RETENTION GEOMETRY: WHERE THE EVIDENCE SITS DECIDES HOW KV-CACHE COMPRESSORS RANK

Anonymous authors
Paper under double-blind review

ABSTRACT

KV-cache compression methods are compared by one accuracy number per benchmark. That number cannot identify how two methods rank once their position-resolved accuracies cross, and the deployed compressors make crossings common: at a matched budget, what a method keeps, and so whether the model answers, depends on where in the context the evidence sits. We formalise this with a method's evidence retention profile $\rho_m(p; B)$ and a benchmark's evidence-position prior D: a reported pairwise comparison has the sign of $\int \delta \, dD$ for the depth-resolved accuracy contrast δ, and window templates and accumulated attention have profiles that cross the random-eviction null's by construction. On a controlled grid (three models, 8K–32K, ten presses, six matched budgets, up to 21 depths paired within each item, two deployment protocols; 692,960 scored answers), the contrast between published methods changes sign with depth in 1–8 of 15 pairs in each harder-task cell under query-aware compression, certified by a bootstrap band corrected over depths and pairs; the count holds when the model's own greedy output is scored instead of teacher-forced tokens, though most reversals involve the first or last depth. A monotone link from ρ_m predicts a held-out method's position-resolved accuracy at R^2 0.47–0.77, against 0.41–0.58 for depth-averaged retention, an advantage carried by the window templates under query-aware compression and shared by most methods under query-agnostic compression. On natural documents scored by RULER's contains-match, 160 items certify 2 published pairs, only at the largest budget, and 7 under query-agnostic compression; RULER's own 996 QA rows, where depth is observed rather than paired, certify none, and re-weighting the grid by a task's measured prior does not forecast that task's ranking: the result is non-identifiability, not a forecast. An audit of 20,081 benchmark items puts the mean first-occurrence depth of 12 localisable tasks between 0.20 and 0.53, and finds one sub-task whose mean moves by 0.44 across RULER's releases. We recommend depth-resolved reporting and release a one-prefill retention probe. All code and every scored record are in the supplementary material and will be made public at `https://github.com/xxx/xxx` upon acceptance.

1 INTRODUCTION

Serving a language model over a long context is bounded by the key–value (KV) cache, and a large literature prunes that cache at prefill time: StreamingLLM keeps a few sink tokens and a recent window (Xiao et al., 2024), H_2O keeps the tokens that accumulated the most attention (Zhang et al., 2023), SnapKV keeps what the last few query tokens attend to (Li et al., 2024), PyramidKV varies the budget across layers (Cai et al., 2024), TOVA scores by the final query alone (Oren et al., 2024), and a norm rule keeps the keys of smallest L_2 norm (Devoto et al., 2024). These methods are compared the same way: an accuracy-versus-budget curve on LongBench (Bai et al., 2024), RULER (Hsieh et al., 2024a) or a needle-in-a-haystack (NIAH) suite (Kamradt, 2023), one number per method and budget. RULER's limitations section says that it "reports a single number metric for each input length without providing the depth-level performance"; SnapKV's evaluation ran NIAH over depth and length and then, "For simplicity, we aggregated the scores across all depths and lengths". The positional signal is measured and discarded before the ranking is read off.

Fig. 1 shows what is lost. At one budget and one model the accuracy of each compressor is a function of where the evidence sits, the functions have different shapes, and they cross: StreamingLLM is

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.

