none at 16320. On the three synthetic tasks alone, which have windows at every length, the slopes are 0.97 [0.93, 1.02] and 0.87 [0.85, 0.89] on Llama-3.2-3B and 0.85 [0.80, 0.88] and 0.76 [0.75, 0.78] on Qwen3-8B-Base. The last interval lies below 0.8, the only reading that changes; we keep the reading of the five-task set, as fixed before the sweep. Dropping the keys outside the exact set instead of keeping the remainder gives slopes of 0.83 to 0.96.

At a fixed $m = 1024$ the error keeps growing with the history (Figure 16(c)). On the filler windows of the synthetic tasks, the relative error grows from $N = 4096$ to 16320 by a factor of 1.49 [1.38, 1.60] on Llama-3.2-3B and 1.85 [1.57, 2.17] on Qwen3-8B-Base with the block-first anchor, and by 1.65 [1.55, 1.76] and 1.96 [1.81, 2.11] with four anchor rows. On the question rows of the final block ($N = 16352$) it is .22 and .27 with one anchor row and .12 on both models with four.

The question rows are harder to approximate than the filler. At $m = N/4$ their relative error is .124 on Llama-3.2-3B and .157 on Qwen3-8B-Base with the block-first anchor, against .058 and .047 with four anchor rows. On the same 24 synthetic inputs and the same two query heads per layer, the end-to-end runs with four anchor rows at $m = N/4$ (Section 6) show a mean per-layer question-row error of .058 and .048. Reaching that level at the operator level takes $m = 4064$ and 3955 with four anchor rows, about $N/4$ (on Llama-3.2-3B, 16% of the bootstrap replicates do not reach it within the grid); a single block-first anchor does not reach this level within $m \le N/4$. By

contrast, with one anchor row the mean L2/Vmax error over all rows of a filler window at $N = 16320$ is already .026 and .029 at $m = N/128$, which is a relative error of .20 and .24. At $m = N/4$ the operator-level question-row error matches the end-to-end one: the ratio of the means is 1.00 [0.99, 1.01] and 0.97 [0.96, 0.98] with four anchor rows, and 0.92 on both models for one anchor row against the single-anchor run at $m = N/4$, so at this budget the perturbed inputs that a layer receives end to end raise its question-row error by at most about 3% with four anchor rows and 8 to 9% with one. The two measurements use different references: end to end, the error is taken against bf16 SDPA on each layer's own, already perturbed, inputs; at the operator level, against a float64 dense computation on the captures of the exact model. Neither captures the drift that the errors of all layers accumulate in the residual stream, which only the task scores reflect.

G.6 Budget scaling on longer histories

To extend Appendix G.5 beyond 16K keys, we ran the capture pipeline again on synthetic NIAH, VT and FWE inputs of 131,008 tokens for Llama-3.2-3B and 32,704 tokens for Qwen3-8B-Base, eight per task (Qwen3-8B-Base has 32,768 positions without YaRN, which would change the model), on one H200, and analysed every layer as it was computed instead of storing the tensors. Windows of 32 queries end at histories from 1K to 130,944 keys (Llama-3.2-3B) and to 32,640 keys (Qwen3-8B-Base) inside the filler of the same 24 inputs, so every length is measured on the same sequences; the final block of each input, which holds the question, is treated separately, as in Appendix G.5. The rerun reproduces the archived captures: on the twelve synthetic test sequences of 16K tokens, the queries, keys and values of heads 0 and $H_q/2$ agree bit for bit in every layer, and the window errors recomputed from them differ from the archived ones by at most 10^{-10}. Targets, fit and bootstrap follow Appendix G.5; the budget grid adds powers of two up to 16,384, and we interpolate the error linearly in $\log m$ between grid points (the interpolation of Appendix G.5 changes the exponents at a relative error of .05 by at most .03). Before the run we fixed how to read the 95% interval of α: an upper end below 0.9 means that the budget grows more slowly than N, an interval that contains 1 or lies above 0.9 that it grows in proportion to N.

At a relative L2 error of .05 with the zero-order remainder, $m_\varepsilon(N)$ grows as N^α with $\alpha = 0.72$ [0.69, 0.72] for the block-first anchor and 0.67 [0.66, 0.68] for four anchor rows on Llama-3.2-3B (4K to 128K), and 0.73 [0.69, 0.77] and 0.65 [0.63, 0.68] on Qwen3-8B-Base (4K to 32K) (Table 24, Figure 17(a)). These intervals, and those at the other targets of Table 24, lie below 0.9, so by the rule fixed before the run the budget grows more slowly than N on these inputs. The intervals describe only the resampling of the 24 inputs: fitted task by task, the exponents range from 0.65 to 0.72 on Llama-3.2-3B and from 0.57 to 0.83 on Qwen3-8B-Base.

The growth is not a single power law; the split at 16K below was not fixed before the run. On Llama-3.2-3B the slope over 1K to 16K is 0.96 [0.94, 0.97] with the block-first anchor, in proportion to N as on the 16K inputs of Appendix G.5, and 0.90 [0.89, 0.91] with four anchor rows; over 16K to 128K it is 0.50 [0.47, 0.52] and 0.45 [0.43, 0.48]. The budget stays near a fixed fraction of N up to 32K and then grows little: with one anchor row it is 12.8K keys at 32K, 15.9K at 64K and 17.2K at 128K (.39N, .24N and .13N), and with four anchor rows 7.2K, 9.8K and 9.8K (.22N, .15N and .075N). On Qwen3-8B-Base the slope over 1K to 16K inside these inputs is 0.75 [0.73, 0.76] with one anchor row, below the 0.85 [0.80, 0.90] of Appendix G.5; there the windows end next to the question and the tasks include LongBench, here the windows lie inside a long filler, and we have not separated the two differences.

At a fixed $m = 1024$ (Figure 17(c)), the error on the filler windows rises from 4K to 32K and changes little after that: on Llama-3.2-3B with one anchor row it is .078 at 4K, .138 at 16K, .181 at 32K and .175 at 128K. On the same inputs the error at 128K is 1.27 [1.22, 1.32] times the error at 16K with one anchor row and 1.30 [1.26, 1.35] times with four; on Qwen3-8B-Base the error at 32K is 1.42 [1.38, 1.47] and 1.37 [1.34, 1.40] times the error at 16K. We had fixed before the run that a ratio above 2 at 128K would limit every efficiency statement to 16K tokens. On the question rows of the final block, where each length has its own inputs, the error at $m = 1024$ keeps rising, from .14 at 4K to .25 at 32K and .27 at 128K on Llama-3.2-3B with one anchor row. To reach the per-layer question-row error of the end-to-end runs (.058 on Llama-3.2-3B and .048 on Qwen3-8B-Base, Appendix G.5), four anchor rows need m between .24N and .30N up to 16K tokens, .24N and .21N at 32K, .20N at 64K and .13N at 128K; a single block-first anchor does not reach this error within the grid at any length.

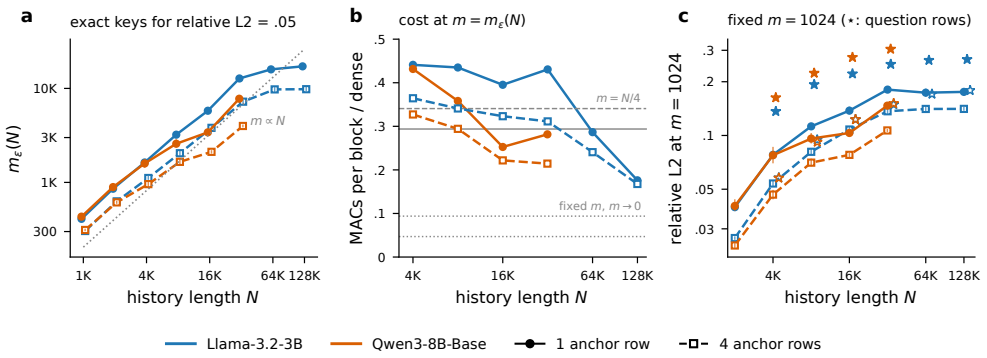

Figure 17: **Exact budget, its cost and the error at a fixed m on long synthetic inputs** (zero-order remainder; filler windows of the same 24 inputs at every length; one anchor row: filled circles, solid lines; four anchor rows: open squares, dashed lines). (a) $m_\varepsilon(N)$ at a relative L2 error of .05, from 1K to 128K on Llama-3.2-3B and to 32K on Qwen3-8B-Base, with 95% cluster-bootstrap intervals (dotted where the upper end is censored in some replicates); the dotted grey line has slope 1. (b) Multiply-accumulates of Equation (21) per block of 32 queries at $m = m_\varepsilon(N)$, divided by those of dense attention; the grey lines give the ratio at $m = N/4$ (Table 21; solid for one anchor row, dashed for four) and the share that every block pays for the whole history at any fixed m (dotted). (c) Relative L2 error at a fixed $m = 1024$ on the filler windows, and on the question rows of the final block (stars; different inputs at each length).

Table 24: **Exponent α of $m_\varepsilon(N) \propto N^\alpha$ on long synthetic inputs** (zero-order remainder; least-squares fit over the filler windows, 95% cluster-bootstrap intervals over the 24 inputs). The first four columns fit N from 4K to 128K on Llama-3.2-3B and from 4K to 32K on Qwen3-8B-Base, one column per error target; the last two split the relative-L2 fit at 16K, with the windows at 1K and 2K included. A dash marks a range that the Qwen3-8B-Base inputs do not reach. The intervals describe resampling of the inputs only.

		rel .05	rel .10	L2/Vmax .01	L2/Vmax .02	rel .05, 1K–16K	rel .05, 16K–128K
Llama-3.2-3B	1 anchor row	0.72 [0.69, 0.72]	0.67 [0.65, 0.69]	0.68 [0.66, 0.70]	0.66 [0.62, 0.70]	0.96 [0.94, 0.97]	0.50 [0.47, 0.52]
	4 anchor rows	0.67 [0.66, 0.68]	0.62 [0.60, 0.64]	0.63 [0.62, 0.64]	0.61 [0.58, 0.64]	0.90 [0.89, 0.91]	0.45 [0.43, 0.48]
Qwen3-8B-Base	1 anchor row	0.73 [0.69, 0.77]	0.61 [0.55, 0.66]	0.61 [0.54, 0.67]	0.38 [0.28, 0.48]	0.75 [0.73, 0.76]	–
	4 anchor rows	0.65 [0.63, 0.68]	0.53 [0.49, 0.57]	0.54 [0.49, 0.59]	0.35 [0.27, 0.42]	0.70 [0.68, 0.71]	–

Figure 17(b) turns these budgets into cost. At $m = m_\varepsilon(N)$ for a relative error of .05, the multiply-accumulates of Equation (21) per block are .44, .40, .43, .29 and .18 of dense attention at 4K, 16K, 32K, 64K and 128K with one anchor row on Llama-3.2-3B and .36, .32, .31, .24 and .17 with four; on Qwen3-8B-Base they are .43, .25 and .28 with one anchor row and .33, .22 and .21 with four at 4K, 16K and 32K. At $m = N/4$, as in our experiments, the ratio is .29 to .30 with one anchor row and .34 with four at every length (Table 21), and on Llama-3.2-3B with one anchor row $N/4$ keys miss this error on the filler windows up to 32K. On Llama-3.2-3B, holding the operator error fixed thus costs a nearly constant fraction of dense attention up to 32K and a falling fraction beyond; on Qwen3-8B-Base the fraction already falls between 4K and 16K. The fraction cannot fall below the terms that every block pays for the whole history, .047 and .094 of dense attention (the fixed-m share above), so prefill attention remains quadratic in the prompt length: a slower-growing budget lowers its constant at long histories and does not change its order.

These are operator-level errors on synthetic inputs, measured on two query heads per layer against a float64 dense computation on the captures of the exact model. They do not show that a given fraction of the history preserves task scores at these lengths, and we have not measured natural documents of this length.

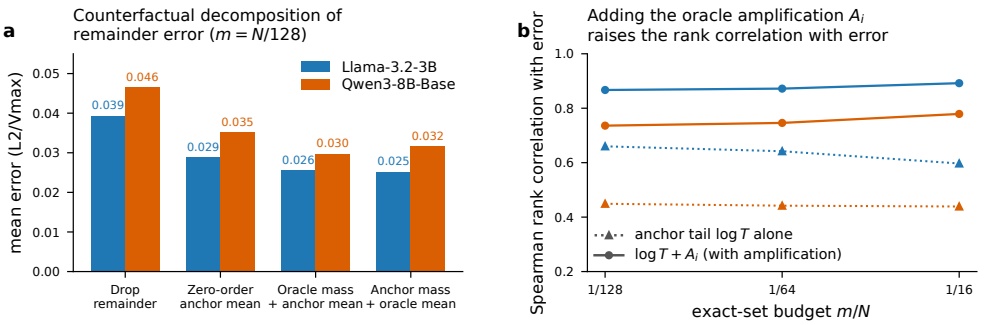

Figure 18: **Analytical decomposition of remainder error and tail amplification** (5120 **windows, development split).** (a) Counterfactual ablation of Proposition 2 components at $m = N/128$. Discarding the unselected keys entirely yields substantial error (.0393 / .0465). Replacing them with the zero-order anchor mean cuts error to .0288 / .0352. Supplying the oracle tail mass (π_i) or the oracle conditional mean (μ_i) shows how much each correction alone reduces the error; the oracle-mass error is exactly the first term of Proposition 2, but the two corrections do not split the error into its two terms. (b) Spearman rank correlation between error and tail metrics across budgets $m \in \{N/128, N/64, N/16\}$. While anchor tail mass $\log T$ correlates moderately with error (dotted), incorporating the within-block amplification factor A_i (the window mean of the log history-conditional share $\log T + A_i$, an oracle quantity that uses each query's full interaction with the history) raises the correlation to .87–.89 on Llama and .74–.78 on Qwen (solid), showing that residual error tracks the amplified tail mass more closely than anchor tail mass alone.

Table 25: **Counterfactual remainder decomposition across budgets.** Mean L2/Vmax error under four treatments of the unselected remainder, compared against the gate with an exact-exponential remainder on eight local directions, a projection-only reference that is neither a bound nor deployable.

Model	Budget m/N	Drop Remainder	Zero-Order Rem.	Oracle Mass	Oracle Mean	Exact Exp. Diag.
Llama-3.2-3B	1/128	.0393	.0288	.0256	.0252	.0174
	1/64	.0343	.0258	.0226	.0231	.0161
	1/16	.0235	.0188	.0161	.0176	.0125
Qwen3-8B-Base	1/128	.0465	.0352	.0298	.0316	.0215
	1/64	.0383	.0300	.0255	.0275	.0192
	1/16	.0234	.0196	.0168	.0185	.0137

H Remainder Error, Centering and Tail Amplification

H.1 Counterfactual evaluation of the error decomposition

In Section 5, Proposition 2 decomposed the output error into a value mean error term $\pi_i(\hat{\mu} - \mu_i)$ and a mass estimation error term $(\hat{\pi}_i - \pi_i)(\hat{\mu} - \nu_i)$. We evaluated counterfactuals on the development split that replace one remainder statistic at a time by its true value:

1. **Drop remainder**: Set unselected keys to zero ($\hat{\pi}_i = 0$). Error: .0393 (Llama) and .0465 (Qwen) at $m = N/128$.
2. **Zero-order anchor mean**: Our standard operator ($\hat{\mu} = S_1/S_0$, $\hat{\pi}_i$ estimated via anchor decomposition). Error: .0288 (Llama) and .0352 (Qwen).
3. **Oracle remainder mass**: Supply the true tail mass π_i while keeping the anchor mean $\hat{\mu}$. Error drops to .0256 (Llama) and .0298 (Qwen).
4. **Oracle remainder mean**: Supply the true conditional value mean μ_i while keeping the anchor mass estimate $\hat{\pi}_i$. Error drops to .0252 (Llama) and .0316 (Qwen).

