WHETHER A KV CACHE CAN BE EVICTED BLIND IS A PROPERTY OF ITS QUERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Two summary statistics of an attention head's queries, their average direction and how tightly they cluster around it, predict whether the head's KV cache can be evicted blind, that is, cut to one set of kept tokens chosen before any query arrives, for queries that continue the document. For questions appended after the document, the model family decides: under a decision rule recorded before the analysis, the prediction holds on Olmo-3 and fails on Llama, Qwen2.5 and Qwen3.5. Blind eviction works on a head only when its queries agree on which cached tokens matter, and the two statistics rank heads by that agreement, with nothing fitted, at a Spearman correlation of at least $+0.89$ on held-out queries in open-weight models from three families. A serving system can therefore tell before decoding which heads it can evict blind on continuation-style workloads. A scoring rule built from the average direction is competitive with the strongest recent methods on held-out continuation queries: it adds less output error than each of them as released, and Expected Attention overtakes it once its query rotation is averaged over the held-out positions. On the LongBench benchmark, whose questions are appended, it is level with Expected Attention at a quarter of the cache kept and trails it at half. The measurements also explain why RoCo's division of accumulated attention by the number of positions that could see each token, which kvpress already applies, cuts that excess error for H2O by 45%. They show that a guarantee met on the queries that chose the kept set breaks on many unseen ones, and they yield five practical implications for eviction. Code: https://anonymous.4open.science/r/kv-incompressibility-A706/

1 INTRODUCTION

A language model reading a long document stores a key and a value per token, so that later tokens can attend to earlier ones. This cache often outgrows the model's weights and is the main obstacle to an affordable long context. The cache must shrink while the text is still being read, but the question comes later, so most methods score each token on the text read so far and discard the low scorers.

Every such method makes the same bet: that the tokens one query needs are the tokens the next query will need. If two queries want different tokens, no single choice of what to keep serves both, and the method is limited by how much the queries agree before it is limited by how clever its score is. The question this paper answers is how far the queries of each attention head agree about which cached tokens matter, and what that agreement allows a method that never sees them.

Terms. A method is query-free if it fixes the retained set before any query it serves arrives. A unit is one layer and one query head. Dropping a cached token changes a unit's output by an amount we compute exactly (Eq. 1), and a query's retention set is the B tokens it would miss most. The agreement of two queries is the Jaccard overlap of their retention sets. Regret is the relative output error of a retained set minus that of the query's own retention set at the same budget. A continuation query is a position inside the document, and an appended question is asked after it.

Agreement as the quantity to measure. An elementary observation says why agreement is the quantity to measure. Whatever score chose it, a fixed retained set overlaps the queries' own retention sets by at most $(1 + \bar{J})/2$ on average, where $\bar{J}$ is their mean pairwise agreement at the same budget

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.

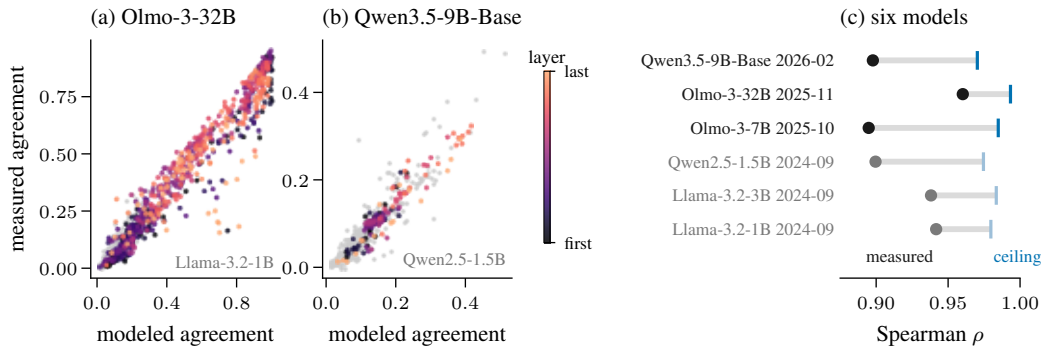

Figure 1: **(a, b)** Measured against modeled cross-query agreement per (layer, head) unit on the content-separated split, by relative depth, over a 2024 model in gray. Points rise with the modeled value, which overstates the level. The queries that estimated each unit's two moments share nothing with the queries that measured its agreement. **(c)** Content-separated Spearman ρ (black) and its reliability ceiling (blue) for the six models, newest first. The value of ρ attains 0.91 to 0.97 of its ceiling on every model.

(Lemma 1). We measure agreement for every unit whose cache grows with the context in six models released from 2024 to 2026. On the models we map head by head it is low everywhere, and it varies more than tenfold between heads of one model (Figure 2). Some heads want nearly the same tokens whatever they are asked, and others reorder their cache completely.

Prediction from two query moments. Two properties of a head's queries say which heads agree: the direction they point on average, and how tightly they cluster around it. Fed into a simulation against the head's own keys and values, they predict which heads agree on held-out queries on all six models, at Spearman $+0.895$ to $+0.960$ (Figure 1). It holds within layers as well as across them, near the ceiling our measurement noise allows. The model ranks heads correctly but overstates their agreement. For continuation-style workloads a serving system can thus rank heads for blind eviction at prefill. On held-out continuation queries, a query-free score that asks one query along the head's mean direction halves the regret of H2O as released, the best of the four 2023 to 2024 policies we ported. It also beats each of the five 2025 to 2026 methods we ported, as released, and only Expected Attention with its rotation averaged over the held-out positions overtakes it. On appended questions the model family decides whether prefill moments predict agreement: they pass on Olmo-3 and fail on Llama, Qwen2.5 and Qwen3.5 (Section 4.1).

The visibility correction to H2O. H2O as released by its authors sums attention over every prefill row. Under a causal mask it therefore counts an early token over more rows than a late one, a bias AhaKV traces to the mask (Gu et al., 2025). RoCo's mean attention score divides each token's accumulated score by its visibility count, the number of rows that could see it (Ren & Zhu, 2024), and the kvpress implementation of H2O already applies this division. What we add is why it works. The bound scores a retained set by an average over queries, and the division turns H2O's sum into an average over the rows that see each token. It cuts the regret of H2O as released by 45% on held-out continuation queries of Llama-3.2-1B.

Contributions. (1) The exact measurement of cross-query agreement per head on six models from three families, and two measured moments of a head's queries, nothing fitted, that predict it on held-out queries at Spearman $+0.895$ to $+0.960$ from one to thirty-two billion parameters, and on held-out documents on Llama-3.2-1B (Section 4). (2) A measured scope condition: whether those moments transfer from continuation queries to appended questions depends on the model family, passing on Olmo-3 and failing on Llama, Qwen2.5 and Qwen3.5, so any query model estimated at prefill has to be validated on the model that uses it (Section 4.1). (3) A query-free score from the mean direction that halves the regret of H2O as released and beats every ported 2025 to 2026 method as released on continuation queries, and that on LongBench, the task-level test, is level with Expected Attention within its interval (Section 5). (4) The division of H2O's score by the

visibility count is RoCo's, and kvpress already applies it. We contribute the account of why it helps, with the regret it removes and the share of the score's gain it carries (Section 5). (5) Five practical implications for building and evaluating eviction policies, each with the measured cost of ignoring it, including held-out violation of droppability guarantees on all six models (Section 6).

2 THE AGREEMENT BOUND AND THE QUERY MODEL

Setup. Fix a layer and a query head with head dimension d. The cache holds keys $k_1, \dots, k_N \in \mathbb{R}^d$ and values $v_1, \dots, v_N$. A query q attends with weights $w_t(q) = \exp(z_t)/\sum_j \exp(z_j)$, $z_t = q^\top k_t/\sqrt{d}$, and produces $o(q) = \sum_t w_t(q)\, v_t$. The exact cost of evicting token t under renormalization is

$$c_t(q) \;=\; \big\|o(q) - o_{-t}(q)\big\| \;=\; \frac{w_t(q)}{1 - w_t(q)}\,\big\|v_t - o(q)\big\|, \tag{1}$$

which we use throughout as the measured target, never as a proposed score. Each query q in a set Q induces the retention set A_q of its B costliest tokens, and cross-query agreement is the overlap $J(A_q, A_r) = |A_q \cap A_r|/|A_q \cup A_r|$ of two such sets.

The agreement bound. A query-free method commits to one set S of B tokens before any query in Q arrives. Agreement among the queries caps how well any single set can serve them.

Lemma 1 (elementary)**.** *Fix the budget B, and let $\bar{J}$ be the mean of $J(A_q, A_r)$ over ordered pairs of distinct queries in Q at that budget. For every set S, $\frac{1}{|Q|}\sum_{q\in Q} J(S, A_q) \;\le\; \frac{1}{2}\,(1 + \bar{J})$.*

Proof. Jaccard distance $1 - J$ is a metric on finite sets, so $1 - J(A_q, A_r) \le \big(1 - J(A_q, S)\big) + \big(1 - J(S, A_r)\big)$ for every pair. Averaging over ordered pairs of distinct queries, each query appears equally often on either side, which gives $1 - \bar{J} \le 2\big(1 - \frac{1}{|Q|}\sum_q J(S, A_q)\big)$. □

The bound holds per head and for every set, even one chosen with the queries in hand. In plain terms, when a head's queries disagree about which tokens matter, no fixed cache can serve them all well. It is a loose ceiling. Two disjoint retention sets attain it, but $|Q|$ pairwise disjoint ones allow only $1/|Q|$, since $J(S, A_q)$ is convex in the share of S drawn from A_q, so what a fixed set retains of an unseen query's cost is measured, not derived (Section 4). The lemma's $\bar{J}$ is a mean over pairs at one budget, and the reported statistic a median averaged over three budgets, which rank units alike (Appendix D).

The query model. Let $\hat{u} = q/\|q\|$ denote the query direction. We model the direction distribution of a head by its first two moments: a mean direction μ (the normalized mean of observed unit queries) and a mean pairwise cosine $\bar{c}$, realized as

$$\hat{u} \;=\; \sqrt{\bar{c}}\,\mu \;+\; \sqrt{1 - \bar{c}}\; e, \qquad e \sim \mathrm{Unif}\big(\mathbb{S}^{d-2}(\mu^\perp)\big). \tag{2}$$

Both parameters are measured from prefill queries, and nothing is fitted. Under (2) the logit of token t is, across queries,

$$z_t \;\sim\; m_t + \sigma_t\,\varepsilon_t, \qquad m_t = \tfrac{\|q\|}{\sqrt{d}}\,\sqrt{\bar{c}}\,(k_t^\top \mu), \qquad \sigma_t = \tfrac{\|q\|}{\sqrt{d}}\,\sqrt{\tfrac{1-\bar{c}}{d-1}}\,\big\|P_\perp k_t\big\|, \tag{3}$$

with ε_t zero mean, unit variance, and correlated across tokens through the shared draw of e. The first moment m_t coincides, up to normalization, with the point estimate used by expectation models of future attention (Devoto et al., 2025). The second, σ_t, is the query-induced dispersion that no first-moment score observes, and it sets how far the retention sets of individual queries disagree, which is what $\bar{J}$ in Lemma 1 summarizes.

Scope. The model rests on three assumptions, a fixed norm per head, an isotropic residual in $\mu^\perp$, and parameters estimated from prefill queries, and Appendix B gives the measurement that bounds each. The second is consistent with the model's one systematic failure (Section 4), and the third is measured rather than assumed (Section 4.1).