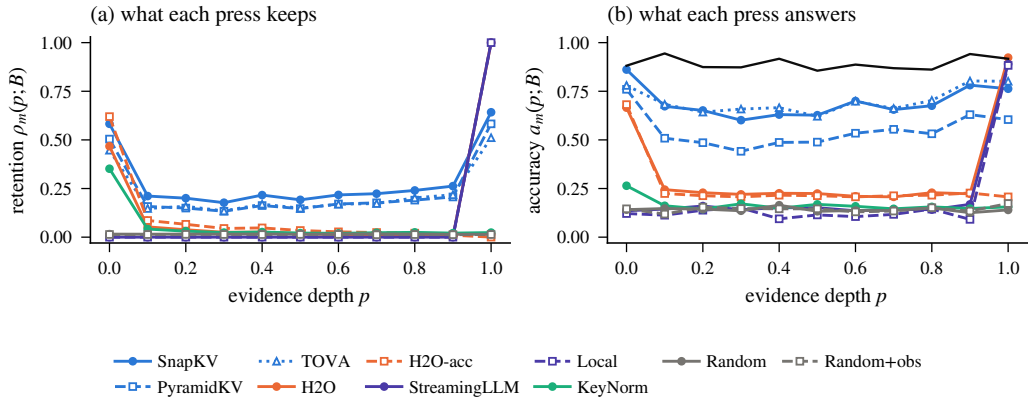

Figure 1: One budget, one model, ten compressors: Qwen2.5-7B at a 16,454-token context compressed to $B = 256$ tokens per layer (1.6% of the cache), query-aware protocol. (a) Evidence retention profile $\rho_m(p; B)$: the fraction of the evidence a method keeps when the evidence sits at depth p. (b) Position-resolved accuracy on the same items; the thin black line is the uncompressed model. RULER's own sub-task priors rank StreamingLLM 6th–7th of ten here and 5th–10th at $B = 512$ under $\mathrm{P}_{\mathrm{early}}$ (Fig. 2); a prior concentrated at the final tokens ranks it 2nd. Task H, teacher-forced token accuracy (Section 3).

second best when the evidence is in the final tokens and sixth to tenth everywhere else, level with the random and local presses; SnapKV is first or second everywhere else and fifth there. The scalar a benchmark reports averages such a curve under an evidence-position distribution that is neither controlled nor stated. Two benchmarks with different evidence placement can rank the same methods differently, and both rankings are then correct.

This paper treats that as a measurement problem. We model a compressor (a *press*) as a position-dependent retention operator and define two objects: the *evidence retention profile* $\rho_m(p; B)$, the fraction of evidence tokens a press keeps at budget B when the evidence lies at depth p, and the *evidence-position prior* D, the distribution of evidence depth induced by a benchmark. Three things follow. A reported comparison has the sign of $\int \delta \, dD$ for the depth-resolved contrast $\delta(p)$, which lies in $[\min_p \delta, \max_p \delta]$ on a depth grid, so a benchmark scalar is a functional of the pair (method, D), not of the method (Proposition 1). Under a two-regime answering model $\operatorname{sign} \delta(p) = \operatorname{sign}(\rho_{m_1}(p) - \rho_{m_2}(p))$ wherever the link separates the two retention values, so a per-depth ranking can reverse only where the profiles cross (Proposition 2), and the profiles cost one uncompressed prefill and no benchmark. And the standard presses have *structural* positional priors: accumulated-attention scores are early-biased under a uniform-attention null and their profiles measured so in every cell, window templates are steps with an interior hole at every budget, and random eviction is the position-unbiased null (Proposition 3), so a template's profile crosses the null's at every budget whose tail holds the evidence and an early-biased profile crosses a template's, and where the link separates the levels on both sides the ranking is depth-contingent (Corollary 2).

We test every step on a controlled multi-key retrieval grid whose one haystack carries evidence at 11 or 21 depths, so comparisons across depth are paired within an item: three models from two families, 8K to 32K, ten presses at six matched budgets, two protocols, two task difficulties, 692,960 scored answers, and the same design rebuilt from natural paragraphs. The accuracy contrast between published presses changes sign with depth in 1–8 of their 15 pairs in each harder-task cell under query-aware compression and a bootstrap band corrected over every depth and every pair, and about as often when the model's own output is scored without teacher forcing; a family of 18 priors, most of them point masses, produces up to 18 distinct rankings; re-weighting the same measurements by the empirical priors of RULER's own sub-tasks re-orders the ten presses, though under query-aware compression it rarely changes the leader. Under query-aware compression a single monotone link from ρ_m predicts the accuracy surface of a press the fit never saw at R^2 0.47–0.77 against 0.41–0.58 for depth-averaged retention and beats every depth-permuted refit. On RULER's own QA rows that re-weighting forecasts the measured ranking no better than the un-reweighted grid (Appendix X.3).

Where a press dominates at every depth the ranking is stable; Llama-3.1-8B on the easier task under query-aware compression comes closest, and exposes a mechanism the retention model does not cover, answer information in the retained query window.

Our contributions are a framework, *retention geometry*, with prior-invariant summaries (the worst-depth score W_m and lead fraction Φ) and structural positional priors for the standard press families (Section 2); a position-controlled measurement of it at scale, on synthetic grids, on natural documents and on a released benchmark's own rows (Sections 4 and 5); and a string-match audit of gold-evidence position in 15 benchmark tasks, with a reporting protocol and a one-prefill retention probe (Section 6). We do not propose a press: the claim is that the reported ordering is a property of the benchmark's evidence placement as well as of the presses, and can be reported without depending on it.