Supplying the true mass leaves exactly the first term of Proposition 2, so the oracle-mass column is the mean of $\|\pi_i(\hat{\mu} - \mu_i)\|_2 / V_{\max}$. Supplying the true mean does not isolate the second term: it leaves $(\hat{\pi}_i - \pi_i)(\mu_i - \nu_i)$, which differs from $(\hat{\pi}_i - \pi_i)(\hat{\mu} - \nu_i)$ by $(\hat{\pi}_i - \pi_i)(\hat{\mu} - \mu_i)$. As Figure 18(a) and Table 25 show, correcting either statistic alone closes only part of the gap to the exact-exponential diagnostic. The two counterfactuals do not split the error into the two terms, which are not orthogonal and can partly cancel (Appendix B), so a correction can raise the error of an individual row.

On the development split at $m = N/128$ the two terms are of similar size, with mean $\|a_i\|_2/V_{\max}$ of .0256 (Llama) and .0298 (Qwen), which is the oracle-mass column of Table 25, against mean $\|b_i\|_2/V_{\max}$ of .0260 and .0317, and often point in partly opposite directions: over the rows after the first of each block their cosine has median $-.192$ and $-.256$ and is negative on 72.9% and 78.5% of rows, so supplying the true mass raises the error on 20.5% and 25.2% of rows even though it lowers the mean error; on the test split the corresponding values are .0238 and .0282 against .0244 and .0304, with median cosine $-.179$ and $-.246$, negative on 71.0% and 77.8% of rows, and the true mass raises the error on 19.9% and 24.4% of rows. Both terms are computed directly in float64 from the captured queries, keys and values.

H.2 Predicting tail mass amplification

At $m = N/128$ the anchor-gated operator assigns the remainder a mass $\hat{\pi}_i$ whose window mean has median .037 over development windows, whereas the true remainder mass π_i under the full causal softmax has median window mean .097 in Llama and .206 in Qwen. The anchor tail mass $T = \sum_{j \notin \mathcal{H}} p_a(j)$ itself has median .073 and .113.

As derived in Appendix B, the remainder's share of the history-conditional distribution is $\pi_i^{\text{hist}} = Te^{A_i}$ (Equation (14); median .132 and .288), and its full causal mass is $\pi_i = h_i Te^{A_i}$ (Equation (15)), where $h_i < 1$ is the full-softmax mass on the history. In Figure 18(b), we compare this share with the output error. On the development split, the anchor tail mass $\log T$ alone correlates with error at .660 (Llama) and .449 (Qwen) at $m = N/128$; adding the log-amplification factor A_i, that is, ranking windows by the mean of $\log \pi_i^{\text{hist}} = \log T + A_i$, increases the Spearman correlation to .867 and .736 at $m = N/128$, and to .892 and .779 at $m = N/16$.

Ranking windows instead by the mean over the rows of $\log \pi_i$, with the full causal mass $\pi_i = h_i\, \pi_i^{\text{hist}}$ of Equation (15) in place of the history-conditional share, gives .864 and .715 at $m = N/128$ and .881 and .754 at $m = N/16$; on the test split the history-conditional share gives .868 and .720 at $m = N/128$ and .878 and .759 at $m = N/16$, and the full mass .863 and .702 and .866 and .739.

The window statistic is the mean over the rows of $\log T + A_i$, the log of the geometric mean of π_i^{hist} over the window rather than the log of its arithmetic mean. It is also an oracle quantity: A_i uses each query's full interaction with the history, so $\log T + A_i$ is the log of the true history-conditional remainder share, and the rise in correlation shows what the error tracks, not what the anchor row can predict. Estimates that do not use the true interaction follow.

Furthermore, consider the closed-form Gaussian (second-order) estimate of A_i,

$$\hat{A}_i \;=\; u_i^\top \bar{w}_{\mathcal{O}} + \frac{1}{2}\, u_i^\top\, (\Sigma_{\mathcal{O}} - J)\, u_i, \tag{22}$$

where $\bar{w}_{\mathcal{O}} = \sum_{j \in \mathcal{O}} \hat{q}(j)\, w_j$ and $\Sigma_{\mathcal{O}} = \sum_{j \in \mathcal{O}} \hat{q}(j)\, (w_j - \bar{w}_{\mathcal{O}})(w_j - \bar{w}_{\mathcal{O}})^\top$ are the mean and the centered covariance of the w_j under the anchor base measure $\hat{q} = p_a/T$ of the remainder, and J (Appendix B.2) is their tilted covariance over the whole history, which needs no centering because the w_j have tilted mean zero. This estimate correlates with the true history-conditional amplification factor A_i at $\rho = .480$ on Llama and .707 on Qwen (test split, $m = N/128$, Spearman correlation over windows of the within-window means). $\hat{A}_i$ truncates the cumulant expansions of both log-moment generating functions in $A_i = \delta_i - \log M_i$ after the second cumulant. Scaling the displacement to λu_i gives

$$A_i(\lambda) \;=\; \lambda\, u_i^\top \bar{w}_{\mathcal{O}} + \frac{\lambda^2}{2}\, u_i^\top\, (\Sigma_{\mathcal{O}} - J)\, u_i + R_{3,i}(\lambda), \qquad R_{3,i}(\lambda) \;=\; \frac{\lambda^3}{6}\, (\kappa_{3,\mathcal{O}} - \kappa_3) + O(\lambda^4),$$

where $\kappa_{3,\mathcal{O}}$ and κ_3 are the third cumulants of $u_i^\top w_j$ under $\hat{q}$ and p_a; on a finite key set $A_i(\lambda)$ is analytic in λ, so the expansion holds for small λ. The estimate sets $\lambda = 1$ and drops $R_{3,i}(1)$, which we do not control, so it is justified only for small displacements, and the correlation above measures ranking, not calibration.

The estimate is not calibrated. Over test rows at $m = N/128$, excluding the block-first row, the least-squares slope of A_i on $\hat{A}_i$ is .195 and .349 (95% intervals $[.181, .207]$ and $[.310, .389]$), where a calibrated estimate would give 1; the median of $|A_i - \hat{A}_i|$ is .258 and .302 against a median $|A_i|$ of .570 and .716, and the error exceeds $|A_i|$ itself on 36.7% and 30.9% of rows. The error grows with the gap between the third cumulants: its median is .035 and .049 in the lowest and 1.58 and 1.47 in

the highest tertile of $|\kappa_{3,\mathcal{O}} - \kappa_3|$.[1] For displacements scaled by $\lambda \le 1/4$ it falls as λ^s with $s = 2.97$ and 2.98, as the cumulant expansion predicts (median over the rows of 64 test windows per model, stratified by layer and effective sample size). At the displacements that occur in our data, however, the expansion is not in its asymptotic range: the third-cumulant term $(\kappa_{3,\mathcal{O}} - \kappa_3)/6$ is of the size of the error (median ratio of the error to it .75 in both models), but adding it to $\hat{A}_i$ leaves a median .60 and .87 of the error and enlarges it on 33.8% and 45.3% of the same rows; and over all test rows the first-order value term $\mathrm{Cov}_{\hat{q}}(v_j, w_j)\, u_i$ of Appendix B.6 leaves a median fraction .61 and .71 of $\|\mu_i - \hat{\mu}\|_2$ unexplained. The Gaussian estimate should therefore be read as a ranking device rather than an approximation with a controlled error. Figure 19(c) shows A_i against $\hat{A}_i$.

On the test split at $m = N/128$, the first-order term $u_i^\top \bar{w}_{\mathcal{O}}$ alone correlates with A_i at .574 and .605, and ranking windows by $\log T + \hat{A}_i$ instead of $\log T + A_i$ gives error correlations of .645 and .556, against .660 and .445 for $\log T$ alone. Within-block tail drift is therefore only partly predictable from these moments, and on Llama the Gaussian estimate does not improve on $\log T$ for ranking windows by error.

The linear term $u_i^\top \bar{w}_{\mathcal{O}}$ of $\hat{A}_i$ equals the mean omitted interaction $\bar{x}_i$ of Appendix B.6. It contributes to the mass error only because the per-query factor is centered at the full-history mean $\bar{k}$; centering it at the remainder mean $\bar{k}_{\mathcal{O}}$ removes that contribution without enlarging the state (Proposition 5 and the paragraph after it). All counterfactuals and correlations above use the full-history center.

H.3 Bounds and estimates against measured values

Figure 19 sets the bounds of Appendix B and the estimates of this appendix against the measured quantities on all 5120 test windows. Panel (a) concerns the restricted basis of Section 4 (Appendix D.2).

Panel (b) sets the bound of Proposition 4 against the actual error. Over the 317,440 rows of both models at $m = N/128$ and $N/16$ the bound is never violated; its ratio to the actual error has 90th percentiles of 45 and 62 at $m = N/128$ and 33 and 44 at $m = N/16$ (medians in Appendix B.6), and it grows with $\chi^2(P_{i,\mathcal{O}} \,\|\, \hat{q})$, with Spearman ρ of .75 and .74 at $m = N/128$ and median ratios of 7.1 and 8.1 in the lowest and 29.1 and 36.2 in the highest tertile of χ^2. The mass term of the bound is the exact norm of the second term of Proposition 2, so the slack comes from the triangle inequality and from the Cauchy–Schwarz step Equation (13). Panel (c, left) sets the log-odds distortion δ_i of Equation (16) against two estimates from the tilted moments of the remainder keys: the linear $\bar{x}_i = u_i^\top \bar{w}_{\mathcal{O}}$ and the Gaussian $\hat{\delta}_i = \bar{x}_i + \frac{1}{2}\, u_i^\top \Sigma_{\mathcal{O}}\, u_i$, which is the part of Equation (22) that estimates δ_i. At $m = N/128$ the Gaussian estimate ranks δ_i with Spearman ρ of .973 (Llama) and .965 (Qwen) over rows, and the least-squares slope of δ_i on $\hat{\delta}_i$ is 1.02 and 1.04; the linear estimate reaches .644 and .759. Because Equation (16) fixes the mass error once δ_i and $\pi_i(1 - \hat{\pi}_i)$ are known, the mass error that it predicts from $\hat{\delta}_i$ keeps a partial Spearman correlation of .94 and .92 with the actual one when $\pi_i(1 - \hat{\pi}_i)$ is held fixed, against .16 and .26 with $\bar{x}_i$ in place of $\hat{\delta}_i$. Panel (c, right) sets A_i against $\hat{A}_i$. The Gaussian estimate is not calibrated: regressing A_i on $\hat{A}_i$ over windows gives slopes .154 and .350 (95% intervals $[.130, .175]$ and $[.301, .399]$), and the median absolute error is .359 and .360. The two panels differ because $A_i = \delta_i - \log M_i$ also contains the log-normalizer of the whole history: in $A_i - \hat{A}_i = (\delta_i - \hat{\delta}_i) - (\log M_i - \frac{1}{2}\, u_i^\top J u_i)$ the second term has median magnitude .228 and .266 over rows against .066 and .112 for the first and is the larger of the two on 73.7% and 68.5% of rows, and the least-squares slope of $\log M_i$ on $\frac{1}{2}\, u_i^\top J u_i$ is .376 and .597. The second-order expansion thus tracks the log-normalizer of the remainder closely and that of the whole history poorly.

Panel (a) shows why Appendix B.3 bounds the captured variance rather than the principal angles. The Davis–Kahan ratio $\|J_{\mathcal{O}}\|_2/\gamma_8$ exceeds one in 100% (Llama) and 99.8% (Qwen) of test windows at $m = N/128$, so the $\sin\Theta$ reference $\min(1, \cdot)$ of the panel sits at its trivial value, and the ratio is nearly unrelated to the extra diagnostic error of the restricted basis over the full-history basis (Spearman ρ of .07 and .06).

[1] Across the tertiles of $\max_{j \in \mathcal{O}} |x_{ij}|$ the median error rises from .048 and .064 to 1.49 and 1.22; across those of $\|u_i\|_2$ it rises from .216 to .400 on Qwen but is not monotone on Llama (.239, .181, .376).

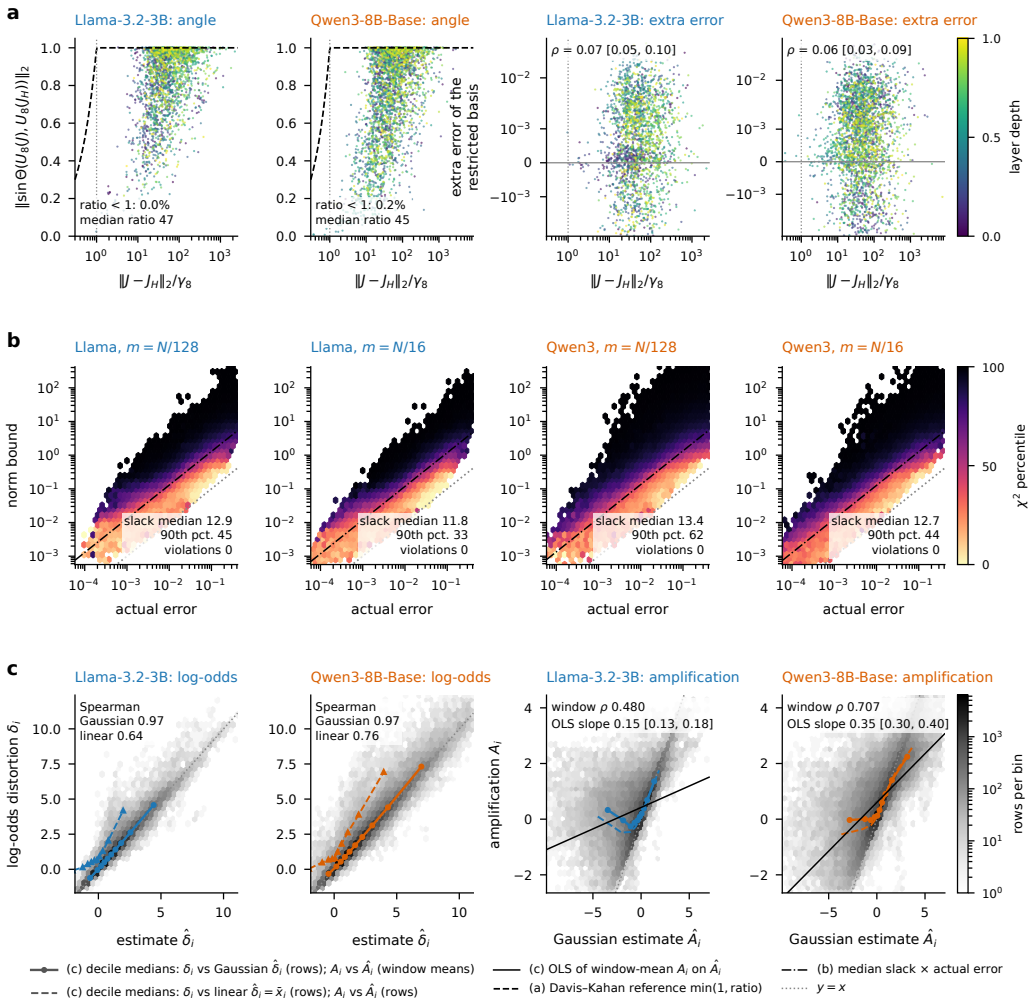

Figure 19: **Bounds and estimates against measured values (test split, rows after the first of each block).** (a) Davis–Kahan ratio $\|J - J_{\mathcal{H}}\|_2/\gamma_8$ at $m = N/128$ against $\|\sin\Theta\|_2$ between the top-8 subspaces of J and $J_{\mathcal{H}}$ (dashed: the sin-theta reference $\min(1, \cdot)$), and against the extra diagnostic error of the restricted basis over the full-history basis; one point per window, colored by relative layer depth. (b) Actual error of the zero-order remainder against the bound of Proposition 4, per row, at $m = N/128$ and $N/16$; color is the median percentile of $\chi^2(P_{i,\mathcal{O}} \,\|\, \hat{q})$ in each hexagon; dotted: equality; dash-dotted: the median ratio. (c) Left: log-odds distortion δ_i against its Gaussian estimate $\hat{\delta}_i$ per row, with decile medians for the Gaussian (solid) and the linear (dashed) estimate. Right: amplification A_i against its Gaussian estimate $\hat{A}_i$ per row, with decile medians over windows (solid) and rows (dashed) and the least-squares line over windows. The axes of (b) omit at most 0.63% and those of (c) at most 2.2% of the rows; all statistics use every row.