3 SETUP

Models and captures. We instrument six pretrained base decoders (Vaswani et al., 2017) from three families, in two sets of three that fill the same roles: a primary model, a second vendor, and a second scale of the primary's family. The 2024 set is Llama-3.2-1B and Llama-3.2-3B (Grattafiori et al., 2024) (16/28 layers, 32/24 query heads sharing 8 KV heads, $d = 64/128$) with Qwen2.5-1.5B (Qwen et al., 2024) (28 layers, 12 query heads sharing 2 KV heads, $d = 128$), all released in 2024-09. The 2025 to 2026 set is Olmo-3-32B (2025-11) and Olmo-3-7B (2025-10) (Team Olmo, 2025) (64/32 layers, 40/32 query heads sharing 8/32 KV heads, $d = 128$), with Qwen3.5-9B-Base (2026-02) (Qwen Team, 2026) (32 layers, 16 query heads sharing 4 KV heads, $d = 256$). All use rotary position embeddings (Su et al., 2024). The 2025 to 2026 models are hybrids. Qwen3.5-9B-Base has three linear-attention layers, which hold a fixed-size recurrent state and no KV cache, for each full-attention layer. Olmo-3 has three sliding-window layers, whose eviction by position the architecture fixes, for each full-attention layer. The captures cover every layer whose cache grows with the context, which is every layer of the 2024 models and the 8/16/8 full-attention layers of Olmo-3-7B / Olmo-3-32B / Qwen3.5-9B-Base, and a unit index counts those layers. The same document is prefilled to $N = 8192$ tokens on every model and we record, at every captured layer, the post-RoPE keys and values per KV head and the post-RoPE query of each probe position per query head. Every experiment reads these frozen captures, except that on the hybrid stacks the question regime of Section 4.1 forwards its questions against a fresh prefill cache of the document. Slashed values are per model, in the order Llama-3.2-1B / Qwen2.5-1.5B for pairs and Llama-3.2-1B / Llama-3.2-3B / Qwen2.5-1.5B for triples, and Table 1 carries every model. The question regime adds Qwen3.5-2B-Base (2026-02) (Qwen Team, 2026), a hybrid of the Qwen3.5-9B-Base family.

Query set. One rule, set before any analysis and shared by every experiment and model, chooses the probe positions used as queries. A grid position p enters the query set iff its visible cache satisfies $p + 1 \geq 3200$, so that a standard sink-plus-recency band of $64 + 256 = 320$ tokens covers at most 10% of what the query can see. The rule yields 33 queries per model. Token positions scored for eviction are the mid cache: everything visible except the first 64 (sink) and last 256 (recency) positions, which index-based heuristics already keep (attention share of this region in Appendix E).

Measured and simulated agreement. For a unit (layer ℓ, query head h), each query's mid-cache costs are the exact leave-one-out values of Eq. (1) against the unit's own keys and values. Each query induces a top B retention set, and the per-unit statistic is the mean over $B \in \{32, 64, 128\}$ of the median Jaccard overlap over 528 fixed query pairs (self pairs and same-position pairs excluded, robustness in Appendix D). The model side replaces real queries by draws from Eq. (2) with μ, $\bar{c}$, and the median norm measured per unit, zero fitted parameters, as a Monte Carlo over 330 query events allocated round robin over the real positions, because the shared residual draw makes per-token marginals insufficient (Appendix D). Cost, B grid and pair rule share code on both sides.

Units and inference. The primary analysis is at unit granularity. Units within a layer share weights and inputs and are not independent, and treating them as independent would be the language-as-fixed-effect fallacy (Clark, 1973; Judd et al., 2012), so no test statistic uses unit-level degrees of freedom. All confidence intervals are cluster bootstrap by layer (2000 resamples, percentile intervals). Correlations are read against a reliability ceiling $\sqrt{\rho_{xx}\rho_{yy}}$ measured on the actual estimator (Cronbach et al., 1972), and scores against information-free controls (Hewitt & Liang, 2019). Cross-unit correlations are reported raw and layer demeaned, and comparisons between dependent correlations as the difference Δr with its own interval.

4 CROSS-QUERY AGREEMENT AND THE QUERY GEOMETRY

Held-out queries on six models. The model orders measured agreement across units at ρ from $+0.895$ to $+0.960$ over the six models when parameters and evaluation share no query (Figure 1, Table 1). The 2025 to 2026 set overlaps the range of the 2024 set, and Olmo-3-32B, the largest model, has the highest value. The content-separated split, the design of record, alternates eight contiguous blocks of a dense 72-position grid between the halves that estimate and evaluate, so each query sits hundreds to thousands of tokens from the nearest query of the other half, which removes

			content separated		reliability		held-out violation, $E = 10^{-2}$		
Model	released	units	ρ	demeaned	ceiling	attained	rate (%)	k^*	droppable
Llama-3.2-1B	2024-09	512	$+0.942$	$+0.942$	0.980	0.96	18.3	2.5	0.27
Llama-3.2-3B	2024-09	672	$+0.938$	$+0.949$	0.983	0.95	15.2	2	0.28
Qwen2.5-1.5B	2024-09	336	$+0.900$	$+0.908$	0.975	0.92	18.5	2	0.26
Olmo-3-7B	2025-10	256	$+0.895$	$+0.910$	0.985	0.91	11.8	2	0.14
Olmo-3-32B	2025-11	640	$+0.960$	$+0.949$	0.993	0.97	8.9	1.25	0.09
Qwen3.5-9B-Base	2026-02	128	$+0.898$	$+0.829$	0.970	0.93	26.1	4	0.11

Table 1: Main result. Units: layers whose cache grows with the context, times query heads. Demeaned: each layer's mean removed first. Ceiling: $\sqrt{\rho_{xx}\rho_{yy}}$ measured on these estimators. Attained: ρ as a fraction of it. Held-out violation (Section 5): the share of unseen (head, query) blocks exceeding the budget at face value, the inflation k^* restoring a rate near 5%, and the droppable mid-cache fraction at k^*.

shared position. The cross-document design removes shared content instead, estimating the moments on document i and predicting the agreement measured on document j for all 72 ordered pairs on Llama-3.2-1B: median $\rho = +0.932$ [$+0.867$, $+0.955$]. Against the same-document diagonal the paired cost is $+0.019$ [$+0.013$, $+0.026$] within a declared margin of 0.05. A head's query geometry is therefore fixed by the head together with the distribution its queries are drawn from, rather than by the text. In practice, two statistics measured once per head tell a serving system which heads it can evict blind for continuation queries.

Correlation against the reliability ceiling. The headline reaches 0.91 to 0.97 of ceilings from 0.970 to 0.993 (Table 1, Appendix D). With all queries shared, the in-sample construction, run on four of the six models, matches the content-separated values, so shared sampling contributes little. The fractions are floors, as model-side reliability excludes parameter noise.

Retained cost of an unseen query. Lemma 1 bounds set overlap, and an eviction method pays in cost. Per unit we build the query-free ordering from the retained queries as the mean leave-one-out cost, select the top B, and measure on the held-out query the fraction of its mid-cache eviction cost that the selected set captures. Predicted agreement orders that fraction at $+0.868$ (Llama-3.2-1B) and $+0.875$ (Qwen2.5-1.5B) at a quarter of the mid cache, and $+0.822$ to $+0.904$ across both models and every budget tested, and a marginals-preserving shuffle collapses it to $+0.219$ or below (Appendix I).

Controls on the prediction. Replacing the mean direction by a random one of matched spread gives $+0.821$ / $+0.782$ and an isotropic query $+0.107$ / $+0.371$. Nonlinear regressions on head-level summaries still trail the full model, and removing profile concentration leaves $+0.944$ / $+0.900$, so the prediction reduces to no scalar we tested (Appendix I). With parameters remeasured per document on Llama-3.2-1B the unit-level correlation stays within [+0.943, +0.962] over 9 documents from five registers (Appendix G).

Overprediction of the agreement level. The model predicts the ordering of units far better than the level of agreement: in sample, simulated top B overlap exceeds measured overlap by roughly 1.6 to $2.3\times$. The cause is the second of the model's assumptions: the residual directions are anisotropic in $\mu^{\perp}$. On the content-separated split of Llama-3.2-1B, where median simulated over median measured agreement is 1.34, drawing the residual from its measured covariance moves that ratio to 0.99, where an isotropic null of the same sample size leaves it at 1.36 (Appendix J).

Dependence on query spacing. Measured agreement depends on the spacing of the queries compared (Figure 2): median pairwise agreement falls from ≈ 0.11 at close spacing to a stable ≈ 0.05 beyond a thousand tokens, on both models. Close queries share nearly their whole cache, so comparisons at that spacing measure near-duplicate stability.

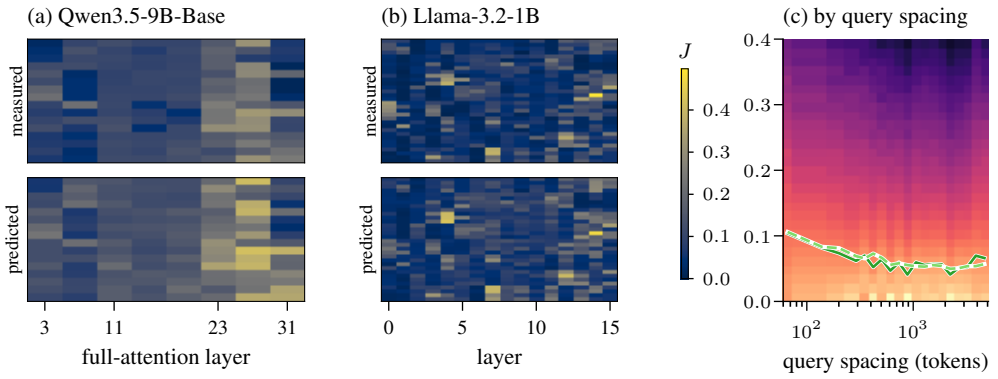

Figure 2: **(a, b)** Held-out cross-query agreement as a layer × head map, measured (top) over predicted (bottom) on a shared scale. **(c)** Agreement against query spacing on Llama-3.2-1B, as the conditional density of pairs, with the median for Llama-3.2-1B (solid) and Qwen2.5-1.5B (dashed).

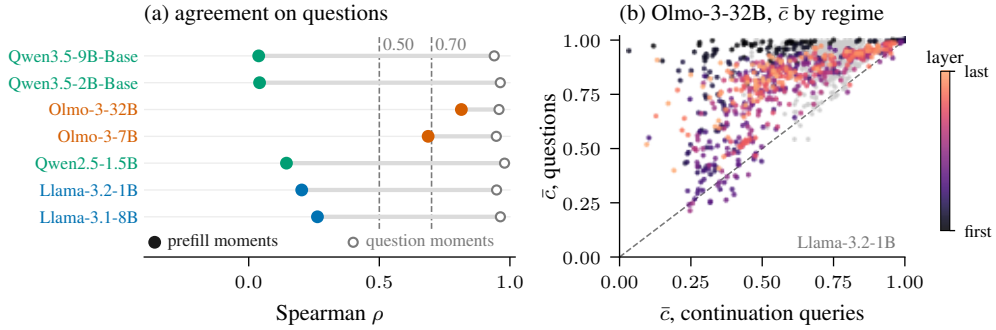

Figure 3: **(a)** Spearman ρ of agreement on appended questions with its prediction from prefill moments (filled) and question moments (open), registered bars dashed. **(b)** Per-head query concentration $\bar{c}$, continuation queries against questions, over Llama-3.2-1B in gray.