2 RETENTION GEOMETRY

Setup. An instance is a context of T tokens, a query and a gold answer; its *evidence set* $E \subseteq [T]$ is the set of context positions carrying the answer and its *relative depth* is $p = \min E/T$. A *press* m with budget B maps the prefilled cache of a pressed prefix of P tokens to a retained index set $S_m^{(\ell,h)} \subseteq [P]$ for every layer $\ell \in [L]$ and KV head $h \in [H]$, under the matched-budget constraint $\sum_\ell |S_m^{(\ell,h)}| = LB$ for every h. Two deployment protocols are usually conflated: under P_{late} the press sees the whole prompt including the query (the published setting of SnapKV, H_2O and PyramidKV); under $\mathrm{P}_{\text{early}}$ it sees the context alone, as in prefix caching.

Definition 1 (Retention profile, accuracy, prior). $\rho_m(p; B) = \mathbb{E}_{\text{depth}=p}\big[\frac{1}{LH}\sum_{\ell,h} |E \cap S_m^{(\ell,h)}|/|E|\big]$; $a_m(p; B) = \mathbb{E}[\text{score} \mid \text{depth} = p]$ under the task's metric; and $A_m(D) = \int_0^1 a_m(p; B)\, dD(p)$ where the *evidence-position prior* D is the distribution of evidence depth a benchmark induces. ρ_{cue} and ρ_{ans} are the same quantity on the cue tokens of the evidence (what the query matches) and its answer tokens (what the model must emit).

Given E, the profile is computable from one uncompressed prefill and the press's scoring rule, with no decoding and no accuracy; on a probe corpus E is known by construction, on a benchmark it must be annotated (Section 6). Every published number on a task with localisable evidence is an estimate of $A_m(D)$ for a D the paper does not state.

2.1 A REPORTED RANKING IS A FUNCTIONAL OF THE PRIOR

Let $\delta(p) = a_{m_1}(p; B) - a_{m_2}(p; B)$ and let $\mathcal{D}$ be the probability measures on $[0, 1]$ or a finite grid $\mathcal{P}$.

Proposition 1. $\operatorname{sign}\big(A_{m_1}(D) - A_{m_2}(D)\big) = \operatorname{sign}\int \delta\, dD$, *and* $\{\int \delta\, dD : D \in \mathcal{D}\}$ *is an interval with closure* $[\inf_p \delta,\ \sup_p \delta]$, *equal to it on a finite grid. Hence (i) if* $\delta \geq 0$ *everywhere (*depth-robust dominance*) no D ranks m_2 strictly above m_1; (ii) if δ changes sign, there are priors, among them point masses and finite grids concentrated where δ keeps one sign, that order the pair either way.*

The proof (Appendix A) evaluates the affine map $D \mapsto \int \delta\, dD$ at point masses. We call $[\inf_p \delta, \sup_p \delta]$ the *ranking-instability interval* $\mathrm{RII}(m_1, m_2; B)$; a comparison is *depth-contingent* when 0 is interior to it and *depth-robust* otherwise.

Corollary 1. *(Non-identifiability) A scalar per method on an unstated depth grid is a functional of the pair (method, D), and for a pair with* 0 *interior to its* RII *it does not determine the ordering. (Prior-invariant summaries) A summary of $a_m(\cdot)$ on a stated grid $\mathcal{P}$ that does not weight depths, such as $W_m = \min_p a_m(p)$ or $\Phi(m_1, m_2) = |\{p : \delta(p) > 0\}|/|\mathcal{P}|$, is unchanged by re-weighting.*

2.2 RETENTION MEDIATES POSITION-RESOLVED ACCURACY

Assumption 1. (Two regimes) The expected score is $\alpha(p)$ when the evidence is available in the cache and $\beta < \alpha(p)$ otherwise, with α and β shared by all presses and budgets of a cell. (Monotone availability) Availability is $g(\rho)$ for a non-decreasing $g : [0, 1] \to [0, 1]$, also shared. Section 4.4 measures the cells in which the shared-β half fails.