H.4 Remainder ablations on the test split

Tables 26 to 28 evaluate the zero-order remainder of Section 5 on all 5120 test windows, with the same selector and exact sets as Section 5 and the statistics of Appendices A and A.1; the column n gives the number of the 20 test items whose difference has the sign of the estimate (Appendix A.1). Table 26 replaces the full-history center $\bar{k}$ of the per-query factor by the remainder mean $\bar{k}_{\mathcal{O}}$ of Appendix B.6, which keeps the state at $d + d_v + 1$ numbers, and adds two diagnostics: the first-order moment $M_1 = \sum_{j\in\mathcal{O}} \hat{q}(j)(k_j - \bar{k}_{\mathcal{O}})v_j^\top$, which corrects the value term $\mathrm{Cov}_{\hat{q}}(v_j, w_j)u_i$ of Appendix B.6 while the new center corrects the mass term, and the Gaussian mass factor of Equation (22). Centering at $\bar{k}_{\mathcal{O}}$ lowers the error at every exact-set size m and with four anchor rows, on both models, in at least 19 of the 20 items. With this center the zero-order factor is a lower bound on the remainder

Table 26: **Centering the remainder (test split).** Mean L2/Vmax error of the zero-order remainder centered at the full-history mean $\bar{k}$ and at the remainder mean $\bar{k}_{\mathcal{O}}$, the paired difference with its 95% cluster-bootstrap interval and the number n of the 20 items on its side, and three diagnostics on top of $\bar{k}_{\mathcal{O}}$: the first-order moment M_1 ($d \times d_v$ numbers per block, about 2 times the arithmetic of dense attention to build), its projection on the eight restricted directions of Section 4 (1,024 numbers, about a quarter of dense), and M_1 with the Gaussian mass factor (another $d(d+1)/2$ numbers, about the dense cost). Four anchors: the selector of Section 6. Only the two centers keep the $d + d_v + 1$ numbers of the state of Section 5.

Model	m/N	$\bar{k}$	$\bar{k}_{\mathcal{O}}$	$\bar{k}_{\mathcal{O}} - \bar{k}$ [95% CI]	n	$+M_1$	M_1 effect [95% CI]	compact M_1	M_1 + Gauss.
Llama-3.2-3B	1/128	.0272	.0261	$-.0011$ $[-.0012, -.0010]$	19	.0235	$-.0026$ $[-.0027, -.0024]$	.0246	.0166
	1/64	.0244	.0230	$-.0013$ $[-.0015, -.0012]$	20	.0210	$-.0021$ $[-.0022, -.0019]$	.0219	.0145
	1/32	.0212	.0197	$-.0015$ $[-.0017, -.0013]$	20	.0181	$-.0016$ $[-.0017, -.0015]$	.0188	.0120
	1/16	.0175	.0159	$-.0015$ $[-.0017, -.0014]$	20	.0147	$-.0012$ $[-.0013, -.0011]$	.0153	.0094
	1/8	.0133	.0119	$-.0014$ $[-.0016, -.0013]$	20	.0110	$-.0008$ $[-.0009, -.0008]$	.0114	.0070
	1/4	.0088	.0076	$-.0012$ $[-.0013, -.0011]$	20	.0072	$-.0005$ $[-.0005, -.0004]$	.0074	.0045
	1/16 (4 anchors)	.0111	.0104	$-.0007$ $[-.0008, -.0006]$	19	.0096	$-.0009$ $[-.0009, -.0008]$	–	–
	1/4 (4 anchors)	.0048	.0043	$-.0005$ $[-.0006, -.0005]$	20	.0040	$-.0003$ $[-.0003, -.0003]$	–	–
Qwen3-8B-Base	1/128	.0338	.0313	$-.0025$ $[-.0028, -.0023]$	20	.0288	$-.0025$ $[-.0027, -.0023]$	.0301	.0241
	1/64	.0288	.0265	$-.0023$ $[-.0025, -.0021]$	20	.0246	$-.0019$ $[-.0020, -.0017]$	.0257	.0201
	1/32	.0240	.0220	$-.0020$ $[-.0022, -.0019]$	20	.0206	$-.0014$ $[-.0015, -.0013]$	.0214	.0165
	1/16	.0192	.0175	$-.0017$ $[-.0018, -.0016]$	20	.0165	$-.0010$ $[-.0011, -.0010]$	.0171	.0133
	1/8	.0144	.0131	$-.0013$ $[-.0014, -.0013]$	20	.0124	$-.0007$ $[-.0007, -.0007]$	.0128	.0103
	1/4	.0094	.0085	$-.0009$ $[-.0009, -.0009]$	20	.0080	$-.0004$ $[-.0004, -.0004]$	.0083	.0072
	1/16 (4 anchors)	.0112	.0102	$-.0010$ $[-.0011, -.0009]$	20	.0094	$-.0007$ $[-.0008, -.0007]$	–	–
	1/4 (4 anchors)	.0042	.0037	$-.0005$ $[-.0005, -.0005]$	20	.0034	$-.0003$ $[-.0003, -.0003]$	–	–

Table 27: **The remainder against dropping under four error measures** (test split, $m = N/128$, same exact set). L2/Vmax is the measure of the paper; relative L2 divides by $\|o_i\|_2$, value-RMS by the root-mean-square value norm on the prefix, absolute is $\|\hat{o}_i - o_i\|_2$. Means are cluster means; the column $\bar{k}_{\mathcal{O}}$ is the recentered remainder of Table 26, lower than the zero-order remainder under every measure; q90 and q99 are row quantiles (drop / zero-order); the last column restricts the paired difference to the question rows of the late windows of the synthetic tasks (12 items). The relative error is large on rows with a small $\|o_i\|_2$; n is the number of the 20 items on the side of the estimate. At every budget from $N/128$ to $N/4$ and with four anchor rows the zero-order remainder is below dropping under all four measures, in all 20 items, and also on the question rows. Only the drop, zero-order and $\bar{k}_{\mathcal{O}}$ arms were recomputed under these measures (the equal-cost drop arms of Table 28 were recorded per window but are not aggregated here); the other operators were stored in L2/Vmax only, so the orderings of the gate against P_2 (Figure 3(d)) and against per-query selection (Table 2) are checked in L2/Vmax only.

Model	Measure	drop	zero-order	difference [95% CI]	n	$\bar{k}_{\mathcal{O}}$	q90	q99	question rows [95% CI]
Llama-3.2-3B	L2/Vmax	.0377	.0272	$-.0105$ $[-.0110, -.0100]$	20	.0261	.0950 / .0702	.2504 / .2223	$-.0051$ $[-.0053, -.0049]$
	relative L2	.3325	.2279	$-.1046$ $[-.1081, -.1011]$	20	.2155	.7899 / .5800	1.2127 / 1.0587	$-.0610$ $[-.0632, -.0586]$
	value-RMS	.0659	.0471	$-.0188$ $[-.0196, -.0180]$	20	.0450	.1683 / .1238	.3966 / .3584	$-.0095$ $[-.0098, -.0092]$
	absolute	.233	.168	$-.066$ $[-.069, -.063]$	20	.161	.615 / .445	1.477 / 1.347	$-.031$ $[-.032, -.030]$
Qwen3-8B-Base	L2/Vmax	.0450	.0338	$-.0112$ $[-.0115, -.0109]$	20	.0313	.1131 / .0844	.2832 / .2557	$-.0054$ $[-.0056, -.0052]$
	relative L2	.3620	.2742	$-.0878$ $[-.0900, -.0857]$	20	.2511	.8867 / .7441	1.3780 / 1.1767	$-.0440$ $[-.0454, -.0425]$
	value-RMS	.1012	.0768	$-.0244$ $[-.0250, -.0238]$	20	.0713	.2562 / .1990	.5901 / .5546	$-.0126$ $[-.0129, -.0122]$
	absolute	.982	.744	$-.238$ $[-.249, -.229]$	20	.704	2.299 / 1.744	9.982 / 8.440	$-.129$ $[-.137, -.119]$

mass by Jensen's inequality ($\rho_i \geq 0$); the mean log-mass gap that remains at $m = N/128$ is 1.46 and 1.77 nats, against a mean omitted interaction $\bar{x}_i$ of -0.16 and $+0.46$, so most of the remaining mass error comes from the spread of the remainder logits rather than from their mean. The first-order moment and the Gaussian factor lower the error further but cost more than dense attention to build; an eight-direction compact version of the moment recovers about half of its gain at about a quarter of the dense cost.

Table 27 repeats the comparison of the zero-order remainder with dropping under three further error measures and on the question rows; the ordering does not change under any of them, and under each measure it holds in all 20 items at every exact-set size m and with four anchor rows. Table 28 gives the equal-cost comparison of Section 5 at every exact-set size m. FWE items carry a large part of the advantage of the remainder there: leaving FWE out moves every estimate toward dropping, and under the MAC count on Qwen the estimate then favors dropping at every budget of the single-anchor selector.

Table 28: **The remainder against dropping at equal cost** (test split). Each cell is the paired difference, zero-order remainder at m minus drop at m', with its 95% cluster-bootstrap interval and the number n of the 20 items on its side; m' spends the construction cost of the remainder on more exact keys, in the key-pair account of Appendix G (kp) or in the MAC count of Equation (21) (MAC), and m'/m is its mean over the Llama windows (within .01 on Qwen). The drop arm uses the same selection scores. †: fails the criterion of Appendix A.1 (t interval excludes zero and sign-flip $p < .05$). Last two columns: the estimate with FWE left out.

				Llama		Qwen		without FWE	
Selector	m/N	count	m'/m	difference [95% CI]	n	difference [95% CI]	n	Llama	Qwen
1 anchor	1/128	kp	3.01	$-.0028$ $[-.0033, -.0023]$	14	$+.0011$ $[+.0008, +.0015]$	15†	$-.0007$	$+.0036$
		MAC	5.03	$+.0011$ $[+.0006, +.0017]$	14†	$+.0063$ $[+.0057, +.0069]$	16	$+.0029$	$+.0082$
	1/64	kp	2.00	$-.0034$ $[-.0038, -.0030]$	15	$-.0009$ $[-.0012, -.0006]$	7†	$-.0017$	$+.0011$
		MAC	3.00	$-.0003$ $[-.0008, +.0002]$	10†	$+.0031$ $[+.0027, +.0035]$	16	$+.0012$	$+.0046$
	1/32	kp	1.50	$-.0035$ $[-.0039, -.0030]$	18	$-.0017$ $[-.0018, -.0015]$	11	$-.0020$	$-.0001$
		MAC	1.99	$-.0011$ $[-.0016, -.0007]$	11†	$+.0010$ $[+.0008, +.0012]$	16†	$+.0000$	$+.0022$
	1/16	kp	1.25	$-.0029$ $[-.0032, -.0026]$	20	$-.0017$ $[-.0018, -.0016]$	20	$-.0019$	$-.0005$
		MAC	1.49	$-.0014$ $[-.0018, -.0011]$	15	$-.0002$ $[-.0003, -.0000]$	5†	$-.0005$	$+.0009$
	1/8	kp	1.13	$-.0020$ $[-.0022, -.0018]$	20	$-.0013$ $[-.0013, -.0012]$	20	$-.0013$	$-.0004$
		MAC	1.24	$-.0012$ $[-.0014, -.0009]$	17	$-.0004$ $[-.0005, -.0003]$	6†	$-.0006$	$+.0003$
	1/4	kp	1.06	$-.0010$ $[-.0011, -.0009]$	20	$-.0007$ $[-.0007, -.0007]$	19	$-.0007$	$-.0002$
		MAC	1.11	$-.0006$ $[-.0008, -.0005]$	18	$-.0003$ $[-.0004, -.0002]$	6†	$-.0003$	$+.0002$
4 anchors	1/16	kp	1.25	$-.0038$ $[-.0041, -.0035]$	20	$-.0026$ $[-.0026, -.0025]$	20	$-.0028$	$-.0012$
		MAC	1.49	$-.0026$ $[-.0029, -.0023]$	20	$-.0012$ $[-.0013, -.0011]$	10†	$-.0017$	$+.0000$
	1/4	kp	1.06	$-.0016$ $[-.0017, -.0014]$	20	$-.0012$ $[-.0012, -.0011]$	20	$-.0012$	$-.0005$
		MAC	1.11	$-.0013$ $[-.0014, -.0012]$	20	$-.0009$ $[-.0009, -.0009]$	20	$-.0009$	$-.0003$

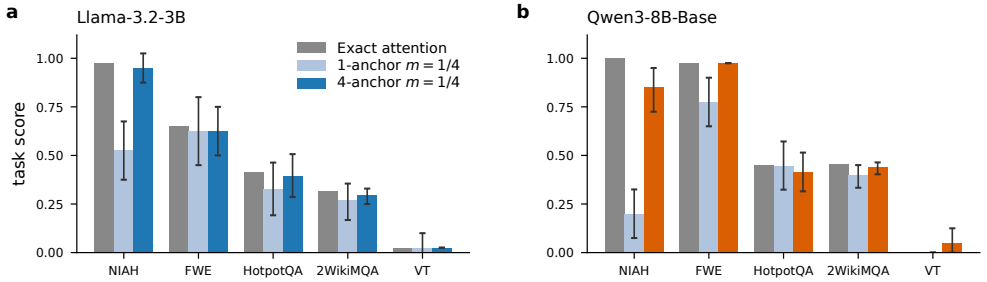

Figure 20: **Per-task end-to-end evaluation across five benchmark tasks (200 prompts of up to 16K tokens per model, full models).** Comparison among dense bf16 SDPA (Exact attention), the 1-anchor baseline at $m = N/4$, and the gated representation with the 4-anchor selector at $m = N/4$. (a) Llama-3.2-3B. (b) Qwen3-8B-Base. Bars on the two gated arms are 95% intervals of the paired per-task difference against exact attention, not adjusted for the five tasks, drawn around the arm's score. A single anchor degrades sharply on some tasks (e.g., NIAH falls from .975 $\rightarrow$ .525 on Llama and 1.000 $\rightarrow$ .200 on Qwen). The 4-anchor selector recovers most of this loss, but task-specific gaps to exact attention remain, the largest being NIAH on Qwen (.850 against 1.000), the only per-task interval of the 4-anchor arm that excludes zero.