4.1 Continuation queries against appended questions

Every query used so far is a continuation query, whose query vector predicts the document's next token. On Llama-3.2-1B we appended 48 questions to the same document, six per eighth of it, each answerable from that eighth, with the thresholds and the reading of either outcome recorded in the repository before the capture (Appendix R).

Within the question regime the model holds: with μ and $\bar{c}$ measured on the questions of four eighths and agreement measured on the questions of the other four, sharing no question and no source text, it predicts which heads agree at $+0.949$ against a bar of 0.70. The primary test asks the deployment question: whether the moments a serving system can measure at prefill predict agreement among the questions it will later be asked. The answer is no ($+0.202$ against a bar of 0.50), and the measured side is reliable enough to rule out a noisy target ($+0.927$ split half). Which heads agree is not the same ordering in the two regimes (Appendix L). For a serving system, statistics gathered while reading a document cannot be assumed to describe the questions later asked about it.

On Llama-3.2-1B the reason is that the query geometry itself differs between the two regimes (Figure 3). Median $\bar{c}$ is 0.720 under continuation queries and 0.955 under questions, the mean directions differ, and per-head $\bar{c}$ rank correlates across the two regimes at $+0.178$. It is not an artifact of the questions sitting further from the scored region: distance-matched late continuation queries give median agreement 0.055, against 0.497 for the questions.

Further models. The same statistics and code, on the same document and 48 questions, were run on six further models and on a fresh capture of Llama-3.2-1B, which reproduces the original value

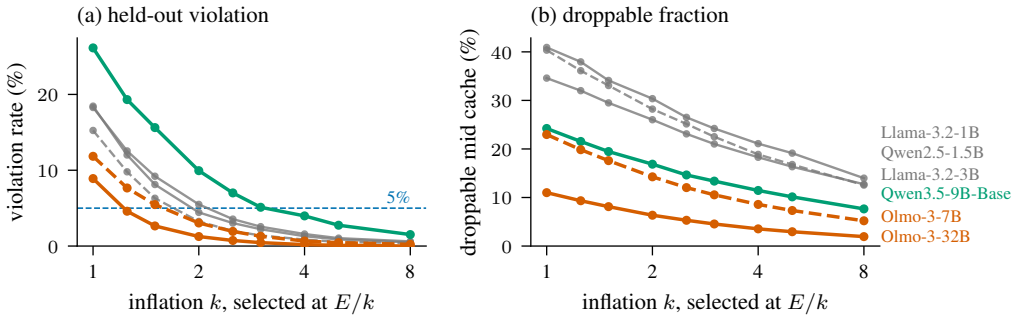

Figure 4: **(a)** Held-out violation rate and **(b)** droppable mid-cache fraction at $E = 10^{-2}$ against the inflation factor k (selected at E/k, measured against E) on all six models, 2024 models in gray, the smaller model of a family dashed, the 5% level in blue.

(+0.203). All entered the design before their captures, the larger checkpoints through an amendment (Appendix R). The primary fails on 5 of 7 configurations and passes on Olmo-3-7B and Olmo-3-32B, reaching up to +0.814 against the bars of 0.50 (partial) and 0.70 (established) (Table 10 in Appendix L). Every configuration that passes comes from one family, and the Llama, Qwen2.5 and Qwen3.5 models fail at every size tested. Within the models tested the transfer therefore follows the family, and scale does not predict it. Qwen3.5-9B-Base is a hybrid captured on the same path as the Olmo-3 models, so neither stack nor capture path accounts for the split, and the dense Llama-3.1-8B (2024-07) gives +0.263. We read the split through the query geometry: per-head $\bar{c}$ rank-correlates across the regimes at +0.178 to +0.719, more closely on every model that passes than on any that fails. Which heads agree keeps its ordering across the regimes on Olmo-3-32B alone.

5 Consequences for deployed eviction

Held-out violation of droppability guarantees. Per (layer, KV head) we grow a dropped set until the worst query row loses attention mass E (Figure 4, null in Appendix H). A set chosen on 32 queries at budget E cannot exceed it on those queries, but on all six models 8.9% to 26.1% of unseen (head, query) blocks exceed it at $E = 10^{-2}$ (Table 1). Selecting at E/k^* holds unseen-query violations near 5% while 0.09 to 0.28 of the mid cache stays droppable. In practice, a guarantee checked only on the queries that set it overstates safety, and a uniform margin restores it.

Absolute output error and mid-cache mass. Agreement orders absolute output error with a positive sign on each 2024 model, against a predicted negative one, because mid-cache attention mass sets how much a head has to lose (Appendix H). Regret normalizes that stake out, and conditional on mass share predicted agreement orders regret at $\rho = -0.362$ [-0.449, -0.264] on Llama-3.2-3B.

Published policies at matched budget. We compare four policies of 2023 to 2024, H2O, SnapKV, PyramidKV and StreamingLLM, each ported from released code and checked against the authors' own classes on identical tensors before scoring (all reproducing the retained set exactly, Appendix N). Each selects from prefill alone and is scored by regret on held-out continuation queries at the budget of Table 2. Every method, ours included, keeps the sink and recency band of Section 3 and differs only on the mid cache. H2O is the one method that scores every prefill row and the one whose retained set resembles the mean leave-one-out ordering of Section 4. Predicted agreement orders regret for H2O at -0.527 and for none of the other three, against a bar of three in four set for the comparison (Appendix N). The five released 2025 to 2026 policies of Table 2, ported from kvpress (NVIDIA, 2024) and checked against it, all have higher regret than the score of this section, which differs from the best of them, Expected Attention, by -0.00035 [-0.00052, -0.00023] paired by unit. With its rotation averaged over the held-out positions, which its release does not do, it falls below the score (sensitivity row). With one retained set per KV head, shared by its query heads as grouped-query attention needs for a memory saving, the score still has lower regret than Expected Attention, which exceeds it by +0.00031 [+0.00024, +0.00040] paired by unit (fourth block).

retained set chosen from prefill alone	median regret	90th percentile
PyramidKV (Cai et al., 2024)	0.0275	0.0891
SnapKV (Li et al., 2024)	0.0259	0.0885
StreamingLLM (Xiao et al., 2024)	0.0218	0.0810
H2O as released (Zhang et al., 2023)	0.0155	0.0503
H2O divided by visibility count (Ren & Zhu, 2024)	0.0086	0.0272
mean-direction query, attention weight	0.0067	0.0239
mean-direction query, eviction cost	0.0066	0.0230
random-direction query, eviction cost (control)	0.0221	0.0733
CurDKV (Sengupta et al., 2025)	0.0224	0.0748
KeyDiff (Park et al., 2025)	0.0204	0.0680
CapKV (Yang et al., 2026)	0.0140	0.0460
KVzip (Kim et al., 2025)	0.0129	0.0409
Expected Attention (Devoto et al., 2025)	0.0073	0.0254
rotation over held-out positions (sensitivity)	0.0057	0.0203
mean-direction query, eviction cost, same run	0.0065	0.0235
Expected Attention, one set per KV head	0.0085	0.0281
mean-direction query, one set per KV head	0.0077	0.0258
exact ordering from held-out real queries	0.0061	0.0215

Table 2: Regret, in relative output error against the per-query oracle, on held-out continuation queries at one matched budget, 1040 of 2880 mid-cache tokens, on 512 units of Llama-3.2-1B, one document and 33 evaluation queries. LongBench is the task-level test. The last row uses real queries and is a ceiling. The fourth block shares one retained set among the query heads of each KV group, the form that saves memory under grouped-query attention. The third block, per query head, comes from one run on a second extraction of the capture, in which the score and H2O reproduce the first two blocks to within 0.7%. Sensitivity: rotation averaged over the held-out positions.

The visibility correction of H2O. The division by the visibility count is RoCo's (Ren & Zhu, 2024), a correction of the mask bias that Wang et al. (2025b) make by a vote, and the kvpress implementation of H2O already applies it. Our contribution is why it works: the division turns H2O's sum into the average over queries by which the bound scores a retained set. It cuts the regret of H2O as released by 45% on its own (Table 2).

A query-free score from the mean direction. A score that sums attention over prefill rows sees the first moment of Eq. (3). Our score asks one query along μ at the head's median query norm and ranks by the exact eviction cost of Eq. (1), top B. It needs the head's own keys and values, which prefill produces anyway, and two scalars per head measured on a different document. It halves the regret of H2O as released, the best of the four 2023 to 2024 ports, from 0.0155 to 0.0066, and closes 95% of the distance to the ordering that sees the real queries. Each of its three steps, the visibility normalization, the single mean-direction query and reading the eviction cost, closes part of the gap from H2O with an interval excluding zero (Appendix O). A random direction of matched norm, all else unchanged, raises regret 3.4×. The second moment governs the bound instead, and predicts on Llama-3.2-1B which heads break the guarantee above (-0.450 [-0.538, -0.359]), with an interval covering zero on Qwen2.5-1.5B, though a per-head inflation factor set from it frees no more cache than a uniform one (Appendix Q).

Task accuracy. On LongBench with the instruct Llama-3.1-8B at a quarter of the context cache kept, in a comparison recorded before compute, the division leaves the LongBench average of H2O as released unchanged within its interval (-0.6 [-3.4, $+2.1$] points), and the mean-direction score is level with Expected Attention within its interval (-0.2 [-3.0, $+2.6$], Appendix P). With half the cache kept it trails Expected Attention by 2.9 points (-2.9 [-5.1, -0.9]). LongBench asks appended questions, where Section 4.1 finds that prefill moments do not predict agreement on the Llama family, so neither regret gain reaches task accuracy there.

6 PRACTICAL IMPLICATIONS FOR EVICTION SYSTEMS

Each item carries its measured cost, with full versions in Appendix M.

Five implications for query-free eviction

1. **Divide accumulated attention by the visibility count**, as RoCo does (Ren & Zhu, 2024) and kvpress already applies. It cuts the regret of H2O as released by 45% (Table 2).
2. **State guarantees on held-out queries.** On unseen queries 8.9% to 26.1% of blocks break a budget no in-sample block can break, and 1.25 to 4× inflation restores it.
3. **Report the query spread.** Close queries agree at 0.11, distant ones at 0.05 (Figure 2).
4. **Analyze per head rather than per layer.** Most of the variance in query geometry lies within layers, and with each layer's mean removed the prediction keeps $\rho \geq +0.829$ (Table 1).
5. **Measure query statistics on the queries you will serve.** Question moments predict agreement among questions at +0.940 or above, and prefill moments fail on 5 of 7 configurations (Section 4.1).

7 RELATED WORK

Query-free KV cache selection. Besides paging the cache (Kwon et al., 2023), one large family of remedies selects tokens with query-free scores: fixed sink plus recency (Xiao et al., 2024), accumulated attention mass (Zhang et al., 2023; Liu et al., 2023), per-head adaptive policies (Ge et al., 2024), prefill attention patterns (Li et al., 2024), key norms (Devoto et al., 2024), key diversity (Park et al., 2025), value-guided CUR selection (Sengupta et al., 2025), retained information capacity (Yang et al., 2026), a recurrent state view (Oren et al., 2024), and budget allocation by layer or head (Cai et al., 2024; Feng et al., 2025; Xiao et al., 2025; Mao et al., 2026), with query-aware decode-time selection (Tang et al., 2024) and quantization (Liu et al., 2024; Hooper et al., 2024) as complementary regimes. RoCo divides accumulated attention by each token's visibility count (Ren & Zhu, 2024), AhaKV traces the bias this removes to the causal mask (Gu et al., 2025), and Section 5 explains the division from the bound. KVzip scores tokens by how well the cache reconstructs the context, so that one compressed cache serves many queries (Kim et al., 2025), and SCBench finds that sub-linear memory methods suffer when many queries share a context (Li et al., 2025). Position-based heuristics match or exceed learned importance scoring (Steele, 2026), and under query-agnostic protocols few published methods beat trivial baselines (Luo et al., 2026).