Proposition 2 (Retention mediation). *Under Assumption 1, $a_m(p; B) = \beta + (\alpha(p) - \beta)\, g(\rho_m(p; B))$ and $\delta(p) = (\alpha(p) - \beta)[g(\rho_{m_1}(p)) - g(\rho_{m_2}(p))]$, so $\operatorname{sign}\delta(p) = \operatorname{sign}(\rho_{m_1}(p; B) - \rho_{m_2}(p; B))$ wherever $g(\rho_{m_1}(p; B)) \neq g(\rho_{m_2}(p; B))$, and $\delta(p) = 0$ otherwise.*

Both halves are tested in Section 4. Two consequences: the per-depth contrast changes sign only where the profiles cross, which one prefill locates within a grid step in four of nine cells (Appendix H); and at each depth, ranking presses by ρ ranks them by accuracy up to ties where g is flat.

2.3 THE STANDARD PRESSES HAVE STRUCTURAL POSITIONAL PRIORS

Let A be the causal, row-stochastic prompt attention of one (layer, head) and $c_i = \sum_{q \geq i} A_{q,i}$ the accumulated score that H_2O ranks by; (b) and (d) are written for a pressed prefix of $P = T$ tokens (P_{early}); under P_{late} the last $P - T$ positions of the tail are the query, so (b)'s boundary becomes $p \geq (P - B + s)/T$ and (d) reads B/P.

Proposition 3. (a) *Under the uniform-attention null $A_{q,i} = 1/(q+1)$ for $i \leq q$, $c_i = H_T - H_i \approx \ln(T/i)$, strictly decreasing in i and independent of content; if rows are exchangeable over their causal support, $\mathbb{E}[c_i] = H_T - H_i$ as well.* (b) *For a template that keeps the first s and last $B - s$ positions ($s < B < T$), a contiguous evidence span of $|E| \leq B - s$ tokens not straddling the tail boundary has $\rho(p) = |E \cap [0, s)|/|E| + \mathbf{1}[p \geq 1 - (B - s)/T]$: 1 on the tail, 0 on the open middle, which holds a whole span whenever $B \leq T - |E|$, and on the head the fraction of the evidence inside the s sink tokens, which is 0 for evidence starting at or after position s. An evidence set straddling the tail boundary takes the intermediate value $|E \cap [T - (B - s), T)|/|E|$.* (c) *Query-anchored presses (SnapKV, PyramidKV, TOVA) are content-driven; SnapKV and PyramidKV additionally keep the last w positions.* (d) *Random eviction retains each position with probability B/T, so $\rho_{\text{rand}} \equiv B/T$; any position-unbiased press has this marginal, which makes random the canonical per-depth null.*

Part (a) is the uniform-attention form of the "query-domain mismatch" Zhang et al. (2026c) describe: early positions are inflated because more queries have seen them. Our measurement finds H_2O-acc's profile ρ_{acc} falling with depth in every cell, and c_i in every layer of the cell probed, but rejects the $\ln(T/i)$ form, so real attention is not exchangeable and (a) is used only as a monotonicity statement (Section 4.1 and Appendix X.1).

Corollary 2 (Crossings are structural)**.** *For every $s < B < T$ with $|E| \leq B - s$ (under P_{late}, $|E| \leq B - s - (P - T)$), (b) is below (d)'s B/T at some depth off the tail and $1 > B/T$ on the tail, so template and random profiles cross; the accumulated press's profile under the null of (a), which keeps exactly the first B positions, crosses (b) too. By Proposition 2 the accuracy contrast changes sign at a crossing wherever g separates the retention levels on both sides of it, and by Proposition 1 the ranking is then depth-contingent; no budget in that range removes the retention crossing, and g decides whether it becomes an accuracy reversal (Section 4).*

Two further results are deferred: the iso-accuracy sets of the surface $a_m(p; B)$, which for a template press are bounded by the retained tail, so budget buys nothing at depths below it (Proposition 4); and whether compressing a length-T context behaves like reading a shorter one, which (b)'s hole rules out for a template press once the uncompressed profile is measured flat (Appendix J).