I Full-Model Evaluation: Task Breakdown and Causal Selectors

Section 6 reported overall task scores and relative question errors for the end-to-end evaluation. Here we provide the detailed breakdown across all five evaluated benchmarks:

- **Needle In A Haystack (NIAH)** (Hsieh et al., 2024): Eight key-number statements are placed at different depths of a 16K-token book text, and the question asks for the number of one key.
- **Frequent Words Extraction (FWE)** (Hsieh et al., 2024): A text of 40 common English words separated by dots and drawn with skewed frequencies; the answer is the three most frequent words, which are separated in count from the fourth.
- **HotpotQA** (Bai et al., 2024): Multi-hop question answering across concatenated documents requiring joint entity linking.
- **2WikiMQA** (Bai et al., 2024): Multi-hop reasoning requiring synthesis across Wikipedia articles.

- **Variable Tracking (VT)** (Hsieh et al., 2024): Twelve chains of four variable definitions are scattered through book text; the question asks for the number that the last variable of one chain resolves to after three references.

Figure 20 and Table 29 report individual task scores. Under a single block-first anchor (1 anchor, $m = N/4$), retrieval-heavy tasks degrade substantially: NIAH drops from .975 to .525 on Llama, and from 1.000 to .200 on Qwen. For the three synthetic tasks the question starts at token 16,357 to 16,370, after the first query of the last block (16,352), so the block-first anchor is a filler token rather than a question token. This is consistent with a coverage problem, but positional coincidence alone does not identify the cause of the task loss.

The 4-anchor selector scores every history key under four query rows of the block (positions 0, 10, 21 and 31) and keeps the m keys with the largest of the four tilted masses, equivalently the union of the keys whose tilted mass under some of the four rows exceeds a common threshold (Section 6). For the synthetic tasks, rows 21 and 31 of the last block (and row 10 when the question starts early enough) fall inside the question, so the selector sees question tokens there, which is also why it relies on looking ahead within the block. It recovers most of the single-anchor loss but not all of it: NIAH recovers to .950 (vs .975) on Llama and .850 (vs 1.000) on Qwen, where 6 of the 40 inputs that exact attention answers correctly now fail (four of them return the number of a different needle in the same input and two return 1234567, and their mean question-row relative error, .063, is the same to three decimals as that of the 34 inputs answered correctly, so this error does not flag them); FWE reaches .625 (vs .650) on Llama and .975 (identical to exact) on Qwen; HotpotQA reaches .395 (vs .416) on Llama and .415 (vs .448) on Qwen; 2WikiMQA reaches .295 (vs .314) on Llama and .439 (vs .458) on Qwen. Variable Tracking baseline scores are near zero across all methods (.025 on Llama, .000 on Qwen), reflecting intrinsic model difficulty rather than approximation degradation.

All arms of the run and non-inferiority. The run behind the first three arms of Table 1 had eight arms, all on the same 200 inputs per model and with the same code, and Table 29 lists all of them. Below, a point is 0.01 of the task score. The lower end of each 95% interval is the one-sided 97.5% lower bound of the paired difference, so the four-anchor arm at $m = N/4$ would pass a non-inferiority test at that level only with margins above 5.6 and 6.4 points, or above 7.0 and 8.9 points without VT, on which exact attention scores .025 and .000. No margin was fixed before the run, and we do not claim non-inferiority. Per task, only NIAH on Qwen3-8B-Base has an interval that excludes zero under this arm. Its sign test, $p = .031$, is not adjusted; a Holm correction over the three tasks scored by accuracy gives .094, and a Holm correction of the five per-task bootstrap p values gives .020. The sign test is the appropriate one at a ceiling: with exact attention at or near 1.0, resampling the observed paired differences cannot produce discordant pairs in exact attention's disfavor, so the smaller bootstrap value is not evidence that the sign test is conservative.

Question rows. The mean question-row error hides variation across heads: for the four-anchor arm at $m = N/4$, over layers and heads its median is .048 and its 90th percentile .111 on Llama-3.2-3B (.032 and .097 on Qwen3-8B-Base), and in every input at least one question row of some head has a relative error above 1, that is, an error larger than the norm of that row's exact output; the same holds for every gated arm. On LongBench the question rows are the last 48 tokens of the prompt, and the last $N \bmod 32$ of them fall in the final partial block, which every arm computes with exact attention (Appendix A): 63 and 58 of the 80 LongBench inputs have such rows, on average 11.8 and 10.7 of the 48 over all 80 inputs. The logs keep only the mean and the maximum of the error over the question rows of each head, so we bound the mean over the gated rows from above by setting the error of the exact rows to zero. The ratio between the single-anchor and the four-anchor arm at $m = N/4$ is then 2.35 and 3.27 over the gated question rows, against 2.35 and 3.28 over all of them, and setting the exact rows to the bf16 error of their layer instead gives 2.36 and 3.28.

Budget and remainder. Against the four-anchor arm at $m = N/4$, keeping $m = N/8$ changes the total by -4.3 [-9.6, $+0.4$] points on Llama-3.2-3B and -11.7 [-15.8, -7.7] points on Qwen3-8B-Base. At $m = N/8$ the arm that drops the keys outside the exact set pairs with the remainder arm on the same set (Table 30); the remainder lowers the question-row error on every input of both models but does not raise the total score measurably, and on FWE it scores lower than dropping on both models (.525 against .650 and .950 against 1.000).

Tail-mass arms. A tail-mass rule picks for each block the smallest of the six budgets from $N/128$ to $N/4$ at which each of the four anchor rows leaves at most a fraction τ of its tilted mass outside

the exact set, and computes the block densely when no budget qualifies. With $\tau = .05$ it scored lower than the fixed quarter budget at a higher cost: .429 and .488 at cost .423 and .431, -2.9 [-8.0, $+2.2$] and -5.8 [-10.3, -1.0] points against $m = N/4$, with 26% and 29% of all blocks computed densely. With $\tau = .1$ it scores .297 and .372 at cost .282 and .307.

An earlier single-anchor run. An earlier run on the same inputs used a single block-first anchor row and an earlier version of the gate code, which is not part of the current reference implementation (its hash and outputs are archived, and every statistic below is recomputed from those outputs); it did not log question-row errors. At the same m, keeping the zero-order remainder instead of dropping the unselected keys changes the total score by $+3.6$ [$+0.1$, $+7.1$] points at $m = N/128$ and $+11.8$ [$+6.9$, $+16.9$] points at $m = N/32$ on Llama-3.2-3B, and by -1.3 [-4.9, $+1.7$] and -1.3 [-4.5, $+1.9$] points on Qwen3-8B-Base, at key-pair cost .058 and .082 against .043 and .066 for dropping (Table 30); all four arms stay more than 23 points below exact attention. The lower end of the first interval is so close to zero that its sign depends on the order in which the bootstrap draws the tasks, so we do not read it as a significant gain. The exact-attention arm of that run differs from the one in Table 1 on two of the 200 Llama-3.2-3B generations and in one score. That run also tried $m = N/8$ with the remainder, three tail-mass thresholds ($\tau \in \{.05, .1, .2\}$) and three error targets ($\epsilon \in \{1\%, 3\%, 5\%\}$) for an auxiliary budget predictor restricted to eight features that the gate computes. Each of these arms scored at least 15 points below exact attention except the predictor at $\epsilon = 1\%$, which scored .495 and .576 against .481 and .576 but computed 60% and 55% of the blocks densely, at cost .656 and .623.

Prespecified paired comparisons. The drop arms and the strictly causal selectors of Table 1 come from two later runs on the same 200 inputs per model (Table 31 lists all their arms). A note written before these runs fixed the margin of 5 points, the primary endpoint E4 (the equal-weight mean over NIAH, FWE, HotpotQA and 2WikiMQA; VT leaves it because exact attention scores .025 and .000 on it), the comparisons and how each result would be read, and the analysis code; each run recorded the SHA-256 of the note and of the code before its first unit, and the supplementary material contains both. The new arms are paired with the exact, single-anchor and four-anchor arms of the first run on the same inputs. They ran with the same software on H100 GPUs (the first run had used H100 and H200 GPUs), and before any pairing we reran archived arms there. On the first new GPU, exact attention and the four-anchor arm reproduced the archived prediction on 20 of 20 inputs per model and the single-anchor arm on 39 of 40, with identical costs and question-row errors; on the further GPUs, 60 rerun archived units gave identical costs and question-row errors, and all but one of them the same generated text; the exception is an input whose next token is nearly tied, which two processes on the same GPU decoded differently. A paired difference therefore carries this small decoding noise. The drop arms ran the archived gate code; the causal arms ran an extended version of it that gives bit-identical outputs for the archived arms in its unit tests. Differences below are in points on E4 unless marked as the five-task total (E5).

Strictly causal selectors. The progressive selector is compared with exact attention on E4 using the prespecified five-point non-inferiority margin and Holm correction across the two models (Table 33). Its differences are -5.1 [-11.4, $+1.0$] and -8.0 [-12.5, -3.7] points ($p_{\mathrm{ni}} = .51$ and .91): neither establishes five-point non-inferiority. Against four-anchor selection, the differences are -2.8 [-8.1, $+2.2$] and -3.0 [-7.5, $+1.3$]. These intervals do not prove that look-ahead is necessary or that all causal selectors fail. Progressive selection recovers 10.2 of 15.3 and 18.5 of 26.5 points of the single-anchor loss. The alternatives in Table 31 use the first row plus three rows from the preceding block, spend the three scoring rows on a larger single-anchor exact set, or shorten blocks to eight queries. None improves the five-task total over the progressive selector. The short-block arm was an exploratory follow-up, not part of the primary test family. Progressive selection builds and selects four states per block, whereas the four-anchor arm selects and builds one; this additional selection overhead is not captured by key-pair counts.

Remainder against drop at $m = N/4$. Table 32 lists the registered comparison of the zero-order remainder with dropping the unselected keys. Its primary contrast is the one at equal MACs and its secondary contrast the one at the same m; on both models neither excludes zero (Section 6), so the experiment does not establish a task-score benefit of the remainder beyond the selector. An interval containing zero is not evidence of equivalence or absence of an effect. The five-task totals give the same reading, $+1.3$ [-2.7, $+5.5$] and $+1.7$ [-0.5, $+4.0$] points at the same m and $+1.2$ [-2.5,

$+5.2$] and -0.5 [-2.5, $+1.4$] points at equal MACs. With the single-anchor selector the remainder changes E4 by $+1.1$ [-2.5, $+5.0$] and $+0.3$ [-3.1, $+3.9$] points at the same m and by -0.5 [-3.7, $+2.9$] and -2.1 [-6.3, $+2.3$] points at equal MACs, and the interaction of selector and remainder, (4 anchors $-$ 1 anchor) $\times$ (remainder $-$ drop) at the same m, is 0.0 [-6.0, $+6.0$] and $+1.8$ [-2.5, $+6.2$] points, so we do not claim that the remainder matters more under one selector than under the other. With the four-anchor selector, the remainder arm has the lower question-row error on every input of both models against the drop arm of the same m, but against the equal-MAC drop arm on Qwen3-8B-Base only on 99 of the 200 inputs; with the single-anchor selector it has the lower error on 200 and 161 of the 200 inputs at the same m and on 196 and 1 at equal MACs (Table 32). In a comparison we did not register, keeping $m = 71N/256$ instead of $N/4$ in the drop arm raises the five-task total of Qwen3-8B-Base by $+2.2$ [$+0.6$, $+4.2$] points.

Prefix invariance. We tested causality at three levels. In one attention layer (random queries, keys and values at 256 positions, float32) we perturbed the query, key and value at in-block position t of one block and compared every earlier row. For 15 causal configurations (the block-first row alone, the block-first row with three rows of the previous block, progressive selection, 8-query blocks and a fixed number of exact keys, each with the remainder and with drop; the remainder centered at $\bar{k}_{\mathcal{O}}$ with the first two; and a single-anchor tail-mass rule at three thresholds), with 48 perturbations each, every earlier row stays bit-identical. Under the four-anchor selector, with the remainder or with drop, the four-anchor tail-mass rule, the budget predictor and the centered four-anchor variant, rows 1 to $t-1$ change in every case, while perturbing a non-anchor position, 15, leaves every earlier row bit-identical even under the four-anchor selector: the dependence enters through rows 10, 21 and 31. Row 0 changes only within rounding under the zero-order remainder, by at most 7×10^{-7} of its norm, because at $q_i = q_a$ the remainder reproduces the history contribution whatever keys are selected; with drop it changes by up to 16% of its norm. In tiny random Llama and Qwen3 models (three layers, float32), changing the tokens after position p of a block leaves the logits and every layer's keys and values at positions up to p bit-identical for exact attention and all ten causal arms, while each of the four look-ahead arms changes, in every case, at least 15 of the compared per-position values (the keys and values of each layer and the logits) by more than 10^{-5}, with the keys and values of the first layer unchanged. In both full models (bf16) we replaced the prompt tokens at positions 22 to 31 of one block in two NIAH inputs. Under exact attention, the block-first row alone, the previous-block rows and progressive selection, the attention outputs, keys and values of positions 0 to 21 and of the previous block stay bit-identical in every layer. Under the four-anchor selector the attention outputs at positions 1 to 21 change in the first layer by up to 99% and 33% of their norm in the most affected head (Llama-3.2-3B, Qwen3-8B-Base), and their keys and values, unchanged in the first layer, change in every later layer, by up to 104% and 63%. Position 0 changes by at most one bf16 unit in the last place in the first layer; in the later layers the rounding differences grow, up to 13% for the attention output and 3% for the keys and values, which is of the order of the median change at positions 1 to 21, so it is unchanged only within rounding in the first layer and not bit for bit in the later ones. The generated answers were not part of this test.

J Longer Contexts and Failure Interpretation

J.1 Protocol and paired task results

We evaluate four-anchor zero-order prefill at $m = N/4$ and at a fixed $m = 4096$ on 32K-token inputs for both core models and 64K-token inputs for Llama. The generator, selector, decoding limits and strict scoring follow Appendix A. Exact attention is first evaluated on 40 inputs per synthetic task. NIAH and FWE form the paired primary endpoint; VT is kept as a structural check at 32K (10 inputs per approximate arm) because the exact baseline is near zero. Each approximate arm uses 40 inputs per primary task at 32K and 30 at 64K. The exact scores in Table 34 are on those same paired inputs, not the larger screening pool. No context-extension modification is applied to Qwen, and 64K is not tested for it.