Expectation models and compressibility theory. Expected Attention scores a cached token by its mean attention under a Gaussian model of unseen queries (Devoto et al., 2025), and lookahead methods score the cache with pseudo or predicted future queries (Wang et al., 2025a; Ahn et al., 2026). Lemma 1 bounds any single ordering such estimates produce by the agreement of the queries served (further work in Appendix A).

8 SCOPE OF THE FINDINGS

The findings hold for continuation queries. Every correlation outside Section 4.1 and every regret uses positions inside the document, and on real questions prefill moments predict agreement only on Olmo-3-7B and Olmo-3-32B. Output quantities are changes in a head's attention output against a per-query oracle. Retained sets are per query head except in the last block of Table 2, where one set per KV head keeps the score ahead of Expected Attention. Task accuracy enters through LongBench on the instruct Llama-3.1-8B, with a secondary check on Llama-3.2-3B-Instruct (Appendix P). The 48 questions follow one template about one document, and their agreement sits near its own null. Agreement covers six base checkpoints at $N = 8192$.

9 CONCLUSION

Two measured properties of a head's queries predict, before decoding, which heads can be evicted blind on continuation queries, and the model family decides whether they carry to appended questions. The mean-direction score halves the regret of H2O as released on continuation queries, and Section 6 draws five implications for eviction systems.

ETHICS STATEMENT

This work measures, on public text, how far the queries of open-weight language models agree about KV cache eviction cost. No model is trained, fine-tuned, or steered. No capability is elicited, and no human subjects, personal data, or annotation labor are involved. The deployability analysis argues for more conservative eviction guarantees (held-out violation rates and inflated budgets) than current practice, which if anything reduces the risk of silently degraded model behavior in deployed systems. The nine-document corpus consists of public-domain literature, an economics treatise, open-source code, and a public-domain French novel. No private or sensitive text is used.

REPRODUCIBILITY STATEMENT

The anonymous repository at https://anonymous.4open.science/r/kv-incompressibility-A706/ holds the analysis code, every result file and the record of predictions with its amendments (Appendix R). No statistic in the paper is typed by hand: each one is produced by that code from a result file in the repository, through one code path for the 2024 models and the 2025 to 2026 models alike. The checkpoints are public and are listed in Section 3 and Appendices C, L and P. Post-RoPE keys, values and probe-position queries are captured once and then read by every experiment, apart from the further question-regime configurations (Appendix L) and the single ported policy that needs forward passes (Appendix N). Context length is $N = 8192$ for the six captures and $N = 4096$ for the 8B consistency check. Each unit has a fixed seed, and each control gives its number of draws. For every experiment in the record, the statistic, null, thresholds and permitted repairs were committed to the repository before it ran, a record that is not independently timestamped, and models added later follow the same rules. Documents 1 to 8 come from public Project Gutenberg texts and public source code. Captures and evaluations were run on NVIDIA A100 GPUs, and the smaller models on a single 16 GB GPU.

AI USE STATEMENT

Large language model assistants served this work in three ways. Writing: they edited the text for clarity, concision and consistency. Research ideation and execution: they helped generate and refine hypotheses and experimental designs, wrote and debugged the capture, baseline-port and analysis code, and helped interpret results. Drafting: they produced first versions of some sections, which we rewrote. The authors checked every hypothesis, design, analysis and claim, and every reported number is emitted by the released code run on the released result files. The assistants were used for nothing else. The authors are responsible for the full content of the paper, including any text, claims and artifacts produced with generative AI.

REFERENCES

Jinwoo Ahn, Ingyu Seong, Akhil Kedia, Junhan Kim, Hyemi Jang, Kangwook Lee, and Yongkweon Jeon. LookaheadKV: Fast and accurate KV cache eviction by glimpsing into the future without generation. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. doi: 10.48550/arXiv.2603.10899. arXiv:2603.10899.

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, 2024. doi: 10.18653/v1/2024.acl-long.172.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024. doi: 10.48550/arXiv.2406.02069.

Herbert H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359, 1973. doi: 10.1016/S0022-5371(73)80014-3.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, 1972. ISBN 0-471-18850-6.

Alessio Devoto, Yu Zhao, Simone Scardapane, and Pasquale Minervini. A simple and effective L_2 norm-based strategy for KV cache compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 18476–18499, 2024. doi: 10.18653/v1/2024.emnlp-main.1027.

Alessio Devoto, Maximilian Jeblick, and Simon Jégou. Expected attention: KV cache compression by estimating attention from future queries distribution. *arXiv preprint arXiv:2510.00636*, 2025. doi: 10.48550/arXiv.2510.00636.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 125462–125498, 2025. doi: 10.52202/085713-3773. URL https://arxiv.org/abs/2407.11550.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2310.01801.

Raghavv Goel, Junyoung Park, Mukul Gagrani, Dalton Jones, Matthew Morse, Harper Langston, Mingu Lee, and Chris Lott. CAOTE: KV cache selection for LLMs via attention output error-based token eviction. *arXiv preprint arXiv:2504.14051*, 2025. doi: 10.48550/arXiv.2504.14051.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783.

Yifeng Gu, Zicong Jiang, Jianxiu Jin, Kailing Guo, Ziyang Zhang, and Xiangmin Xu. AhaKV: Adaptive holistic attention-driven KV cache eviction for efficient inference of large language models. *arXiv preprint arXiv:2506.03762*, 2025. doi: 10.48550/arXiv.2506.03762.

Themistoklis Haris and Krzysztof Onak. Compression barriers for autoregressive transformers. *arXiv preprint arXiv:2502.15955*, 2025. doi: 10.48550/arXiv.2502.15955.

Lukas Haverbeck, Carmen Amo Alonso, Andres Felipe Posada-Moreno, Sebastian Trimpe, and Marco Pavone. The risk of KV cache compression. *arXiv preprint arXiv:2607.01520*, 2026. doi: 10.48550/arXiv.2607.01520.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. doi: 10.48550/arXiv.2401.18079. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/028fcbcf85435d39a40c4d61b42c99a4-Abstract-Conference.html.

Charles M. Judd, Jacob Westfall, and David A. Kenny. Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1):54–69, 2012. doi: 10.1037/a0028347.

Beomgu Kang, SoJin Yun, Hojoon Kim, and Hyunseok Seo. Risk-controlled KV-cache eviction: From memory budgets to risk targets. *arXiv preprint arXiv:2609.27981*, 2026. doi: 10.48550/arXiv.2609.27981.

Norbert L. Kerr. Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217, 1998. doi: 10.1207/s15327957pspr0203_4.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W. Lee, Sangdoo Yun, and Hyun Oh Song. KVzip: Query-agnostic KV cache compression with context reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.48550/arXiv.2505.23416. arXiv:2505.23416.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. SCBench: A KV cache-centric analysis of long-context methods. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2412.10319. arXiv:2412.10319.

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 37 (NeurIPS)*, 2024. doi: 10.48550/arXiv.2404.14469. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html. arXiv:2404.14469.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. doi: 10.48550/arXiv.2305.17118. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a452a7c6c463e4ae8fbdc614c6e983e6-Abstract-Conference.html.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. doi: 10.48550/arXiv.2402.02750. URL https://proceedings.mlr.press/v235/liu24bz.html.

Daming Luo, Christy Liang, and Junyu Xuan. How query visibility changes KV-cache compression rankings: A matched-budget audit. *arXiv preprint arXiv:2607.11942*, 2026. doi: 10.48550/arXiv.2607.11942.

Weian Mao, Xi Lin, Wei Huang, Yuxin Xie, Tianfu Fu, Bohan Zhuang, Song Han, and Yukang Chen. TriAttention: Efficient long reasoning with trigonometric KV compression. *arXiv preprint arXiv:2604.04921*, 2026. doi: 10.48550/arXiv.2604.04921.

NVIDIA. kvpress: LLM KV cache compression made easy. Software library, 2024. URL https://github.com/NVIDIA/kvpress.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 18724–18741, 2024. doi: 10.18653/v1/2024.emnlp-main.1043.

Junyoung Park, Dalton Jones, Matthew J. Morse, Raghavv Goel, Mingu Lee, and Chris Lott. KeyDiff: Key similarity-based KV cache eviction for long-context LLM inference in resource-constrained environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.48550/arXiv.2504.15364. arXiv:2504.15364.

Qwen et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Siyu Ren and Kenny Q. Zhu. On the efficacy of eviction policy for key-value constrained generative language model inference. *arXiv preprint arXiv:2402.06262*, 2024. doi: 10.48550/arXiv.2402.06262.

Ayan Sengupta, Siddhant Chaudhary, and Tanmoy Chakraborty. Value-guided KV compression for LLMs via approximated CUR decomposition. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 88198–88229, 2025. doi: 10.52202/085713-2656. URL https://arxiv.org/abs/2509.15038.

Brady Steele. On the limits of learned importance scoring for KV cache compression. *arXiv preprint arXiv:2601.14279*, 2026. doi: 10.48550/arXiv.2601.14279.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

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 (ICML)*, 2024. doi: 10.48550/arXiv.2406.10774. URL https://proceedings.mlr.press/v235/tang24l.html.

Team Olmo. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. doi: 10.48550/arXiv.2512.13961.

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 30 (NeurIPS)*, pp. 5998–6008, 2017. doi: 10.48550/arXiv.1706.03762. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Yixuan Wang, Shiyu Ji, Yijun Liu, Yuzhuang Xu, Yang Xu, Qingfu Zhu, and Wanxiang Che. Lookahead Q-Cache: Achieving more consistent KV cache eviction via pseudo query. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025a. doi: 10.18653/v1/2025.emnlp-main.1732.

Zhican Wang, Hongxiang Fan, Haroon Waris, Gang Wang, Zhenyu Li, Jianfei Jiang, Yanan Sun, and Guanghui He. VEDA: Efficient LLM generation through voting-based KV cache eviction and dataflow-flexible accelerator. In *Proceedings of the 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7, 2025b. doi: 10.1109/DAC63849.2025.11132861. URL https://arxiv.org/abs/2507.00797.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2309.17453. arXiv:2309.17453.

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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2410.10819.

Jiaming Yang, Chenwei Tang, Liangli Zhen, and Jiancheng Lv. Rethinking KV cache eviction via a unified information-theoretic objective. *arXiv preprint arXiv:2604.25975*, 2026. doi: 10.48550/arXiv.2604.25975.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. doi: 10.48550/arXiv.2306.14048. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

APPENDIX CONTENTS

A Further related work . . . 14
B The model's three stated limits . . . 14
C The in-sample per-unit scatter . . . 14
D Estimator details and the insufficiency of per-token marginals . . . 15
E Attention share of the scored region . . . 15
F Relation to each published finding . . . 16
G Per-document replication in full . . . 16
H Deployability in full . . . 16
I Controls in full . . . 18
J Measurements on the level deficit . . . 19
K The allocation experiment in full . . . 19
L The query regime in full . . . 20
M The practical implications in full . . . 20
N The published policies and their verification . . . 21
O The score and the decomposition of its gain . . . 22
P Task accuracy on LongBench . . . 23
Q Per-head inflation factors from predicted agreement . . . 24
R The registered record . . . 24

A FURTHER RELATED WORK

Lookahead Q-Cache scores the cache with low-cost pseudo lookahead queries (Wang et al., 2025a), and LookaheadKV trains modules that predict the importance a future response would assign (Ahn et al., 2026). The minimax risk of KV compression is characterized in expectation over a query distribution (Haverbeck et al., 2026), where our account is per query. Attention-based generation needs $\Theta(nd)$ space in the worst case without structural assumptions (Haris & Onak, 2025), where our bound is stated per head in a measured statistic of its queries. Risk-controlled eviction certifies a policy on calibration data with a finite-sample guarantee (Kang et al., 2026), the kind of held-out statement Section 5 finds necessary. Our leave-one-out cost coincides with CAOTE (Goel et al., 2025) for one token.

B THE MODEL'S THREE STATED LIMITS

(i) The model is a direction model with a fixed norm per head. Measured per-head norm variation is small (median CV 0.065) and is not modeled. A query norm mixture, a candidate mechanism for the level deficit, is excluded by its own registered gate on that number. (ii) The residual e is isotropic in $\mu^{\perp}$. It is not, and this is the limit that the level overprediction of Section 4 is made of: the measured residual covariance has participation ratio 10.6 against 21.9 for a matched isotropic null, and restoring it closes the deficit (Appendix J). (iii) All quantities are estimated from prefill queries. Whether they transfer across documents is measured rather than assumed, the cross-document matrix of Section 4 is that measurement, and it transfers the moments only, never the cache.

C THE IN-SAMPLE PER-UNIT SCATTER

Figure 5 is the in-sample scatter for Llama-3.2-1B and Qwen2.5-1.5B, with parameters and evaluation sharing every query. Its Spearman correlations, $+0.950$ and $+0.903$, sit within 0.01 of the content-separated values of Table 1, the measurement in Section 4 that shared sampling contributes little on this design. A capture of Llama-3.1-8B at half the context reproduces the ordering ($+0.914$ over 1024 units) at a query spacing where agreement is inflated, so it is a consistency check rather than a scale result, and the six models of Table 1 reach 32B at full context. Centering each head's agreement across the nine documents of Appendix G leaves only content-induced movement, which the model tracks at $+0.539$ on disjoint query halves against $+0.876$ when the halves share queries.

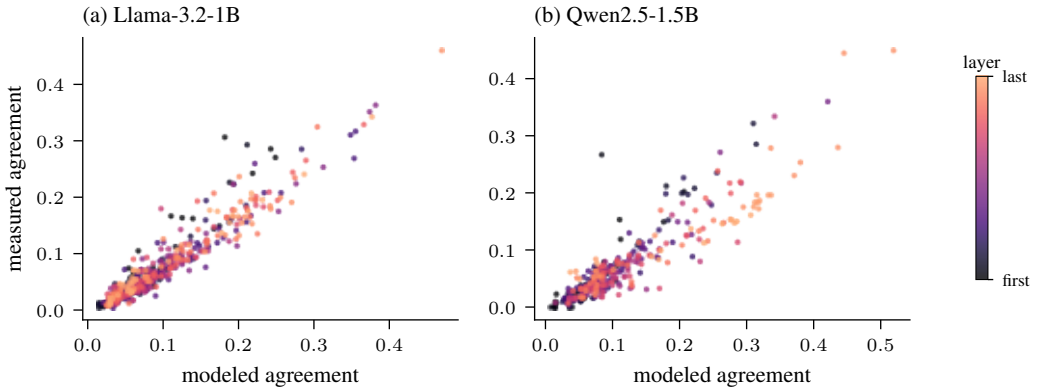

Figure 5: In-sample measured against modeled cross-query agreement per (layer, head) unit, **(a)** Llama-3.2-1B and **(b)** Qwen2.5-1.5B, colored by relative depth, parameters and evaluation sharing the 33 queries of Section 3. Content-separated values of record: $+0.942$ / $+0.900$. The model has zero fitted parameters.

D Estimator details and the insufficiency of per-token marginals

The per-unit statistic averages the median pairwise Jaccard over $B \in \{32, 64, 128\}$, and pairs exclude self pairs and same-position pairs. Two robustness checks follow. The position grid ends in a short run of consecutive positions (8 queries, 28 of 528 pairs). Excluding those pairs changes the statistic by $+0.0019$ at rank correlation 0.9981, so the tail cluster is not influential under the median over pairs estimator. Replacing the median over pairs by the mean gives rank correlation 0.986 with the reported statistic.

Why per-token marginals are insufficient. Under Eq. 3 the logits of two tokens under one query event share the residual draw e: $\mathrm{Cov}(z_s, z_t) \propto (1 - \bar{c})\,(P_\perp k_s)^\top (P_\perp k_t)$, the Gram structure of the perpendicular keys. A top B set is a joint order statistic of $\{z_t\}$, and joint order statistics are not determined by marginals: two heads with identical per-token means and variances but different perpendicular key Gram matrices produce different cross-query agreement, because positively correlated logits move together across query events (stable rankings, high overlap) while uncorrelated logits reorder freely. This is why the simulation draws query events, one residual per event, shared across tokens, rather than sampling each token's marginal, and why no per-token closed form over (m_t, σ_t) can substitute for it. The scalar baseline results of Section 4 are the empirical face of this argument: head-level summaries, including nonlinear functions of them, do not recover what the joint simulation predicts. Simulated events are allocated round robin over the real query positions so that simulated agreement inherits real visibility windows.

E Attention share of the scored region

The share of attention the mid cache carries varies by more than an order of magnitude across heads (p10 to p90 over the three 2024 models 0.004 to 0.127, medians 0.019 / 0.030 / 0.022). Share correlates with agreement at $+0.469$ / $+0.474$, so the result is not confined to heads carrying little attention: top-decile heads by share have $3.5\times$ / $6.6\times$ the median agreement of bottom-decile heads and $21\times$ / $39\times$ the median damage, so the same heads both transfer better and have far more to lose. At the low extreme, 5 Qwen2.5-1.5B heads carry under 10^{-4} of their attention in the scored region and are pure sink and recency routers.

F RELATION TO EACH PUBLISHED FINDING

Section 7 positions this paper against the literature in prose. Table 3 does it claim by claim, stating for each published finding what our measurements add and at what width the evidence supports it. The last column separates supplying a mechanism that predicts an observation from proving its cause, and supplying a quantity from showing that it helps.

Published finding	What our measurement adds	Relation
Position heuristics match or exceed learned importance scoring (Steele, 2026)	A query-free ordering is bounded by cross-query agreement, which no first moment or position score observes. Agreement varies by more than $10\times$ across heads	Mechanism consistent with their negative
Rankings flip between query-aware and query-agnostic protocols (Luo et al., 2026)	Agreement depends on query spread, and densely sampled query positions measure near-duplicate stability (Section 4)	Mechanism supplied
Future attention estimated from a query distribution (Devoto et al., 2025)	Their estimate coincides with our first moment up to normalization. The object governing agreement is the second	Positioned rather than contested
Per-head budget allocation named open (Mao et al., 2026)	Per-head agreement is measurable, predictable from two moments, and predicts what per-head selection retains (Section 4). It is not an allocation dial (Section 5)	Quantity supplied for selection. Negative for allocation

Table 3: Published claims and what this paper's measurements add. "Mechanism consistent with" means we supply an account that predicts their observation. It does not mean we prove their cause.

G PER-DOCUMENT REPLICATION IN FULL

Table 4 gives the unit-level correlation for each of the nine documents, with cluster bootstrap intervals and each document's median $\bar{c}$. The near constancy of $\bar{c}$ across documents (0.708 to 0.727) is reported as an observation: head query geometry appears to be set principally by the model rather than by the text.

Document	ρ	95% CI	ρ (demeaned)	median $\bar{c}$
doc 0 (QA)	$+0.950$	$[+0.926, +0.963]$	$+0.951$	0.720
doc 1	$+0.951$	$[+0.933, +0.959]$	$+0.943$	0.713
doc 2	$+0.947$	$[+0.930, +0.955]$	$+0.938$	0.720
doc 3	$+0.962$	$[+0.943, +0.972]$	$+0.943$	0.720
doc 4	$+0.952$	$[+0.932, +0.962]$	$+0.942$	0.727
doc 5	$+0.949$	$[+0.934, +0.959]$	$+0.940$	0.715
doc 6	$+0.947$	$[+0.928, +0.960]$	$+0.942$	0.725
doc 7 (code)	$+0.947$	$[+0.921, +0.961]$	$+0.948$	0.708
doc 8 (French)	$+0.943$	$[+0.925, +0.955]$	$+0.929$	0.717

Table 4: Cross-document replication, Llama-3.2-1B, parameters remeasured per document.

H DEPLOYABILITY IN FULL

Absolute output error. The bound constrains what a set shares with a query's needs and says nothing about how much a head has to lose, so agreement need not order absolute output error. It orders it with a positive sign on each of the three 2024 models, against a registered negative one: with profile concentration partialled out the correlation is $+0.139$ to $+0.281$ on Llama-3.2-3B. Mid-cache attention mass is the stake, correlating on Llama-3.2-1B and Qwen2.5-1.5B with damage at $+0.58$ to $+0.91$ and with agreement at $+0.469$ / $+0.474$. Conditional on mid-cache mass share, predicted agreement orders regret at -0.244 to -0.395 across budgets on Llama-3.2-3B with every interval excluding zero (exploratory in origin, confirmed on a held-out model, and the confirming estimate is the one reported).

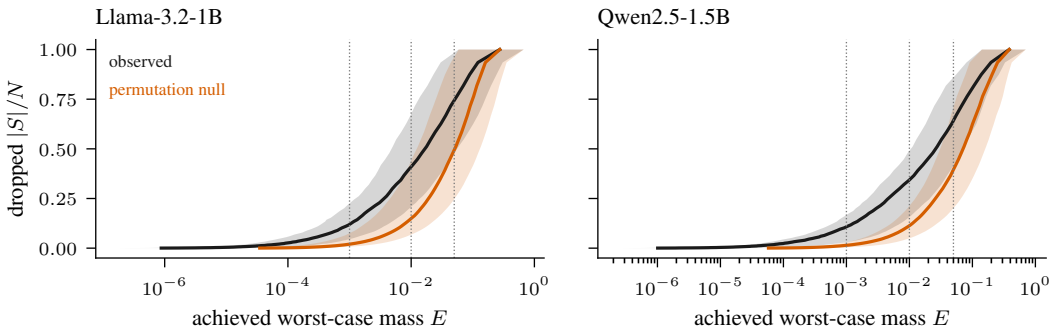

Figure 6: Droppable fraction against achieved worst-case cumulative attention mass per (layer, KV head) on Llama-3.2-1B and Qwen2.5-1.5B, median and 10th to 90th percentile band, observed (black) and per-row permutation null (orange). Dotted verticals: the three reported budgets.

Droppable fraction against its null. At $E = 10^{-2}$ the median droppable fraction is 0.407 (Llama) and 0.342 (Qwen) under the worst-case mass ordering, against per-row permutation nulls of 0.149 / 0.113: query-invariant droppability is real structure rather than marginals. A greedy ordering that grows the set by the smallest marginal increase of the max row cumulative reaches 0.508 / 0.418 at the same budget, so the sup-based convention understates the droppable set by roughly a fifth.