The 32K Llama interval is too wide to establish five-point non-inferiority. At 64K, keeping a quarter of the history preserves NIAH accuracy on the paired sample but loses 40 points on FWE; its two-task average loses 20 points. Fixing the exact set at 4096 keys loses more. Thus the operator-level budget trends in Appendix G.6 are not an end-to-end preservation guarantee.

Table 29: **All eight arms of the reported run, on inputs of up to 16K tokens.** (a) Cost is the key-pair account of Appendix G as a fraction of dense attention; Fallb. is the share of blocks computed densely, counting every block, including the first block of each input and any final partial block, which every arm computes densely (0.4% of blocks); rel_q is the relative error on the question rows; Total is the equal-weight mean of the five task scores. Δ is the paired difference against exact attention on the same inputs, with a 95% interval from 2000 bootstrap resamples of the inputs within each task; the lower end of each interval is the one-sided 97.5% lower bound, the smallest non-inferiority margin that the data would pass at that level. Tail arms choose the budget per block (Appendix I); drop discards the keys outside the exact set. (b) Scores are accuracy for NIAH, FWE and VT and F1 for HotpotQA and 2WikiMQA. For the two arms of Table 1 the rows below the scores give the per-task paired differences and their 95% intervals, which are not adjusted for the five tasks.

(a) Cost, fallback, question-row error and total score

Arm	m	Cost	Fallb.	rel_q	Total	Δ Total [95% CI]
Llama-3.2-3B						
Exact attention	–	1.000	1.000	–	.476	–
1 anchor	$N/4$	.300	.004	.139	.353	$-.123$ [$-.180$, $-.068$]
4 anchors	$N/4$	.393	.004	.059	.458	$-.018$ [$-.056$, $+.020$]
4 anchors	$N/8$	.269	.004	.101	.414	$-.062$ [$-.123$, $-.005$]
4 anchors, drop	$N/8$	.253	.004	.169	.407	$-.069$ [$-.131$, $-.009$]
4 anchors	$N/32$	.175	.004	.197	.263	$-.213$ [$-.273$, $-.155$]
4 anchors, tail τ=.05	adaptive	.423	.261	.053	.429	$-.047$ [$-.097$, $+.002$]
4 anchors, tail τ=.1	adaptive	.282	.104	.092	.297	$-.179$ [$-.234$, $-.122$]
Qwen3-8B-Base						
Exact attention	–	1.000	1.000	–	.576	–
1 anchor	$N/4$	.300	.004	.149	.364	$-.212$ [$-.255$, $-.166$]
4 anchors	$N/4$	.393	.004	.045	.546	$-.030$ [$-.064$, $+.003$]
4 anchors	$N/8$	.268	.004	.083	.429	$-.147$ [$-.189$, $-.107$]
4 anchors, drop	$N/8$	.253	.004	.095	.442	$-.134$ [$-.174$, $-.092$]
4 anchors	$N/32$	.175	.004	.183	.300	$-.276$ [$-.323$, $-.228$]
4 anchors, tail τ=.05	adaptive	.431	.285	.046	.488	$-.088$ [$-.134$, $-.039$]
4 anchors, tail τ=.1	adaptive	.307	.138	.079	.372	$-.204$ [$-.255$, $-.155$]

(b) Per-task scores

Arm	m	NIAH	FWE	HotpotQA	2WikiMQA	VT
Llama-3.2-3B						
Exact attention	–	.975	.650	.416	.314	.025
1 anchor	$N/4$	.525	.625	.324	.268	.025
Δ vs exact		$-.450$	$-.025$	$-.092$	$-.046$	.000
95% CI		[$-.600$, $-.300$]	[$-.200$, $+.150$]	[$-.224$, $+.047$]	[$-.146$, $+.041$]	[$-.075$, $+.075$]
4 anchors	$N/4$	.950	.625	.395	.295	.025
Δ vs exact		$-.025$	$-.025$	$-.021$	$-.019$	.000
95% CI		[$-.100$, $+.050$]	[$-.150$, $+.100$]	[$-.130$, $+.091$]	[$-.064$, $+.015$]	[.000, .000]
4 anchors	$N/8$	.750	.525	.417	.330	.050
4 anchors, drop	$N/8$	.625	.650	.399	.311	.050
4 anchors	$N/32$	.200	.400	.376	.290	.050
4 anchors, tail τ=.05	adaptive	.825	.625	.367	.278	.050
4 anchors, tail τ=.1	adaptive	.300	.500	.327	.309	.050
Qwen3-8B-Base						
Exact attention	–	1.000	.975	.448	.458	.000
1 anchor	$N/4$	.200	.775	.445	.400	.000
Δ vs exact		$-.800$	$-.200$	$-.003$	$-.057$	.000
95% CI		[$-.925$, $-.675$]	[$-.325$, $-.075$]	[$-.124$, $+.124$]	[$-.124$, $-.007$]	[.000, .000]
4 anchors	$N/4$	.850	.975	.415	.439	.050
Δ vs exact		$-.150$	.000	$-.033$	$-.018$	$+.050$
95% CI		[$-.275$, $-.050$]	[.000, .000]	[$-.133$, $+.067$]	[$-.055$, $+.006$]	[.000, $+.125$]
4 anchors	$N/8$	.425	.950	.390	.380	.000
4 anchors, drop	$N/8$	.400	1.000	.406	.380	.025
4 anchors	$N/32$	.050	.800	.301	.349	.000
4 anchors, tail τ=.05	adaptive	.700	.925	.375	.391	.050
4 anchors, tail τ=.1	adaptive	.225	.925	.360	.353	.000

Table 30: **Same exact-set size, zero-order remainder minus drop.** Each row pairs the arm that keeps the zero-order remainder with the arm that drops the keys outside the same exact set, on the same 200 inputs. Δ is the paired difference in the total score with a 95% interval (2000 bootstrap resamples within task); Cost gives both arms (remainder / drop); Same is the share of inputs on which the two arms generate identical text. The earlier run used a single block-first anchor row and an earlier version of the gate code that is not part of the current reference implementation; its statistics are recomputed from the archived outputs, and it did not log question-row errors.

m	Model	Δ Total [95% CI]	Cost	Same	rel_q
Earlier run, single block-first anchor row (earlier gate code)					
$N/128$	Llama-3.2-3B	$+.036$ [$+.001$, $+.071$]	.058 / .043	.210	not logged
	Qwen3-8B-Base	$-.013$ [$-.049$, $+.017$]	.058 / .043	.430	not logged
$N/32$	Llama-3.2-3B	$+.118$ [$+.069$, $+.169$]	.082 / .066	.420	not logged
	Qwen3-8B-Base	$-.013$ [$-.045$, $+.019$]	.082 / .066	.580	not logged
Run of Table 1, four anchor rows					
$N/8$	Llama-3.2-3B	$+.007$ [$-.043$, $+.056$]	.269 / .253	.515	.101 / .169
	Qwen3-8B-Base	$-.013$ [$-.045$, $+.017$]	.268 / .253	.810	.083 / .095

J.2 Generation format and truncation

Qwen's 32K NIAH failures require a more specific interpretation than total loss of retrieval. At $m = N/4$, all 40 generations reach the 12-token maximum; 36 end with a numeric prefix of the correct answer after an explanatory opening. The exact arm completes the answer under the same budget. These are real failures under the fixed task interface, but they mix answer content, changed response style and truncation. We retain the original scores rather than retrospectively crediting partial answers. The mean teacher-forced log-probability of the complete gold answer also drops by 3.50 nats, so truncation does not establish that the underlying answer distribution was unchanged. A common longer generation budget would be needed to separate these explanations.

Llama's 64K FWE loss is not removed by that interpretation. The quarter-budget arm has 6 strict successes among 30 inputs versus 18 for exact attention; even a descriptive relaxed check asking whether all target words appear gives only 11 approximate successes. The long-context boundary is therefore task- and model-dependent rather than a single universal retrieval failure.

J.3 Question exposure and block alignment

An exploratory prompt-position study changed the placement of the question relative to the block. It also inserted instructions, removed some body tokens and changed which rows entered the question-error average. The exact FWE baseline itself changed from .650 to .175 in one placement. Such comparisons are compatible with sensitivity to question exposure, but they do not identify a pure causal effect of block alignment. We therefore use the fixed-sequence selector comparisons of Appendix I, not the prompt-position variants, as the evidence for multi-row coverage. A clean alignment intervention would keep the token sequence fixed and move only the block grid.

K Fresh-Input Confirmation

K.1 Frozen design and completeness

The confirmation evaluated fresh inputs on the two development models and on Mistral-7B-v0.3, which was not used during development. The main arm was fixed before confirmation as the four-anchor mean-remainder representation at $m = N/4$; the progressive selector was secondary. A single-anchor arm was additionally evaluated on Mistral. No drop-remainder arm was run, so this experiment does not isolate the value of the remainder.

The endpoint is the equal-weight average of per-task paired score differences on a model-specific frozen task set, not an input-weighted average. NIAH, FWE, VT, and TREC use binary correctness; HotpotQA and 2WikiMQA use answer F1. The task sets were fixed from development exact scores or exact-only pilots using a .30 floor. Llama and Qwen use FWE, 2WikiMQA, HotpotQA, NIAH, and TREC; Mistral uses HotpotQA, NIAH, and TREC. The legacy analysis variable is named E4, but the fresh endpoint contains five or three tasks and is distinct from the four-task development endpoint. Absolute averages across different models are therefore not on identical task sets.

Table 31: **Arms of the prespecified comparisons** (drop control and strictly causal selectors; the first three rows are the reference arms of Table 29), on the same 200 inputs per model. (a) Cost, rel_q and Total as in Table 29; Δ E5 is the paired difference of the five-task total against exact attention, Δ E4 that of the registered endpoint, the mean over NIAH, FWE, HotpotQA and 2WikiMQA, which leaves out VT because exact attention scores below .10 on it; both with 95% intervals from 2000 bootstrap resamples within each task, whose lower end is the one-sided 97.5% lower bound. Arms marked † look ahead within the block; the arms marked causal use at each position only rows already seen. (b) Per-task scores of the added arms; those of the reference arms are in Table 29.

(a) Cost, question-row error, total score and paired differences against exact attention

Arm	m	Cost	rel_q	Total	Δ E5 [95% CI]	Δ E4 [95% CI]
Llama-3.2-3B						
Exact attention	–	1.000	–	.476	–	–
1 anchor	$N/4$	.300	.139	.353	−.123 [−.180, −.068]	−.153 [−.224, −.084]
4 anchors†	$N/4$	.393	.059	.458	−.018 [−.056, +.020]	−.023 [−.070, +.025]
4 anchors†, drop	$N/4$	.377	.090	.445	−.032 [−.079, +.016]	−.033 [−.092, +.028]
4 anchors†, drop	$71N/256$	.405	.080	.446	−.030 [−.077, +.016]	−.038 [−.096, +.021]
1 anchor, drop	$N/4$	.284	.166	.340	−.136 [−.197, −.083]	−.164 [−.240, −.096]
1 anchor, drop	$71N/256$	.311	.151	.352	−.124 [−.183, −.069]	−.149 [−.223, −.080]
4 anchors, progressive (causal)	$N/4$	.393	.088	.440	−.036 [−.087, +.015]	−.051 [−.114, +.010]
4 rows, previous block (causal)	$N/4$	.393	.134	.347	−.129 [−.189, −.073]	−.155 [−.229, −.086]
1 anchor (causal)	$11N/32$	.393	.107	.437	−.039 [−.095, +.016]	−.061 [−.130, +.005]
8-query blocks (causal)	$13N/64$	.392	.087	.425	−.051 [−.107, +.003]	−.070 [−.137, −.006]
Qwen3-8B-Base						
Exact attention	–	1.000	–	.576	–	–
1 anchor	$N/4$	.300	.149	.364	−.212 [−.255, −.166]	−.265 [−.318, −.207]
4 anchors†	$N/4$	.393	.045	.546	−.030 [−.064, +.003]	−.050 [−.089, −.012]
4 anchors†, drop	$N/4$	.377	.051	.529	−.047 [−.086, −.011]	−.071 [−.116, −.031]
4 anchors†, drop	$71N/256$	.405	.045	.551	−.025 [−.060, +.006]	−.044 [−.084, −.009]
1 anchor, drop	$N/4$	.284	.151	.362	−.215 [−.257, −.169]	−.268 [−.321, −.211]
1 anchor, drop	$71N/256$	.311	.141	.381	−.195 [−.237, −.155]	−.244 [−.296, −.193]
4 anchors, progressive (causal)	$N/4$	.393	.076	.522	−.054 [−.093, −.017]	−.080 [−.125, −.037]
4 rows, previous block (causal)	$N/4$	.393	.136	.387	−.189 [−.231, −.145]	−.236 [−.288, −.181]
1 anchor (causal)	$11N/32$	.393	.119	.371	−.205 [−.246, −.167]	−.257 [−.307, −.208]
8-query blocks (causal)	$13N/64$	.392	.079	.432	−.144 [−.191, −.097]	−.193 [−.248, −.136]

(b) Per-task scores

Arm	m	NIAH	FWE	HotpotQA	2WikiMQA	VT
Llama-3.2-3B						
4 anchors†, drop	$N/4$	.850	.675	.378	.320	.000
4 anchors†, drop	$71N/256$	.850	.675	.382	.298	.025
1 anchor, drop	$N/4$	.550	.575	.319	.255	.000
1 anchor, drop	$71N/256$	.550	.625	.318	.268	.000
4 anchors, progressive (causal)	$N/4$	.850	.625	.384	.292	.050
4 rows, previous block (causal)	$N/4$	.475	.525	.442	.294	.000
1 anchor (causal)	$11N/32$	.575	.800	.432	.305	.075
8-query blocks (causal)	$13N/64$	.900	.550	.415	.210	.050
Qwen3-8B-Base						
4 anchors†, drop	$N/4$	.800	1.000	.374	.421	.050
4 anchors†, drop	$71N/256$	.875	1.000	.390	.438	.050
1 anchor, drop	$N/4$	.175	.800	.443	.390	.000
1 anchor, drop	$71N/256$	.175	.875	.444	.410	.000
4 anchors, progressive (causal)	$N/4$	.800	.975	.415	.369	.050
4 rows, previous block (causal)	$N/4$	.275	.800	.413	.447	.000
1 anchor (causal)	$11N/32$	.125	.875	.435	.420	.000
8-query blocks (causal)	$13N/64$	.575	.875	.262	.397	.050

Table 32: **Zero-order remainder minus drop at $m = N/4$** (registered runs). Each row pairs the arm that keeps the remainder with an arm that drops the keys outside the exact set, either at the same m or at $m = 71N/256$, which spends on exact keys the multiply-accumulates of the remainder (equal MACs). Δ E4 is the registered endpoint and Δ E5 the five-task total, both with 95% intervals (2000 bootstrap resamples within task); Same counts the inputs, of 200, on which both arms generate the same text; rel_q gives both arms (remainder / drop), and Lower counts the inputs on which the remainder arm has the lower question-row error. The interaction is (4 anchors − 1 anchor) × (remainder − drop).