Tables 5 and 6 give held-out violation rates and retained set sizes as a function of the budget inflation factor k (select at E/k, measure violations against E) for all six models. The in-sample rate is structurally zero at $k = 1$ and is not tabulated. The Llama-3.2-3B cells come from a re-extraction of its capture on different hardware. The same code, run on a re-extraction of the Llama-3.2-1B capture, reproduces the Llama-3.2-1B cells of record to within 0.31 percentage points with every k^* equal, so the check is a tolerance and not bitwise identity.

Sup ordering and the greedy gap. Ordering tokens by $\sup_r w_{t,r}$ (their worst-case single-row mass) is optimal for a union bound $\sum_{t\in S} \sup_r w_{t,r} \le E$, but the deployed constraint is tighter: $\sup_r \sum_{t\in S} w_{t,r} \le E$, the maximum accumulated mass in any one row. The two coincide only if each token's mass concentrates in the same row, precisely the case of perfectly correlated queries. When different tokens load different rows, a set can be grown far beyond the union bound by combining tokens whose row profiles are disjoint, which is what the greedy ordering (add the token with the smallest marginal increase of the max row cumulative) exploits. The gap between the two orderings ($0.407 \to 0.508$ at $E = 10^{-2}$ on Llama, and $0.342 \to 0.418$ on Qwen) is therefore a direct measurement of cross-query correlation structure in the weight matrix: zero gap would mean fully aligned rows, and the observed ∼20 to 25% understatement means that sup-score orderings understate the droppable set by that fraction. Worst-case selection over a KV group is affected by within-group residual correlation (Table 7): correlated residuals raise the group supremum, so the independent residual model underpredicts droppable mass, the conservative direction.

Model	E	$k = 1.0$	$k = 1.5$	$k = 2.0$	$k = 2.5$	$k = 3.0$	$k = 4.0$
Llama-3.2-1B	10^{-3}	25.7 / 12.2	15.6 / 9.3	10.6 / 7.8	7.8 / 6.8	6.1 / 6.0	4.0 / 5.0
Llama-3.2-1B	10^{-2}	18.3 / 40.9	9.2 / 34.1	5.5 / 30.4	3.6 / 26.5	2.6 / 24.2	1.6 / 21.1
Llama-3.2-1B	$5 \cdot 10^{-2}$	9.8 / 74.5	4.7 / 65.1	2.7 / 58.9	1.9 / 54.4	1.6 / 50.5	1.1 / 44.8
Llama-3.2-3B	10^{-3}	20.5 / 11.0	10.3 / 8.5	6.5 / 7.1	4.5 / 6.1	2.9 / 5.4	1.7 / 4.5
Llama-3.2-3B	10^{-2}	15.2 / 40.4	6.3 / 33.1	3.2 / 28.2	2.0 / 25.2	1.4 / 22.5	0.9 / 18.9
Llama-3.2-3B	$5 \cdot 10^{-2}$	8.9 / 79.5	3.2 / 69.1	1.7 / 61.7	1.1 / 55.9	0.9 / 51.2	0.6 / 44.8
Qwen2.5-1.5B	10^{-3}	20.3 / 10.8	11.2 / 8.6	7.3 / 7.1	5.0 / 6.1	3.6 / 5.3	2.7 / 4.5
Qwen2.5-1.5B	10^{-2}	18.5 / 34.6	8.1 / 29.5	4.4 / 26.0	3.0 / 23.1	2.2 / 21.0	1.3 / 18.3
Qwen2.5-1.5B	$5 \cdot 10^{-2}$	12.8 / 64.3	6.0 / 55.3	3.6 / 49.1	2.1 / 45.4	1.4 / 42.2	0.8 / 37.8

Table 5: Held-out violation rate (%) / retained droppable fraction (%) at budget E/k, evaluated against E, per model and budget, 2024 set.

Model	E	$k = 1.0$	$k = 1.5$	$k = 2.0$	$k = 2.5$	$k = 3.0$	$k = 4.0$
Olmo-3-32B	10^{-3}	9.8 / 1.6	4.1 / 1.1	2.0 / 0.9	1.4 / 0.7	0.9 / 0.6	0.5 / 0.5
Olmo-3-32B	10^{-2}	8.9 / 11.0	2.7 / 8.1	1.3 / 6.3	0.7 / 5.3	0.4 / 4.5	0.2 / 3.5
Olmo-3-32B	$5 \cdot 10^{-2}$	7.7 / 35.3	2.2 / 26.8	0.8 / 22.0	0.4 / 18.5	0.2 / 16.2	0.1 / 13.0
Olmo-3-7B	10^{-3}	13.3 / 4.4	6.7 / 3.3	4.2 / 2.6	2.9 / 2.2	2.0 / 1.8	1.4 / 1.5
Olmo-3-7B	10^{-2}	11.8 / 23.0	5.5 / 17.6	3.0 / 14.3	1.9 / 12.0	1.3 / 10.5	0.7 / 8.6
Olmo-3-7B	$5 \cdot 10^{-2}$	9.3 / 57.5	3.9 / 46.6	2.0 / 39.8	1.1 / 35.2	0.8 / 31.7	0.5 / 26.8
Qwen3.5-9B-Base	10^{-3}	25.0 / 6.7	13.7 / 5.2	9.2 / 4.4	7.4 / 3.7	6.1 / 3.2	4.8 / 2.7
Qwen3.5-9B-Base	10^{-2}	26.1 / 24.2	15.6 / 19.5	9.9 / 16.9	7.0 / 14.6	5.1 / 13.4	4.0 / 11.5
Qwen3.5-9B-Base	$5 \cdot 10^{-2}$	23.2 / 53.4	12.0 / 44.7	8.0 / 39.4	6.0 / 35.5	5.0 / 32.5	3.2 / 27.8

Table 6: As Table 5, for the 2025 to 2026 set, on their full-attention layers.

Model	within KV group	across group, same layer	across layer
Llama-3.2-1B	$+0.286$ $[+0.150, +0.452]$	$+0.060$ $[+0.023, +0.122]$	$+0.055$ $[+0.023, +0.114]$
Qwen2.5-1.5B	$+0.196$ $[+0.087, +0.312]$	$+0.036$ $[+0.009, +0.107]$	$+0.042$ $[+0.014, +0.109]$

Table 7: Cross-head residual correlation at three scopes (median and IQR over groups). Within-group correlation is K/V sharing structure. Across-group values are indistinguishable from the across-layer baseline.

I CONTROLS IN FULL

Table 8 collects the ablation floors and controls of Section 4.

Controls on the retained-cost link. Over the same models and budgets, a marginals-preserving shuffle collapses the correlation between predicted agreement and retained cost to $+0.060$ to $+0.219$, a wrong-head pairing collapses it to $+0.110$ to $+0.264$, and retained cost correlates with measured agreement at $+0.802$ to $+0.909$, below the 0.99 bar declared in advance for a restatement.

Supporting diagnostics. Ranking units after removing each layer's mean gives the same answer as the raw ranking ($+0.951$ / $+0.908$), and within-layer intraclass correlations of measured agreement are low (0.11 / 0.27). A harsher half-data disjoint split ($+0.901$ / $+0.835$) recovers the headline once its parameter estimation noise is accounted for (matched arm $+0.931$ / $+0.880$).

Ablations toward nearby wrong models. A random mean direction (same $\bar{c}$, same norms) drops the correlation to $+0.581$ / $+0.457$. Keeping everything else identical, including the residual subspace, and replacing only where the mean direction points, rescaled so the dispersion of mean logits across tokens matches the measured value, recovers most of the loss ($+0.821$ / $+0.782$) but remains below the full model: the dispersion of $\{k_t^\top \mu\}$ is the dominant channel through which the mean direction acts on agreement, and where μ points contributes a further increment that no scalar summary of it captures. Distant floors are lower still: an isotropic query on the same cache gives $+0.107$ / $+0.371$, and keys only with $\bar{c}$ frozen at its global median gives $+0.714$ / $+0.564$.

Scalar regressions against the simulation. Predicting measured agreement from head-level scalars alone, profile concentration (participation ratio and entropy), $\bar{c}$, mean perpendicular key norm, and mean logit dispersion, reaches a leave-one-layer-out CV Spearman of $+0.743$ / $+0.579$ with gradient boosting (the learner of record, against $+0.712$ / $+0.350$ for a linear fit and $+0.810$ / $+0.392$ for an under-tuned quadratic ridge on logit-transformed agreement), against the simulation's $+0.950$ / $+0.903$. The comparison is tilted toward the scalars: the concentration covariates are computed from the measured cost profiles, so the regression sees target side information the simulation never sees, and still trails by 0.2 to 0.3 of Spearman. Learners ran at fixed hyperparameters (no tuning inside the fold). The gap is far larger than tuning plausibly recovers.

Profile concentration control. Heads with concentrated cost profiles could agree across queries for trivial reasons. The bulk concentration of each unit's mean cost profile (participation ratio)

correlates with measured agreement at -0.353 / -0.301. Partialling it out leaves the model and measurement correlation at $+0.944$ / $+0.900$ (layer demeaned $+0.941$ / $+0.902$). Component arms locate the rest: keys only reaches $+0.714$ / $+0.564$, the scalar $\bar{c}$ alone $+0.638$ / $+0.430$, and the full model exceeds any additive combination of the two, consistent with the prediction living in the interaction of each unit's own keys with its own query moments.

Model	random μ	dispersion-matched μ	GB on scalars	partial | PR	isotropic
Llama-3.2-1B	$+0.581$	$+0.821$	$+0.743$	$+0.944$	$+0.107$
Qwen2.5-1.5B	$+0.457$	$+0.782$	$+0.579$	$+0.900$	$+0.371$

Table 8: Controls and floors: correlation with measured agreement under each construction, per model.

J Measurements on the level deficit

Simulated top B overlap exceeds measured overlap by roughly 1.6 to $1.7\times$ (Llama-3.2-1B) and 2.2 to $2.3\times$ (Qwen2.5-1.5B), stated at no more precision than the pair sampling noise of the estimator supports. At the $E = 10^{-2}$ droppable set, output perturbation exceeds the independent-directions closed form by 3.59 / 3.59. Re-pairing weights with values within the dropped set reproduces this (3.91 / 3.85), so the excess comes from the value geometry alone, dropped centered values share a direction, and nothing about which weight sits on which value contributes. Weight profiles are strongly concentrated throughout (median per-row Gini 0.901 / 0.898). Candidates tested and excluded for the per-head deficit: query norm mixtures (excluded by their registered gate, median CV of per-head query norms 0.065) and cross-head residual correlation (Table 7).

Residual anisotropy. The model draws the residual uniformly on the sphere in $\mu^{\perp}$. Estimating the residual covariance Σ_e per unit of Llama-3.2-1B from the same 33 queries that give μ and $\bar{c}$, its participation ratio is 10.6 of 63 residual dimensions, against 21.9 for 33 isotropic draws estimated the same way, so the residual is anisotropic by a factor of two beyond what the estimator's own small sample produces. Redrawing $e \sim \mathcal{N}(0, \Sigma_e)$ and renormalizing, with every other element of the simulation unchanged, moves the level ratio from 1.34 to 0.99 and the unit ordering from +0.942 to +0.971 on the content-separated split.