Model	Comparison	Δ E4 [95% CI]	Δ E5 [95% CI]	Same	rel_q	Lower
Llama-3.2-3B	4 anchors†, same m	+.010 [−.039, +.059]	+.013 [−.027, +.055]	135	.059 / .090	200
	4 anchors†, equal MACs	+.015 [−.031, +.065]	+.012 [−.025, +.052]	133	.059 / .080	200
	1 anchor, same m	+.011 [−.025, +.050]	+.014 [−.017, +.045]	139	.139 / .166	200
	1 anchor, equal MACs	−.005 [−.037, +.029]	+.001 [−.028, +.030]	141	.139 / .151	196
	interaction, same m	.000 [−.060, +.060]	.000 [−.051, +.051]	–	–	–
	interaction, equal MACs	+.019 [−.040, +.078]	+.010 [−.037, +.058]	–	–	–
Qwen3-8B-Base	4 anchors†, same m	+.021 [−.006, +.050]	+.017 [−.005, +.040]	176	.045 / .051	200
	4 anchors†, equal MACs	−.006 [−.031, +.017]	−.005 [−.025, +.014]	173	.045 / .045	99
	1 anchor, same m	+.003 [−.031, +.039]	+.002 [−.025, +.031]	165	.149 / .151	161
	1 anchor, equal MACs	−.021 [−.063, +.023]	−.017 [−.050, +.019]	143	.149 / .141	1
	interaction, same m	+.018 [−.025, +.062]	+.015 [−.020, +.049]	–	–	–
	interaction, equal MACs	+.015 [−.036, +.065]	+.012 [−.028, +.052]	–	–	–

Table 33: **Strictly causal selectors** (registered run; all at the key-pair cost of the four-anchor arm, .393, except the 8-query blocks at .392). Against exact attention on the registered endpoint E4 with its 95% interval, whose lower end is the one-sided 97.5% lower bound compared with the margin of $-.050$, and the non-inferiority p value; against the look-ahead four-anchor arm on the five-task total (E5); and the part of the single-anchor loss that the arm recovers, arm minus one anchor over exact minus one anchor, in points (E5; E4). The primary comparison is the progressive arm against exact attention, with Holm's correction across the two models; the previous-block and $11N/32$ arms form a secondary family per model, also with Holm's correction; the 8-query-block arm ran only because the primary comparison failed and is descriptive.

Model	Arm	− exact, E4 [95% CI]	p_{ni}	− 4 anchors†, E5 [95% CI]	Recovered (E5; E4)
Llama-3.2-3B	progressive, 4 rows (primary)	−.051 [−.114, +.010]	.515	−.018 [−.063, +.026]	8.7/12.3; 10.2/15.3
	row 0 + 3 rows of previous block	−.155 [−.229, −.086]	.998	−.111 [−.165, −.055]	−0.6/12.3; −0.2/15.3
	1 anchor, $m = 11N/32$	−.061 [−.130, +.005]	.624	−.020 [−.073, +.033]	8.4/12.3; 9.3/15.3
	8-query blocks, $m = 13N/64$	−.070 [−.137, −.006]	.730	−.033 [−.082, +.013]	7.2/12.3; 8.3/15.3
Qwen3-8B-Base	progressive, 4 rows (primary)	−.080 [−.125, −.037]	.907	−.024 [−.062, +.012]	15.8/21.2; 18.5/26.5
	row 0 + 3 rows of previous block	−.236 [−.288, −.181]	1.000	−.159 [−.206, −.111]	2.3/21.2; 2.9/26.5
	1 anchor, $m = 11N/32$	−.257 [−.307, −.208]	1.000	−.175 [−.217, −.134]	0.7/21.2; 0.9/26.5
	8-query blocks, $m = 13N/64$	−.193 [−.248, −.136]	1.000	−.114 [−.161, −.068]	6.8/21.2; 7.2/26.5

The confirmation sample sizes were frozen at 817, 818, and 398 inputs for Llama, Qwen, and Mistral, respectively; 807, 808, and 368 enter their primary endpoints. Each model has ten VT structural-check inputs. Mistral additionally has ten FWE and ten 2WikiMQA structural-check inputs. These were not promoted into the primary endpoint after observing confirmation scores. In particular, Llama's fresh exact 2WikiMQA score falls below .30 but remains included, whereas Mistral's rises above .30 but remains excluded.

Synthetic inputs use a new seed and disjoint targets, but retain the development filler books; fresh inputs do not mean unseen filler corpora. LongBench questions exclude development question IDs. The new TREC task was selected through exact-only pilot calibration before confirmation, and pilot inputs were excluded. All arms of one input ran in the same process and on the same GPU. Jobs were distributed over four H100 GPUs and one H200. Approximation acts in 32-query prefill blocks; decoding is exact over each arm's own prefill cache. Gold-answer teacher forcing is a secondary diagnostic, not the generated-answer endpoint.

All 6,497 planned input–arm units were recovered: 2,451 Llama, 2,454 Qwen, and 1,592 Mistral. There were no missing, extra, duplicate, or failed units. Each unit passed the runner's structural validation and matched the frozen input hash, source hashes, two preregistration-document hashes, gold teacher-forcing setting, and shard assignment. The copied CPU analysis code matches the hashes in the frozen binding. The reproducible analysis and full per-input provenance are supplied with the supplementary artifacts.

Table 34: Long-context task evaluation. Scores are accuracy; Δ is the equal-weight NIAH/FWE difference from exact attention in percentage points, with 95% stratified paired-bootstrap intervals. The zero-width Qwen interval is the empirical bootstrap outcome when all paired task outcomes coincide; it is not population-level certainty.

Model, length	Budget	Paired n	NIAH	FWE	Δ [95% CI]
Llama, 32K	exact	80	.950	.475	–
	$N/4$	80	.850	.450	-6.2 $[-17.5, +5.0]$
	4096	80	.800	.500	-6.2 $[-17.5, +5.0]$
Qwen, 32K	exact	80	1.000	1.000	–
	$N/4$	80	.000	1.000	-50.0 $[-50.0, -50.0]$
	4096	80	.025	.975	-50.0 $[-53.8, -46.2]$
Llama, 64K	exact	60	.967	.600	–
	$N/4$	60	.967	.200	-20.0 $[-30.0, -8.3]$
	4096	60	.833	.133	-30.0 $[-43.3, -16.7]$

Model	Δ [95% CI]	p_{NI}	Holm threshold	Margin
Llama-3.2-3B	-7.1 $[-9.9, -4.4]$	0.9305	0.0083	Not met
Qwen3-8B-Base	-8.9 $[-11.2, -6.6]$	0.9990	0.0125	Not met
Mistral-7B-v0.3	-20.3 $[-24.7, -16.5]$	1.0000	0.0250	Not met

Table 35: Registered primary fresh-input comparison, in score points. Intervals and lower bounds are unadjusted; the decision uses the registered Holm procedure. The step-down procedure stops at its first failure, and all three models fail the five-point criterion.

K.2 Primary endpoint and non-inferiority

The registered margin is five score points. We use paired bootstrap resampling within each task, 2,000 draws and seed 20260726, then average tasks with equal weights. Intervals are percentile 95% intervals; their lower endpoint is also the one-sided 97.5% lower bound. The registered tail quantity is $p_{\mathrm{NI}} = (1 + \#\{b : \Delta_b \leq -0.05\})/2001$. Holm step-down correction is applied across the three main model comparisons at one-sided $\alpha = .025$. The secondary progressive comparison does not replace the frozen main arm.

The confirmation does not establish task preservation: the four-anchor losses are 7.1, 8.9, and 20.3 points on Llama, Qwen, and Mistral. The new model nevertheless retains a benefit from multirow selection: four anchors improve over one by $+13.3$ $[+8.8, +17.8]$ points on the same frozen three-task endpoint. This improvement is descriptive and does not imply equivalence to exact attention or establish a causal role for the remainder. It fixes m, not arithmetic: Mistral's pair-cost fraction increases from .3001 for one anchor to .3934 for four. The progressive selector is not uniformly better than four anchors, and neither is near-exact on these fresh inputs.

K.3 All tasks and registered structural checks

The following tables retain every evaluated task and arm. A dagger marks tasks excluded from the primary endpoint before confirmation. Scores are on a 0–100 scale; confidence intervals are paired differences in score points. Per-task intervals are descriptive and do not themselves establish taskwise non-inferiority. Each comparison uses the frozen stratified bootstrap; small percentile differences between task-only tables and primary-endpoint tables can arise from the registered task-order random-number draws.

K.4 Secondary diagnostics

Ceiling tasks were also frozen from the development scores or pilots: NIAH for Llama and Mistral, and FWE plus NIAH for Qwen. Their separate secondary margin is ten points, with Holm correction within each model. Passing one of these tests does not rescue the failed five-point primary endpoint.

Reproduction. Run `python3 scripts/analysis/analyze_fresh.py` from the root of the supplementary material with the per-input generation files in place (they are omitted from the supplementary material for size; their SHA-256 values are listed in its unit manifest). The wrapper validates the frozen unit set, copies units without moving or altering sources, calls the hash-matched

Model	Arm	Score	Δ versus exact [95% CI]
Llama-3.2-3B	Exact	61.4	—
Llama-3.2-3B	Four-anchor	54.3	−7.1 [−9.9, −4.4]
Llama-3.2-3B	Progressive	49.2	−12.3 [−15.4, −9.4]
Qwen3-8B-Base	Exact	75.4	—
Qwen3-8B-Base	Four-anchor	66.5	−8.9 [−11.2, −6.6]
Qwen3-8B-Base	Progressive	63.4	−12.0 [−14.3, −9.6]
Mistral-7B-v0.3	Exact	74.8	—
Mistral-7B-v0.3	Single-anchor	41.1	−33.6 [−38.1, −29.4]
Mistral-7B-v0.3	Four-anchor	54.4	−20.3 [−24.7, −16.5]
Mistral-7B-v0.3	Progressive	56.3	−18.5 [−22.7, −14.4]

Table 36: All evaluated arms on each model's frozen primary task set. Only the four-anchor comparison is in the primary three-model non-inferiority family; other arms are secondary or descriptive.

Task	n	Exact	Arm	Score	Δ [95% CI]
FWE	210	70.5	Four-anchor	55.7	−14.8 [−22.4, −7.1]
			Progressive	52.4	−18.1 [−25.7, −10.0]
2WikiMQA	134	29.2	Four-anchor	29.7	+0.6 [−4.5, +5.7]
			Progressive	30.9	+1.7 [−3.7, +7.2]
HotpotQA	155	34.6	Four-anchor	34.7	+0.1 [−3.7, +3.9]
			Progressive	33.7	−0.9 [−5.1, +3.1]
NIAH	210	99.5	Four-anchor	93.3	−6.2 [−9.5, −2.9]
			Progressive	73.8	−25.7 [−31.9, −20.0]
TREC	98	73.5	Four-anchor	58.2	−15.3 [−24.5, −7.1]
			Progressive	55.1	−18.4 [−27.6, −9.2]
VT†	10	0.0	Four-anchor	0.0	+0.0 [+0.0, +0.0]
			Progressive	20.0	+20.0 [+0.0, +50.0]

Table 37: Llama-3.2-3B: every evaluated task and approximate arm. † denotes a frozen structural-check task, not a primary-endpoint task.

registered CPU implementation, and regenerates these tables. It aborts on validation failure before exporting manuscript numbers. No new model execution or GPU computation is involved.

L Relation to Existing Attention Approximations

Table 44 compares the two selectors of this paper, the block-first row alone (1 anchor) and four rows per block (4 anchors, Section 6), both with the zero-order remainder, with the closest published methods along eight components: which queries choose the keys (C1); how their scores are combined over the query axis (C2); whether the keys that are not selected still contribute to the numerator, and through which statistic (C3); whether they keep a share of the normalizer in the same softmax (C4); around which point and to which order their contribution is approximated (C5); the stage at which the method acts and whether its selection is rebuilt or evicted (C6); whether it is strictly causal (C7); and the state it keeps besides the KV cache (C8). Every cell of another method is taken from the full text of the arXiv version named in the table, and its bracket gives the version and section (ResKV v1 leaves its main-text sections unnumbered, so for it we number them in order of appearance); 'not stated' marks what the paper leaves open. Every cell of our rows gives the lines of `code/anchor/anchor_gate.py` in the experimental code that implement it and, where one exists, the equation or section of this paper. The table compares designs, not accuracy, and gives a cost only for our rows.

Every column of our two rows has a precedent in the table. QuoKA chooses the keys of each prefill chunk from a few of its queries, takes the maximum of their scores and rebuilds the set for every chunk, so that, like our four-anchor selector, its selection looks ahead within the chunk (C1, C2, C6, C7). MomentKV, ResKV, MuSe and ClusterAttention keep the keys they do not compute exactly through summary statistics in the same normalization (C3, C4), and ResKV and MuSe keep, for each group they summarize, the same three parts as our block state: a mean key, a value summary and a mass (C8). MuSe also tilts its summaries toward a reference query, the centroid of a query cluster, and keeps the value at zero order (C5); it corrects the log-mass around each summary's own tilted

Task	n	Exact	Arm	Score	Δ [95% CI]
FWE	208	99.5	Four-anchor	99.0	-0.5 $[-2.4, +1.0]$
			Progressive	98.6	-1.0 $[-2.9, +0.5]$
2WikiMQA	137	42.4	Four-anchor	38.5	-3.9 $[-8.7, +0.7]$
			Progressive	34.2	-8.2 $[-13.0, -3.6]$
HotpotQA	155	59.0	Four-anchor	52.9	-6.1 $[-10.6, -1.6]$
			Progressive	52.0	-7.0 $[-11.4, -2.5]$
NIAH	208	100.0	Four-anchor	76.0	-24.0 $[-29.8, -17.8]$
			Progressive	65.4	-34.6 $[-41.3, -28.4]$
TREC	100	76.0	Four-anchor	66.0	-10.0 $[-18.0, -3.0]$
			Progressive	67.0	-9.0 $[-17.0, -1.0]$
VT†	10	0.0	Four-anchor	0.0	$+0.0$ $[+0.0, +0.0]$
			Progressive	0.0	$+0.0$ $[+0.0, +0.0]$

Table 38: Qwen3-8B-Base: every evaluated task and approximate arm. † denotes a frozen structural-check task, not a primary-endpoint task.

Task	n	Exact	Arm	Score	Δ [95% CI]
FWE†	10	0.0	Four-anchor	0.0	$+0.0$ $[+0.0, +0.0]$
			Progressive	0.0	$+0.0$ $[+0.0, +0.0]$
			Single-anchor	0.0	$+0.0$ $[+0.0, +0.0]$
2WikiMQA†	10	34.3	Four-anchor	30.0	-4.3 $[-10.6, +0.0]$
			Progressive	30.0	-4.3 $[-10.6, +0.0]$
			Single-anchor	20.0	-14.3 $[-34.4, +0.0]$
HotpotQA	135	51.7	Four-anchor	44.8	-7.0 $[-12.3, -2.0]$
			Progressive	42.8	-8.9 $[-14.9, -2.9]$
			Single-anchor	34.2	-17.5 $[-24.9, -10.5]$
NIAH	125	97.6	Four-anchor	52.8	-44.8 $[-53.6, -36.0]$
			Progressive	64.8	-32.8 $[-41.6, -24.0]$
			Single-anchor	27.2	-70.4 $[-78.4, -62.4]$
TREC	108	75.0	Four-anchor	65.7	-9.3 $[-15.7, -2.8]$
			Progressive	61.1	-13.9 $[-21.3, -7.4]$
			Single-anchor	62.0	-13.0 $[-20.4, -5.6]$
VT†	10	0.0	Four-anchor	0.0	$+0.0$ $[+0.0, +0.0]$
			Progressive	0.0	$+0.0$ $[+0.0, +0.0]$
			Single-anchor	0.0	$+0.0$ $[+0.0, +0.0]$

Table 39: Mistral-7B-v0.3: every evaluated task and approximate arm. † denotes a frozen structural-check task, not a primary-endpoint task.

key mean, where our rows use the full-history tilted mean $\bar{k}$ (Appendix B.6 discusses that choice of center).

No method in the table contains either of our rows as a special case in C1 to C4. The methods that choose keys from representative queries (QuoKA, SnapKV, FlexPrefill, MInference) drop the keys they do not choose. The methods that keep those keys choose by an eviction score, by clusters or per query, and all but MuSe summarize without a tilt: MomentKV and ResKV keep unweighted statistics of the evicted keys and ClusterAttention unweighted cluster centroids. A MuSe summary of one key cluster has the form of our remainder, with the query-cluster centroid in place of the anchor and a different center for the log-mass term, but MuSe chooses its exact keys for each query from key clusters rather than once per block from the block's own rows.

COBS (Tian et al., 2026) reports that expanding its block scores around a calibrated reference query instead of the origin lowers its long-context retrieval score, and attributes this to tilted moments that all query heads of a KV group must share and to queries far from the reference point (§7.7 of v1). Our state is kept for each query head (C8), so sharing across the heads of a group does not apply to the runs we report; it would apply to a variant that shares one state across the heads of a group, which we have not tested. Distance from the reference point does apply to us: a single block-first anchor misses keys that question rows later in the block need (Section 6).

Model	Task	Δ [95% CI]	p_{NI}	Margin
Llama-3.2-3B	NIAH	-6.2 $[-9.5, -3.3]$	0.0205	Pass
Qwen3-8B-Base	FWE	-0.5 $[-2.4, +1.0]$	0.0005	Pass
Qwen3-8B-Base	NIAH	-24.0 $[-29.8, -18.3]$	1.0000	Not met
Mistral-7B-v0.3	NIAH	-44.8 $[-53.6, -36.0]$	1.0000	Not met

Table 40: Secondary ceiling-task comparisons of four anchors versus exact, with a ten-point margin.

Model	Arm	rel_q	Pair cost	Identical
Llama-3.2-3B	Four-anchor	0.0598	0.3935	43.6%
Llama-3.2-3B	Progressive	0.0807	0.3935	33.2%
Qwen3-8B-Base	Four-anchor	0.0454	0.3935	69.7%
Qwen3-8B-Base	Progressive	0.0746	0.3935	65.4%
Mistral-7B-v0.3	Single-anchor	0.1145	0.3001	37.7%
Mistral-7B-v0.3	Four-anchor	0.0606	0.3934	65.1%
Mistral-7B-v0.3	Progressive	0.0802	0.3934	63.6%

Table 41: Descriptive diagnostics averaged over all evaluated inputs, including structural checks. rel_q compares approximate and exact attention on the approximate trajectory's own states. Pair cost includes the registered key-pair accounting, not wall-clock runtime; identical output means identical generated token IDs.

L.1 Position among recent advances

The rows of Table 45 come from keyword searches run on 25 and 26 September 2026: 144 queries on arXiv (computer science, submissions from 1 January 2025 on, plus title look-ups of older work) over seven themes, namely query-aware and native sparse attention, kernel and polynomial attention and their theory, low-rank structure of keys, sparse-plus-low-rank hybrids, attention sinks, KV-cache compression that reconstructs omitted contributions, and long-context evaluation; a local index of NeurIPS, ICML and ICLR proceedings including ICLR 2026; and web searches for venue status. A separate adversarial search looked for prior statements of each of our three contributions. Every row is based on the work's abstract, not its full text, so the second column states what a work reports and not a verified result. The last column names which of our measurements the abstract does not report: (i) the same local coordinates evaluated under an exact exponential and under finite-state realizations, i.e. the geometry–realization gap (Section 3); (ii) the restricted basis rebuilt from the top $N/128$ keys and the exact-set intervention that closes about 95% of the gap, with the projected-only, sink and support controls of Appendix C.2 (Section 4); (iii) matched-storage versus matched-arithmetic allocation, 78 of 112 configurations against none (Section 5); (iv) the question-row budget and the frozen fresh-input non-inferiority result (Section 6).

Several of these works share single ingredients with ours. Worst-case theory shows that low-dimensional scores need not admit a small nonnegative finite-feature realization or a small streaming state (Opazo, 2026; Haris & Onak, 2025; Gupta et al., 2026; Sarlos et al., 2023; Chu et al., 2024; Sui & Zhang, 2026); our measurement concerns the models' own keys at a fixed coordinate system, where it separates what better directions buy from what a finite-state realization recovers. Concurrent systems combine an exact set with a summary of the rest under one normalization (Li et al., 2026a; Huang et al., 2026; Hoshi et al., 2026; Yang & Li, 2026), and Support Basis computes large entries exactly and a polynomial on the rest (Aliakbarpour et al., 2026). None of their abstracts ties the exact set to the local geometry, measures how much of a realization gap the set carries, or compares allocations at matched arithmetic; conversely, we do not claim the operator forms they introduce.

Further related work by theme. Softmax attention (Vaswani et al., 2017) is computed exactly in tiles by FlashAttention (Dao et al., 2022). Query-aware methods estimate which keys a query needs from page or block statistics (Tang et al., 2024; Xu et al., 2025; Lin et al., 2025; Zhu et al., 2026a; Gao et al., 2025) or by sampling (Chen et al., 2025; Desai et al., 2025), trained models learn the selection natively (Yuan et al., 2025; Lu et al., 2025; DeepSeek-AI, 2025), chunked prefill can share one selection across the queries of a chunk (Song et al., 2026), and some heads can keep only sinks and recent tokens (Xiao et al., 2025); top-k attention has also been analysed through the total-variation distance it introduces (Tzachristas et al., 2026). Among kernel and polynomial realizations, RACE Attention replaces the exponential by a sharpened angular kernel (Joshi et al., 2026), log-linear attention replaces the fixed-size state by a logarithmically growing set of states (Guo

Model	Arm	Task	Δ log-probability [95% CI]
Llama-3.2-3B	Four-anchor	NIAH	-0.306 $[-0.368, -0.251]$
Llama-3.2-3B	Four-anchor	VT	$+0.227$ $[-0.010, +0.478]$
Llama-3.2-3B	Progressive	NIAH	-0.879 $[-0.960, -0.793]$
Llama-3.2-3B	Progressive	VT	$+0.534$ $[+0.222, +0.912]$
Qwen3-8B-Base	Four-anchor	NIAH	-0.232 $[-0.264, -0.197]$
Qwen3-8B-Base	Four-anchor	VT	$+0.064$ $[-0.065, +0.184]$
Qwen3-8B-Base	Progressive	NIAH	-0.285 $[-0.317, -0.253]$
Qwen3-8B-Base	Progressive	VT	$+0.098$ $[+0.021, +0.179]$
Mistral-7B-v0.3	Single-anchor	NIAH	-0.576 $[-0.635, -0.519]$
Mistral-7B-v0.3	Single-anchor	VT	-0.427 $[-0.607, -0.262]$
Mistral-7B-v0.3	Four-anchor	NIAH	-0.313 $[-0.377, -0.249]$
Mistral-7B-v0.3	Four-anchor	VT	-0.170 $[-0.276, -0.060]$
Mistral-7B-v0.3	Progressive	NIAH	-0.298 $[-0.368, -0.231]$
Mistral-7B-v0.3	Progressive	VT	-0.302 $[-0.447, -0.166]$

Table 42: Gold-answer likelihood differences (arm minus exact), per-token mean. Both tasks where teacher forcing was defined are retained. Values are log-probabilities, not score points. These are secondary diagnostics and were not substituted for generated-answer scores.

Model	Arm	Task	Δ log-probability [95% CI]
Llama-3.2-3B	Four-anchor	NIAH	-1.226 $[-1.472, -1.006]$
Llama-3.2-3B	Four-anchor	VT	$+0.682$ $[-0.030, +1.435]$
Llama-3.2-3B	Progressive	NIAH	-3.514 $[-3.841, -3.174]$
Llama-3.2-3B	Progressive	VT	$+1.601$ $[+0.666, +2.737]$
Qwen3-8B-Base	Four-anchor	NIAH	-1.854 $[-2.112, -1.573]$
Qwen3-8B-Base	Four-anchor	VT	$+0.384$ $[-0.390, +1.103]$
Qwen3-8B-Base	Progressive	NIAH	-2.281 $[-2.537, -2.024]$
Qwen3-8B-Base	Progressive	VT	$+0.588$ $[+0.124, +1.076]$
Mistral-7B-v0.3	Single-anchor	NIAH	-4.610 $[-5.082, -4.155]$
Mistral-7B-v0.3	Single-anchor	VT	-2.562 $[-3.642, -1.570]$
Mistral-7B-v0.3	Four-anchor	NIAH	-2.507 $[-3.019, -1.993]$
Mistral-7B-v0.3	Four-anchor	VT	-1.018 $[-1.655, -0.357]$
Mistral-7B-v0.3	Progressive	NIAH	-2.384 $[-2.946, -1.850]$
Mistral-7B-v0.3	Progressive	VT	-1.812 $[-2.684, -0.995]$

Table 43: Gold-answer likelihood differences (arm minus exact), answer sum. Both tasks where teacher forcing was defined are retained. Values are log-probabilities, not score points. These are secondary diagnostics and were not substituted for generated-answer scores.

et al., 2026), and control variates (Zheng et al., 2023), symmetry-aware Taylor expansions (Heinsen & Kozachkov, 2026) and polynomial methods for RoPE attention (Alman & Song, 2025) trade state size or time for accuracy; a curvature-conditioned read uses the running key covariance, which is the curvature of the softmax log-partition (Le et al., 2026). Low-rank structure of keys underlies Loki (Singhania et al., 2024), Eigen Attention (Saxena et al., 2024) and latent attention (DeepSeek-AI, 2024). Attention sinks were identified in work on streaming generation (Xiao et al., 2024). Among methods that keep what they do not compute exactly, RazorAttention adds a compensation token for the tokens it drops (Tang et al., 2025), eviction can be scored by its effect on the output rather than on the attention weights (Gupta & Berant, 2021; Feng et al., 2026), and clustered compression keeps cluster summaries (Zhang et al., 2025); in block-sparse video attention, retained attention mass alone does not determine the output error (Taghavi et al., 2026). Long-context evaluation has moved toward realistic multitask suites (Bai et al., 2025; Yen et al., 2025).

Table 44: **Component-by-component comparison, part 1 (C1–C4).** Our two selectors (zero-order remainder, $m = N/4$, as in Table 1) against the closest published methods. Brackets give the arXiv version and section each cell is taken from; for our rows, 'l.' gives lines of `code/anchor/anchor_gate.py` in the experimental code. Appendix L defines the columns.

Method	C1 Selector	C2 Aggregation over queries	C3 Keys not selected in the numerator	C4 Normalizer
This paper (zero-order remainder, $m = N/4$)				
Ours, 1 anchor (block-first row)	Block-first query q_a of each 32-query block; its exact softmax row p_a over the history $j < s_b$, per query head [l. 173–182, 447–449]	One row: the top m keys of p_a, $m = \max(1, \mathrm{round}(fs_b))$ [l. 101–105, 208, 235; Appendix A]	Yes: $S_1 = \sum_{j \notin \mathcal{H}} e^{s_a(j)} v_j$ and S_0, tilted by q_a; value S_1/S_0 [l. 209–216, 232–236, 294–296; Equation (2)]	Yes: one extra key with logit $\log S_0 + c_i$ in the same softmax as $\mathcal{H}$ and the in-block causal keys [l. 297, 346–357; Section 5]
Ours, 4 anchors	Queries at block positions 0, 10, 21, 31; exact softmax rows over the history $j < s_b$, per query head [l. 184–199, `fd_anchor_run.py` l. 81; Section 6]	Max over the four rows' own softmax probabilities per key, then the top m [l. 187–198, 208, 235; Section 6]	As 1 anchor; the tilt uses row 0 only [l. 71, 209, 294–296]	As 1 anchor [l. 297, 346–357]
Prefill key selection				
QuoKA (Jones et al., 2026) arXiv 2602.08722	Per prefill chunk (128 queries), the 16 queries least cosine-similar to the chunk's mean query; keys scored by cosine similarity to them [v1 §3.1–3.2, §4, Alg. 1]	Max over the kept queries, mean over the query heads of a KV group; top B_{SA} keys per KV head [v1 §3.3, Alg. 1]	No: keys not selected are not used for the chunk [v1 §3.4, App. A Alg. 2]	No: softmax over the selected keys and the chunk's own keys [v1 §2.4, App. A Alg. 2]
FlexPrefill (Lai et al., 2025) arXiv 2502.20766	The last 128 prompt queries: pattern (divergence test) and vertical-slash indices; query-aware heads: pooled blocks of all queries [v1 §3, §4.1]	Their softmax summed along vertical and slash lines, or pooled query blocks; smallest set reaching cumulative mass γ [v1 §3 Algs. 3–4, §4.1]	No: entries outside the index set are masked [v1 §2]	No: softmax over the selected entries only [v1 §2]
MInference (Jiang et al., 2024) arXiv 2407.02490	Offline search assigns each head a pattern; online, vertical-slash heads use the last 64 queries, block-sparse heads pooled 64-query blocks [v2 §3.2, Algs. 1–3]	Softmax of the last 64 queries summed along vertical and slash lines; mean pooling for block-sparse heads [v2 §3.2]	No: masked [v2 §3.1]	No: softmax over the sparse index set only [v2 §3.1]
Decoding-time cache compression or block selection				
SnapKV (Li et al., 2024) arXiv 2404.14469	The last L_{obs} prompt queries (32 in the LongBench runs) vote over the prefix [v2 §4, §5.3]	Sum of their softmax weights per head, 1-D max pooling (kernel 7), top k per head [v2 §4.1, §4.3, §5.3]	No: prefix entries not selected are evicted [v2 §4.1, Listing 1]	No: decoding attends to the compressed cache only [v2 §4.1]
MomentKV (Li et al., 2026b) arXiv 2606.01563	Kept set from SnapKV's prefill protocol; in decoding, the entry with the smallest attention-weighted moment residual $\alpha_j \lVert r_j \rVert$ is evicted [v1 §5.1, §4.2, Alg. 1]	Window sum at prefill; the current query in decoding [v1 §5.1, Alg. 1]	Yes: count, key sum, value sum and value-key outer-product sum of the evicted entries [v1 §4.1]	Yes: the evicted set enters as one pseudo-entry of mass $n_e e^{\beta q^\top \bar{k}}$ [v1 §4.2]
ResKV (Zhan et al., 2026) arXiv 2607.29591	Protected recent window plus the top tokens by the base method's score (SnapKV or AdaKV) on an observation window [v1 §5.2, App. B.3, Alg. 1]	Inherited from the base method; not stated further [v1 §5.2, Alg. 1]	Yes: r entries, each the mean key, mean value and count of a k-means cluster of the evicted keys [v1 §5.3, App. B.3]	Yes: the entries join the same softmax with logit $\beta q^\top \bar{k}_c + \log c_c + \log g(q)$ [v1 §5.4]
COBS (Tian et al., 2026) arXiv 2607.09052	No representative query: each decoding query scores every cached key block by $q^\top \bar{k}_b + \frac{1}{2} q^\top \Sigma_b q$ [v1 §5.2–5.5]	Block masses summed over the G query heads of a KV group; one block set per KV head [v1 §3.4, §5.2]	Not in the selection branch (top 16 blocks read exactly); a separate compression branch attends to mean-pooled blocks [v1 §2, §5.2, §7.1]	No single softmax: selection, compression and window branches are gated and summed [v1 §2]
Full-sequence clustered attention				
MuSe (Mitchell & Kersting, 2025) arXiv 2509.10406	Queries and keys k-means clustered per head over the whole sequence; per query, top key clusters, then top spatial blocks, computed exactly [v4 §3, §3.2]	None for selection (per query); summaries tilted by the query-cluster centroid [v4 §3.1–3.2]	Yes: every other (cluster, block) pair as a value centroid tilted by the query-cluster centroid, with a log-mass [v4 §3.1]	Yes: summaries and exact parts merged by logsumexp into one normalization [v4 §3.3]
ClusterAttention (Nordenram & Dittmann, 2026) arXiv 2608.26965 (bidirectional)	Query clusters of 64 from recursive PCA splitting; each ranks all key clusters (128 keys each) by centroid dot product [v1 §3.1–3.3]	Query-cluster centroid (mean query) [v1 §3.3]	Yes: each key cluster not selected, through its key centroid, value centroid and size [v1 §1.2, App. A]	Yes: in the same numerator and denominator [v1 App. A, §3.1]

Table 44: **Component-by-component comparison, part 2 (C5–C8 and cost).** Cost is the key-pair account of Appendix G at $m = N/4$, with dense attention $= 1$, averaged over the 200 end-to-end inputs of each model; the two models agree to three decimals (Table 1). The other methods count cost in other units, which we do not convert here.