The reading is tested against an outcome-level null whose criterion was registered in advance: if a covariance estimated from 33 fresh isotropic draws per unit also brought the ratio to 1.05 or below, the closure would be the rank deficiency of a 33-sample covariance in 63 dimensions rather than anisotropy, and the reading would not hold. That null returns 1.36 with ordering +0.940, which is the base arm's value, so the criterion is not met and the reading holds.

The anisotropic arm does not replace the model of record. Σ_e is a $d \times d$ object estimated per unit, where the model of record is two measured scalars, so the anisotropic arm explains the deficit without being the model the paper proposes.

K The allocation experiment in full

Protocol: Llama-3.2-1B, prefill $4{,}096$ tokens, evict mid cache per (layer, KV head) by a shared observed attention score (64-query observation window, sink and recency always kept), teacher-forced log perplexity over the next 512 tokens. Three documents (narrative QA, Python source, French prose) $\times$ eight nonoverlapping prefill offsets. Registered arms: uniform, layer-wise pyramid, proportional to predicted agreement, and proportional to measured agreement (the oracle ceiling). Table 9 gives the single-offset values. Paired per-arm effects against uniform with bootstrap intervals over the 24 (document, offset) cells are the quantities Section 5 summarizes. A registered numerics control, the full-cache baseline computed with and without the eviction mask, returns exactly zero difference, which bounds fp16 accumulation effects.

Exploratory arms. Two further arms, inverse ($\propto 1/J$) and coverage ($\propto 1-J$), were added after the registered proportional arms lost to uniform, on the hypothesis that disagreeing heads need

budget to cover the union of their queries' needs. They are exploratory in the sense of Kerr (1998) and appear only here: under the paired analysis the inverse arm is significantly worse than uniform at every budget, and the coverage arm is indistinguishable from it (-0.0005 [$-0.0010, +0.0000$] at $F = 0.5$). A small gain of mild eviction over the full cache at a single offset reverses under the 24-cell analysis ($+0.008$ [$+0.004, +0.013$] at $F = 0.5$), so single-offset cells are not reliable.

Doc	F	uniform	pyramid	predicted	oracle	inverse	coverage
doc 0	0.1	$+0.0198$	$+0.0206$	$+0.0210$	$+0.0217$	$+0.0214$	$+0.0194$
doc 0	0.25	$+0.0057$	$+0.0026$	$+0.0099$	$+0.0098$	$+0.0117$	$+0.0035$
doc 0	0.5	-0.0035	-0.0044	-0.0009	-0.0010	-0.0005	-0.0027
doc 7	0.1	$+0.1243$	$+0.1217$	$+0.1334$	$+0.1323$	$+0.1103$	$+0.1245$
doc 7	0.25	$+0.0527$	$+0.0580$	$+0.0848$	$+0.1046$	$+0.0706$	$+0.0535$
doc 7	0.5	$+0.0098$	$+0.0202$	$+0.0364$	$+0.0550$	$+0.0277$	$+0.0085$
doc 8	0.1	$+0.0214$	$+0.0270$	$+0.0207$	$+0.0257$	$+0.0349$	$+0.0225$
doc 8	0.25	$+0.0075$	$+0.0078$	$+0.0073$	$+0.0080$	$+0.0136$	$+0.0073$
doc 8	0.5	-0.0000	$+0.0021$	-0.0011	-0.0022	$+0.0035$	$+0.0002$

Table 9: Δ log perplexity vs. full cache (positive = worse) for six budget allocation arms at matched total budget F, with a shared eviction score in every arm.

L THE QUERY REGIME IN FULL

The 48 questions, six per 1,024-token segment in a fixed form mix, were fixed before capture and are released with the code. Each question's tokens are forwarded against the stored cache itself on every dense stack, loaded from the capture into the model's own key-value cache with positions continuing from N, so both regimes are computed against the same tensors by construction. On the hybrid stacks of the further configurations the stored capture holds no state for the linear-attention and sliding-window layers. Each question is therefore forwarded against a fresh copy of the model's own prefill cache of the document, a path declared before any capture, and the analysis covers the layers whose cache grows with the context. A single forward pass over the concatenated document and question is not used, because it computes the document's keys in the presence of the question, which is not the object a serving system holds, and it does not reproduce the stored cache. A fresh document-only prefill does not reproduce the stored cache bitwise either (max $|\Delta K|$ is 2.1e-02 of key scale), a consequence of bf16 kernel nondeterminism across library versions. The quantity the analysis depends on is unaffected: query directions reproduce at cosine 0.99999 median.

Cross-query structure in the question regime. Measured agreement among the questions is 0.497 against a marginals-preserving null of 0.437, an excess of $+0.060$, where the distance-matched continuation arm sits -0.001 above its own null. So in the question regime most of the top B overlap is the same tokens dominating every column and only a small part is cross-query structure. At $\bar{c} = 0.955$ the questions are nearly parallel and their retention sets largely coincide, which is the near-duplicate regime of Section 4, reached through content homogeneity rather than through position spacing. 48 questions in one template about one document are themselves a narrow query distribution, and part of that concentration is our question set rather than user questions in general. This does not change the registered primary outcome, since an unusually homogeneous and therefore easy question set is still not predicted from prefill moments, but it bounds how far the $\bar{c}$ numbers travel (Section 8).

M THE PRACTICAL IMPLICATIONS IN FULL

Section 6 gives each implication in one line. The full versions carry the numbers.

1. **Divide accumulated attention by the visibility count.** Under a causal mask, H2O as released counts an early token over more prefill rows than a late one. Dividing each token's accumulated score by the number of rows that could see it, the mean attention score of Ren & Zhu (2024), cuts the regret of H2O as released by 45% on held-out continuation queries of Llama-3.2-1B (Table 2) and supplies 77% of the gap from H2O to the mean-direction score (Appendix O). The kvpress implementation of H2O already divides. On LongBench, whose questions are appended, the division leaves the LongBench average of H2O as released unchanged within its interval on

		continuation	appended questions				$\bar{c}$ across
configuration	units	in sample	prefill moments	question moments	head identity	reliability	the two regimes
Llama-3.2-1B, registered capture	512	+0.949	+0.202	+0.949	+0.218	+0.927	+0.178
Llama-3.2-1B, fresh capture	512	+0.948	+0.203	+0.949	+0.222	+0.927	+0.178
Qwen2.5-1.5B	336	+0.902	+0.145	+0.979	+0.183	+0.946	+0.392
Qwen3.5-2B-Base	48	+0.920	+0.042	+0.963	+0.147	+0.992	+0.242
Olmo-3-7B	256	+0.907	+0.687	+0.948	+0.596	+0.984	+0.657
Llama-3.1-8B	1024	+0.947	+0.263	+0.964	+0.305	+0.932	+0.180
Qwen3.5-9B-Base	128	+0.903	+0.038	+0.940	−0.015	+0.977	+0.384
Olmo-3-32B	640	+0.963	+0.814	+0.958	+0.808	+0.988	+0.719

Table 10: The question regime per configuration, Spearman ρ over units. Continuation in sample: the model against agreement among the continuation queries its moments came from, computed on the question-regime captures, so it can differ from the in-sample values of Section 4 in the third decimal. Prefill moments (the registered primary, bars 0.50 partial and 0.70 established), question moments and head identity: the three tests of Section 4.1. Reliability: split-half reliability of question agreement. $\bar{c}$ across the two regimes: rank correlation of per-head $\bar{c}$ between continuation queries and questions.

the instruct Llama-3.1-8B at a quarter of the cache kept (Appendix P), so its gain is measured on continuation queries.

2. **State guarantees on held-out queries.** A selection constraint caps in-sample violation at zero by construction, so a matched in-sample evaluation cannot see the failure mode it exists to prevent. On unseen queries the same budgets are violated on 8.9% to 26.1% of head and query blocks at $E = 10^{-2}$ over the six models. The actionable form is the budget inflation factor restoring a stated guarantee: here 1.25 to 4× for violations near 5%, largest on Qwen3.5-9B-Base.
3. **Report the query spread.** Measured agreement is not invariant to the spacing of the queries compared: at 70-token spacing an evaluation reproduces in-sample values exactly, because two such queries share nearly their whole cache (Section 4). Continuation-style adjacent queries and document-level question answering are different regimes and do not transfer. State the minimum cross-query separation, or report agreement at your own spread.
4. **Analyze per head rather than per layer.** Aggregating units to layers moves the same correlation from +0.950 to +0.772 on Llama-3.2-1B, and 88% of the variance in per-head $\bar{c}$ lies within layer (intraclass correlation of $\bar{c}$, 0.12): layer-level analysis discards most of the signal it seeks. This bears on the choice between layer-wise (Cai et al., 2024) and head-wise (Feng et al., 2025) budget allocation, though what per-head agreement informs is selection rather than allocation (Section 5).
5. **Measure query statistics on the queries you will serve.** Any method that estimates a query model from prefill is estimating it on continuation queries, and those are not the queries it will be asked. Across the two regimes the same heads have median $\bar{c}$ of 0.720 and 0.955 and a rank correlation of +0.178 on Llama-3.2-1B (Section 4.1), so the estimate does not carry there, and on the models of Table 10 whether it carries is a measurement on each model.

N THE PUBLISHED POLICIES AND THEIR VERIFICATION

Provenance. The scoring rules are transcribed from released code rather than from the papers. H2O accumulates attention over every prefill row with no pooling and keeps a recency window. SnapKV scores from the last window of rows, average pooled along the key axis. PyramidKV uses SnapKV's score with a budget that falls linearly with depth. StreamingLLM keeps a sink and a recency band and scores nothing. The sources are the PyramidKV authors' KVCache-Factory, which carries all four with grouped query attention support, and the SnapKV repository for the independent check.

Verification. A verification script runs each ported policy and the corresponding upstream class on the same random tensors and compares the retained index sets exactly. All 4 comparisons return

method	partial ρ(predicted agreement, regret | mass) estimate	95% CI	median regret	overlap with our ordering
H2O	-0.527	$[-0.602, -0.436]$	0.0155	0.27
SnapKV	-0.244	$[-0.341, -0.132]$	0.0259	0.07
PyramidKV	-0.148	$[-0.292, +0.048]$	0.0275	0.05
StreamingLLM	-0.023	$[-0.194, +0.178]$	0.0218	0.14

Table 11: Four published policies at one matched budget on 512 units, ported from released code and verified against it. The registered bar was $\rho \leq -0.25$ with an interval excluding zero for at least three of the four, and 1 of 4 meets it. The last column is the overlap between each method's retained set and the mean leave-one-out ordering of Section 4, which is what the agreement bound is a bound on.

identical sets, including SnapKV against both its own repository and KVCache-Factory's copy of it. Scoring runs only after every comparison passes, so no number in Table 11 comes from a port that has not reproduced its source.

Fixed settings. The model, the document, N, the mid-cache scored region and the registered query set are the paper's own, and the comparison is on Llama-3.2-1B alone. Each method selects from prefill only. The one permitted repair declared in the registration was used: the published default bands are replaced by this paper's 25% budget over a 64-token sink and a 256-token recency band, so that a single budget means the same number of retained slots for every method. No other parameter was tuned and no method was rerun after its number was seen.

Model scope. The registration named the three 2024 models and the comparison ran on Llama-3.2-1B, a deviation from the registration. H2O accumulates attention over every prefill row, so scoring it faithfully needs the query vector at every position rather than on the sparse probe grid the captures of record store, and that capture is 876 MB for this model and several times that for the larger one. The released capture code reproduces it for the other two. The deviation moves the reading in one direction: the registered rule asked for three of four methods across three models, so a primary decided on four comparisons rather than twelve is a weaker negative than the registration intended, and the one positive it carries is a single-model result.

The 2025 to 2026 policies. Expected Attention (Devoto et al., 2025), CapKV (Yang et al., 2026), KeyDiff (Park et al., 2025), CurDKV (Sengupta et al., 2025) and KVzip (Kim et al., 2025) are ported from their classes in NVIDIA kvpress 0.5.4 (NVIDIA, 2024) and checked against them numerically before scoring (16 comparisons, all passing). KVzip scores the cache with forward passes of the model and its released scores are used directly. The protocol is that of Table 11, on the same 512 units, budget and bands, with every policy reading the whole prefill including the evaluation positions. Two deviations are given to the baselines. Expected Attention and CapKV are scored per query head, where their release averages over the query heads of a KV group, and the group-averaged forms have median regret 0.0085 and 0.0150. KeyDiff is run one shot as its press does, and its block-wise form reaches 0.0185. Every row of the third block of Table 2 comes from one run, in which H2O as released (0.0154) and the score (0.0065) reproduce their values of record to within 0.7%, and every paired difference is within that run. The score against Expected Attention is -0.00035 $[-0.00052, -0.00023]$, and against every other arm the interval also excludes zero. The released rule averages the rotary rotation of Expected Attention over the positions after prefill, where appended questions sit. Our held-out queries sit inside the prefill, so a sensitivity arm averages it over their positions instead: it reaches 0.0057, below the score by -0.00044 $[-0.00059, -0.00035]$, and is reported beside the released arm, never in its place. Predicted agreement orders regret given mass at the bar of Table 11 for 6 of the 8 arms, all but one-shot KeyDiff and CurDKV.

O THE SCORE AND THE DECOMPOSITION OF ITS GAIN

Arms. Every arm selects from prefill alone at the same budget, over the same mid cache, with the same sink and recency bands, and is scored by the same regret against the same per-query oracle.

The four published policies are the ports of Appendix N, unchanged. The normalized arm is the same accumulated score divided per token by the number of prefill rows that could see it, which is the mean attention score of Ren & Zhu (2024). The mean-direction arms ask a single query $q = \|q\|_{\text{med}}\,\mu$ at the end of prefill and rank either by its attention weight or by the eviction cost of Eq. (1). The random-direction arm is identical with μ replaced by a draw from the sphere at the same norm, one draw per unit from a fixed seed. The ceiling is the mean eviction cost over 33 real queries of the evaluation document that are disjoint from the evaluation set.

Paired differences. Cost against the normalized H2O, -0.00138 [-0.00170, -0.00117]. Cost against the random direction, -0.01309 [-0.01658, -0.00931]. Cost against the attention weight of the same query, -0.00009 [-0.00016, -0.00006]. The normalized H2O against H2O as released, -0.00587 [-0.00669, -0.00505]. All by unit with a cluster bootstrap interval by layer, all excluding zero.

Parameter source. μ, $\bar{c}$ and the median query norm are measured on document 1 and applied against the evaluation document's own keys, values and query positions, which is the cross-document protocol of Section 4. A split within the evaluation document was registered first and replaced before any compute, because the eligible range is about 5,000 tokens and the design places 33 parameter and 33 evaluation queries in it, so every evaluation query would have sat inside the spacing at which Section 4 measures agreement to be inflated. The published policies keep their own protocol, which reads every prefill row including the evaluation positions, so the more generous protocol goes to them.

Registered record. The score, its first-moment ablation, its tail variant and the ceiling were registered with their thresholds before compute. The two registered mechanism tests were not met. Drawing queries around μ at the measured $\bar{c}$ does not lower median regret below the first-moment ablation, which sets $\bar{c} = 1$ (paired difference +0.00031 [+0.00021, +0.00041]), and the tail arm does not lower the 90th percentile of regret below it (+0.00036 [+0.00026, +0.00045]). The amendment that added the tail arm derived, before the arms were run, that expected damage is linear in each token's mean cost, so that the second moment cannot lower the median, and predicted no difference there. The measured difference is small, but its interval excludes zero. The decomposition was registered after the score's numbers were seen and before any of its own were, with those numbers written into the registration, because a gain of this size against an ordering the paper measures at Spearman 0.992 with accumulated attention required a decomposition before it was reported.

P TASK ACCURACY ON LONGBENCH

Protocol. Registered before compute. Seven English subsets of LongBench (Bai et al., 2024) covering every English metric family: NarrativeQA, Qasper, HotpotQA and 2WikiMQA (token F1), GovReport (ROUGE-L), TREC (classification accuracy) and LCC (code edit similarity), each the first 50 samples of its test split in released order. Official prompts, generation lengths and metrics, contexts middle-truncated to 7,500 tokens, the context compressed once and the question asked against the compressed cache. GovReport and LCC have no question, so their whole prompt, closing instruction included, is the compressed context. Greedy decoding. Every arm runs in NVIDIA kvpress, and the compression ratio is the fraction of the context cache removed. Expected Attention, SnapKV (Li et al., 2024), KeyDiff and KVzip are the released kvpress classes at their defaults. H2O as released and the normalized H2O share one implementation that differs only in the division by each token's visibility count, both keeping a 64-token recency window. The kvpress observed-attention press already divides, so it is the normalized arm and not H2O as released. The checkpoint is an instruct model where the paper's models of record are base models, and the kvpress pipeline applies the chat template to every subset, which the official release does not do for TREC and LCC.

Statistics. Per arm, the mean of the official metric over each subset, and the average over subsets. The two registered comparisons are paired by sample: the average over subsets of the per-subset mean difference, with a 95% interval from 10,000 bootstrap resamples of samples within each subset. The normalized H2O against H2O as released is -0.6 [-3.4, $+2.1$] and the mean-direction score against Expected Attention -0.2 [-3.0, $+2.6$], over 350 paired samples each. At ratio 0.5, which decides nothing, the averages are 30.8 for H2O as released, 30.4 for the normalized H2O,

29.7 for SnapKV, 31.0 for Expected Attention, 37.7 for KeyDiff, 29.8 for KVzip and 28.1 for the score, and the two comparisons are -0.3 [-2.7, $+2.0$] and -2.9 [-5.1, -0.9]. On the secondary model, Llama-3.2-3B-Instruct at ratio 0.75, the normalized H2O against H2O as released is -1.0 [-3.8, $+1.7$] and the score against Expected Attention is $+0.9$ [-2.2, $+3.9$], the same two readings as on the model of record.

	NQA	Qasper	Hotpot	2Wiki	Gov	TREC	LCC	Avg
Full cache	22.4	42.5	41.7	45.6	34.4	16.0	24.5	32.4
H2O as released	23.0	39.4	38.3	42.7	32.6	2.0	21.1	28.5
H2O, normalized	24.2	31.7	33.9	45.6	31.3	5.3	22.8	27.8
SnapKV	18.0	20.0	37.6	37.1	28.0	9.2	22.5	24.6
Expected Attention	21.6	31.8	40.3	34.0	31.0	5.8	22.9	26.8
KeyDiff	20.7	33.6	39.1	36.5	27.6	50.0	20.8	32.6
KVzip	21.5	45.6	41.0	43.3	34.5	14.0	19.0	31.3
Mean-direction score	21.1	34.9	32.0	26.6	29.4	18.7	23.6	26.6

Table 12: LongBench on the instruct Llama-3.1-8B, 50 samples per subset, official metric $\times 100$, a quarter of the context cache kept (ratio 0.75, the registered primary). The full cache row is uncompressed.

Q PER-HEAD INFLATION FACTORS FROM PREDICTED AGREEMENT

Question and design. A violation is a failure of cross-query agreement by construction, since the dropped set is chosen on 32 queries and tested on the 33rd, so the paper's predictor should say which heads need the larger safety factor. Per unit (layer, KV head) a monotone map assigns an inflation factor $k_u = k_0(\bar{J}/\hat{J}_u)^\gamma$ snapped to the paper's existing grid, and the unit selects at E/k_u. k_0 and γ are fitted on the odd layers to reach a 5% violation rate there, and nothing is fitted on the even layers, where everything is measured.

Arms. Uniform is one factor for all units, as in Section 5. Predicted uses the two-moment model. Mass uses the head's mid-cache attention mass share, the scale-invariant scalar that absorbed the two-moment model in output space, and is the strongest simple baseline. Oracle uses each unit's own measured violation rate, which no deployable method has, and is an upper bound. Inverse flips the sign of γ and is an exploratory sanity arm. Uniform is nested inside every other arm at $\gamma = 0$, so the paired difference tests only whether the extra freedom generalizes off the calibration half.

Outcome. The gate is met on Llama-3.2-1B at both budgets and on neither budget for Qwen2.5-1.5B, so predicted agreement orders violation risk on one of the two models. The registered primary is not met on either. On Llama at $E = 10^{-2}$ the dropped fraction moves from 0.261 under a uniform factor to 0.264 under the predicted one at the same violation rate, with a paired per-unit median of +0.0000, against 0.260 for the mass arm and 0.282 for the oracle. The oracle shows headroom that the predictor does not reach. The inverse arm collapses onto uniform at $\gamma = 0$ on every cell, as expected when the grid carries no usable signal.

Relation to the allocation experiment. The allocation experiment of Appendix K moves budget between heads to minimize continuation perplexity at a fixed total budget. The calibration here fixes a per-query guarantee and asks what total cache it costs. The objective and the constraint differ, both results are negative, and Section 5 states the limit once.

R THE REGISTERED RECORD

In this paper, registered means that a prediction or decision rule was committed to the repository before the analysis it concerns. The commit history is the only record of that order, and it is not independently timestamped. The predictions and their amendments are kept in an append-only record released with the code. Every analysis whose registered reading was revised is retained there with the revision and its reason.

The question regime on further models. The question-regime statistics were first registered for Llama-3.2-1B, with the reading of each outcome. A second registration, made before any further capture, kept those statistics and fixed three further models of up to three billion parameters, a second template on Llama-3.2-1B and a fresh capture of the original configuration, with the prediction that the primary stays below the continuation value on every configuration. An amendment made after that registration and before the captures it names added the larger checkpoints of the paper's model set, Olmo-3-7B, Llama-3.1-8B, Qwen3.5-9B-Base and Olmo-3-32B, under the same statistics, prediction and permitted repairs. They were not part of the design first registered. The prediction held on 5 of 7 configurations and failed on Olmo-3-7B and Olmo-3-32B. The registered configurations of Llama-3.2-3B and Llama-3.2-1B under the second template completed after the analysis for this version was fixed, and their result files are in the released repository.

The 2025 to 2026 policies. Registered before compute with the statistics of Appendix N, the per-group and position sensitivity arms declared beside the released ones, and the prediction that predicted agreement orders regret for a method that reads every prefill row and for no method that reads a window or a single query summary. Predicted agreement orders regret for 6 of the 8 arms, several of which read a query summary, so the second half of the prediction did not hold. A rotary check on the capture was written with a float32 tolerance and relaxed to bfloat16 precision before any number of this run was computed.

Task accuracy on LongBench. Registered before compute and reported in Appendix P whatever it showed: the normalized H2O was predicted above H2O as released, and the mean-direction score not ahead of Expected Attention. The first leaves the LongBench average of H2O as released unchanged within its interval, and the second is level with Expected Attention within its interval.