Method	C5 Center and order of the rest	C6 Stage; rebuild or eviction	C7 Strictly causal	C8 Extra state	Cost
This paper (zero-order remainder, $m = N/4$)					
Ours, 1 anchor (block-first row)	Center q_a; value and mass at zero order, mass rescaled per query by e^{c_i}, $c_i = \beta(q_i - q_a)^\top \bar{k}$ with the full-history tilted mean $\bar{k}$ [l. 216–218; Appendix B.6]	Prefill only; rebuilt for every 32-query block; full KV cache kept, decoding exact; first and final partial block dense [l. 62, 142–155, 432–434; Appendix A]	Yes: each row uses q_a, the history $j < s_b$ and the in-block keys up to itself [l. 173–182, 346–347; Section 6]	Per block and query head: $\bar{k}$, S_1/S_0, $\log S_0$ ($d + d_v + 1 = 257$ numbers) and m key indices [l. 217, 294–297, 447–449; Section 5]	.300
Ours, 4 anchors	As 1 anchor; rows 10, 21 and 31 only select [l. 71, 176, 216–218]	As 1 anchor [l. 62, 142–155, 432–434]	No, in-block look-ahead: rows 1–30 use a set chosen with later queries, so the operator is not prefix-invariant [l. 184–199; Section 6]	As 1 anchor [l. 217, 294–297, 447–449]	.393
Prefill key selection					
QuoKA (Jones et al., 2026) arXiv 2602.08722	None (no remainder) [v1 §3]	Chunked prefill; key set rebuilt for every chunk; full KV cache kept; also used in decoding [v1 §3.4, §4.4]	Masks causal; a chunk's set depends on all its queries (in-chunk look-ahead) [v1 App. A Alg. 2]	None across chunks; per-chunk selection memory $O(n_{KV} N_Q T)$ [v1 App. C]	
FlexPrefill (Lai et al., 2025) arXiv 2502.20766	None (no remainder) [v1 §3]	Prefill only; pattern and budget chosen online per input and head; decoding dense [v1 §3, §4.1]	Masks causal; the pattern and the vertical-slash indices come from the last 128 prompt queries (look-ahead to the prompt end) [v1 §3, §4.1]	No persistent state [v1 §3, §4.1]	
MInference (Jiang et al., 2024) arXiv 2407.02490	None (no remainder) [v2 §3]	Prefill only; offline pattern search, online index build per input [v2 §3.2, App. C.2]	Masks causal; vertical-slash indices from the last 64 prompt queries (look-ahead) [v2 §3.2 Alg. 2]	A pattern configuration per head [v2 §3.2, App. C.2]	
Decoding-time cache compression or block selection					
SnapKV (Li et al., 2024) arXiv 2404.14469	None (no remainder) [v2 §4]	Prefill exact; one-shot compression of the prompt cache for all decoding steps [v2 §4]	Yes for decoding: the window precedes every decoding query [v2 §4]	$k + L_{obs}$ cached entries per head [v2 §4, Listing 1]	
MomentKV (Li et al., 2026b) arXiv 2606.01563	Center: unweighted mean evicted key $\bar{k}$; value first order, mass zeroth order [v1 §4.1–4.2, App. A–B]	Decoding-time eviction; statistics accumulate over evictions, never rebuilt [v1 §2, §4.1, Alg. 1]	Yes (decoding) [v1 §2]	$O(d^2)$ per head: a count, two d-vectors and a $d \times d$ matrix [v1 §4.1, App. C.1]	
ResKV (Zhan et al., 2026) arXiv 2607.29591	Center: each unweighted cluster centroid; zeroth order in the within-cluster logit deviation [v1 §5.3, App. A.2]	Decoding; built once after prefill per layer and KV head [v1 §5.5, App. B.3, Alg. 1]	Yes (decoding) [v1 §3.1]	r of the b cache slots hold residual entries [v1 §5.1, §5.5, App. B.3]	
COBS (Tian et al., 2026) arXiv 2607.09052	A selection score, not a remainder: block-key cumulant expansion at $q = 0$ to second order; a calibrated center q_0 scores lower [v1 §4, §5.2, §7.7]	Decoding; block descriptors computed once per filled block and cached; no eviction [v1 §2, §5.6]	Yes (NSA decoder) [v1 §2]	Per key block: mean key and $r = 4$ covariance factors in a query subspace [v1 §5.3–5.6]	
Full-sequence clustered attention					
MuSe (Mitchell & Kersting, 2025) arXiv 2509.10406	Center: query-cluster centroid (exponential tilt); value zeroth order, log-mass first order; a first-order value term is derived, not used in the main runs [v4 §3.2, App. B]	Pretraining and drop-in use; clusters and summaries recomputed in every forward pass [v4 Abstract, §3]	Causal at spatial-block granularity; clustering sees the whole sequence [v4 §3, §3.1]	Key, value and log-mass summaries per cluster pair and spatial block [v4 §3, §3.1]	
ClusterAttention (Nordenram & Dittmann, 2026) arXiv 2608.26965 (bidirectional)	Center: unweighted key-cluster centroids; zeroth order; exact error as a Jensen plus a covariance term [v1 §1.2, App. A]	One forward pass of bidirectional attention; clustering redone per head and pass [v1 Abstract, §3.1]	No: bidirectional attention only [v1 Abstract, §1.1]	No persistent state [v1 §3.2–3.3]	

Table 45: **Position among recent advances, part 1: geometry, realization and small supports.** One row per work or line of work, 2021–2026, most from 2025–2026; the second column follows the abstract. Last column: which of our measurements (i)–(iv) of Appendix L.1 the abstract does not report.

Work	What it establishes	Shared ingredient	What it does not measure, and this paper does
Low-dimensional structure and the cost of a finite-state realization			
Approximation rank (Sui & Zhang, 2026) arXiv 2608.28150v1	Geometric laws for the approximation rank of softmax attention in terms of logit magnitude and an interaction dimension	Interaction dimension as the controlling quantity	(i): a fixed rank-eight coordinate system evaluated under e^t and under finite-state realizations on the heads of trained models
Three tokens (Opazo, 2026) arXiv 2608.11427	A nonnegative kernel-attention head that solves all three-token instances needs $2^{\Omega(m)}$ features, where softmax needs m-dimensional scores	Low-dimensional scores versus the size of a positive finite-feature realization	(i): how much of the error that better directions remove a finite-state realization recovers on measured coordinates; the bound is a worst-case construction
Compression barriers (Haris & Onak, 2025) COLT 2025	Attention-based generation needs $\Omega(de^d)$ space already for head dimension $d = o(\log n)$	Low head dimension does not by itself give a small state	(i) on trained models' keys; (ii) where the missing weight lies
Any temperature (Gupta et al., 2026) ICLR 2026	Subquadratic algorithms for constant head dimension with large entries, and SETH hardness once the dimension grows slightly	The role of logit magnitude	(i): a measured gap between directions and realization at fixed rank
DARKFormer (Farzam et al., 2026) arXiv 2603.04127	Anisotropic pretrained queries and keys inflate the variance of isotropic positive random features; data-aligned random features reduce it	Geometry-aware finite realization	(i): directions and realization separated on the same coordinates against an exact-exponential reference; (ii) which keys carry the realization error
LOCKS (Hwang, 2026) arXiv 2607.24555	Keys are locally low-rank though globally high-rank; page-local rank-r bases rank pages for selection in decoding	Local rank-eight key bases	Anchor-tilted coordinates of a query block; (i) the realization of e^t on them; (ii) the basis rebuilt from a few high-mass keys
Small supports, restricted bases and exact-set corrections			
S^{4}R (Han et al., 2026) arXiv 2608.00528 (AACL-IJCNLP 2026)	Low-rank key and value subspaces built from a selectively sampled subset of prompt tokens, with attention over a sparse reconstruction	Bases built from a token subset	(ii): that the top $N/128$ keys by anchor mass capture about 98% of the optimal rank-eight variance of the full tilted covariance, and that the same keys localize the realization gap
Eigen analysis (Bhojanapalli et al., 2021) arXiv 2106.08823	Most variation among attention scores lies in a low-dimensional eigenspace shared across layers, so partial computation can estimate the rest	Reconstruction from partial computation	(ii) a per-block tilted key covariance rebuilt from high-mass keys; (i)
Support Basis (Aliakbarpour et al., 2026) AISTATS 2026	Exact computation of the sparse large entries and polynomial approximation of the dense rest give subquadratic attention without bounded entries	Exact large entries, a polynomial elsewhere	(ii) on trained heads: the share of a measured realization gap that the top-mass keys carry ($.954$ and $.944$), with projected-only and sink controls; selection by anchor mass rather than entry magnitude
Residual-mass accounting (Hoshi et al., 2026) arXiv 2604.05438	Exact contributions of sink and recent anchors and a retrieved set, a learned positive feature-map estimate for the rest, merged under one normalization	Exact support plus a residual in one normalization	(ii) the local geometry of the exact support; (iii) the arithmetic-matched comparison
CompKV (Huang et al., 2026) arXiv 2609.26300	The residual of block-mean compensation follows block mass and within-block logit variation, so selection targets the blocks that compensation handles worst	Exact set plus a mean-state compensation	(i)–(ii): geometry reconstruction and the realization gap; (iii) storage- versus arithmetic-matched allocation
CEDAR (Li et al., 2026a) arXiv 2609.07237	Exact and summarized chunks share one softmax normalization; high-error chunks are expanded, with an error bound set by within-chunk dispersion	Exact set plus summaries in one normalization	(i)–(ii); (iii) the comparison at matched multiply-accumulates, where a low-dimensional remainder wins in none of the 112 tested configurations

Table 45: **Position among recent advances, part 2: remainders under resource constraints, attention sinks and task-level boundaries.**

Work	What it establishes	Shared ingredient	What it does not measure, and this paper does
Exact sets and compact remainders			
SPLA (Wang et al., 2026) arXiv 2601.22379	Exact attention on blocks chosen by a second-order Taylor metric, a residual linear-attention state for the rest; continual pretraining	Exact set plus a separable remainder	(iii): the reversal between storage- and arithmetic-matched budgets on frozen models; (ii)
LoLA (McDermott et al., 2026) arXiv 2505.23666	Training-free: a sliding window, a sparse cache of hard-to-memorize pairs and a linear-attention state for all other pairs	Exact pairs repairing a separable state	(iii); (ii) an exact set chosen by anchor mass and tied to the local geometry
LESS (Dong et al., 2024) ICML 2024	A constant-size learned recurrent cache lets eviction methods query all tokens	A compact state for evicted tokens	(iii); a zero-order state tilted by the block's own query
Multipole Attention (Hooper et al., 2025) NeurIPS 2025	Exact attention for the most important keys and cluster centroids for the rest, in long reasoning	Exact set plus a summarized remainder	(iii); (ii)
OjaKV (Zhu et al., 2026b) Findings of ACL 2026	First and recent tokens kept at full rank, the others in a context-adapted low-rank basis updated online	Sink tokens exact, a low-rank rest	(ii) sink-only against anchor-mass exact sets (.316 and .418 of the gap against .954 and .944); (iii)
Trust the Mass (Shi & Gu, 2026) arXiv 2608.25230	Keeping the largest weights is near-optimal for eviction; a retrieval margin traces to rankings computed while the question is visible	Mass-based exact sets, storage accounting, question visibility	(iii) matched arithmetic; (iv) the operator budget of the question rows and a frozen non-inferiority test
When is attention compressible? (Yang & Li, 2026) arXiv 2608.21541	Large weights need not hold most of the mass, omitted values can cancel, and preserving the output may not preserve the task; one tail-summary slot per head	A tail summary per head; operator versus task fidelity	(i)–(ii); (iv) the question-row budget and the fresh-input non-inferiority result
Attention sinks			
Sink emergence (Gu et al., 2025; Barbero et al., 2025) ICLR 2025; COLM 2025	Sinks emerge with training and act like key biases that store attention without contributing values; attending to the first token avoids over-mixing	High mass on the first positions	(ii) with positions 0–3 exact in every arm: alone they close .316 and .418 of the gap, and the rest of the top set adds .641 and .533
KVSink (Su & Yuan, 2025) COLM 2025	Keeping the first tokens at full precision protects sinks under KV quantization; sinks can also appear later	Keeping positions 0–3 exact	(ii)
Task-level boundaries			
The Sparse Frontier (Nawrot et al., 2026) Findings of ACL 2026	Tolerable sparsity depends on phase, task and sequence length, so fixed-budget methods are suboptimal	Task-dependent budgets	(iv): an operator-level budget for the question rows (about $N/4$) and a frozen fresh-input non-inferiority test
Query visibility (Luo et al., 2026) arXiv 2607.11942	At matched budgets with a paired bootstrap, making compression query-agnostic reorders methods by how visible the question is to their scores	Question visibility; paired bootstrap	(iv): question-row operator error and frozen fresh inputs
Error certificates (Xie & Alanwar, 2026) arXiv 2607.21475	Poisson-sampled eviction yields per-step error certificates; a pre-registered LongBench study of question-aware eviction	Pre-registration	(iv): a non-inferiority margin for one representation across three models