3 EXPERIMENTAL DESIGN

A within-item paired design over depth. Each item is one haystack that carries K needles at controlled relative depths $p_j = j/(K-1)$, with $K = 11$, and $K = 21$ in cell H4. The haystack is RULER's multi-key construction (Hsieh et al., 2024a), real-word needle keys among pseudo-word filler keys in one shared sentence template (Appendix C). The K queries share one prefilled context, so every comparison across depth is paired within an item and every press starts from the same uncompressed cache (under P_{late} the keep set is recomputed per query). Two answer formats are used: task E, a 5-digit value, and task H, a value `dddd-dddd-dddd`. Task H is multi-token in both tokenizer families (14 Qwen tokens, 8 Llama) and leaves strong retrievers headroom on Qwen2.5-7B; its two hyphens are 0.14 of the graded metric for Qwen, which reproduces them without the evidence and so floors the metric there, and 0.25 for Llama, which does not (Appendix C).

Models, presses, budgets, protocols. Qwen2.5-7B/14B-Instruct (Qwen Team, 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), bfloat16, at 8K, 16K and 32K tokens (Table 1); Table 4 lists every controlled-grid cell and which counts are taken over which. Ten presses at *exactly* equal total budget: the six of Section 1 (StreamingLLM with four sink tokens, H_2O half recent and half

Table 1: The nine cells; each item is K depths $\times$ (10 presses $\times$ 6 budgets + uncompressed) $\times$ 2 protocols scored answers. "Sign": press pairs of 45 whose contrast changes sign with depth under pointwise intervals, in parentheses the same among the 15 pairs of the six methods of Section 1; "rankings": distinct full rankings under $\mathcal{D}_K$, of $K+7$; $\tau_{\min}$: smallest Kendall τ between two of those rankings; R^2: leave-one-press-out for $\rho(p,B)/\bar{\rho}(B)$/budget. Ranges are over budgets under P_{late}; $\mathrm{P}_{\text{early}}$ and the certified counts are in Table 2. T: median context length (spread ≤ 158).

cell	model	T	K	items	records	sign	rankings	$\tau_{\min}$	R^2 $\rho/\bar{\rho}/B$
H1	Qwen2.5-7B	16,454	11	50	67,100	12–24 (3–8)	16–18	−0.24	0.77/0.58/0.05
H2	Qwen2.5-14B	16,452	11	40	53,680	3–10 (1–7)	10–14	0.53	0.62/0.55/0.01
H3	Llama-3.1-8B	16,371	11	50	67,100	2–6 (1–5)	8–16	0.53	0.47/0.41/0.03
H4	Qwen2.5-7B	32,891	21	40	102,480	17–25 (7–11)	19–24	−0.07	0.75/0.55/0.04
E1	Qwen2.5-7B	8,158	11	60	80,520	4–20 (2–7)	16–18	0.00	0.59/0.41/0.03
E2	Qwen2.5-7B	16,368	11	100	134,200	12–20 (2–7)	16–18	−0.09	0.64/0.46/0.02
E3	Qwen2.5-7B	32,717	11	40	53,680	6–20 (0–4)	14–18	−0.16	0.69/0.48/0.02
E4	Llama-3.1-8B	16,332	11	60	80,520	1–4 (1–2)	5–10	0.42	0.01/−0.08/0.00
E5	Llama-3.1-8B	32,680	11	40	53,680	0–3 (0–2)	4–14	0.40	0.05/−0.05/0.00

accumulated, SnapKV with $w = 32$ observation queries and max-pooled scores, PyramidKV on SnapKV scores with an arithmetic per-layer schedule of mean B, TOVA, KeyNorm), plus LOCAL (the last B), H_2O-acc (all accumulated attention; not a deployed method, it isolates the prior of Proposition 3a), RANDOM (the null of Proposition 3d) and RANDOM+OBS (random plus SnapKV's window). Appendix W adds Expected Attention (Devoto et al., 2025) and query-protected variants. Six budgets per context, $B = 64$ to 2048 at 16K (0.4–12.6% of the cache across the primary cells), scaled with T. The retained cache is materialised by gathering B_ℓ entries per layer and KV head ($\sum_\ell B_\ell = LB$), not masked, so the budget is bytes held, the accounting Shi & Gu (2026) find published comparisons to violate; each follows its paper's rule, applied once at prefill and with the deviations Appendix B lists, on one engine verified by a 73-check suite.

Metric, retention, statistics. The primary metric is token accuracy under teacher forcing, the fraction of answer tokens whose argmax equals the gold token given the gold prefix: it is 1 exactly when greedy decoding reproduces the gold token sequence and grades partial answers below that, whereas exact match is near floor on task H for Qwen2.5-7B even uncompressed (Table 9). The 692,960 records of the nine primary cells are scored this way; Section 5 scores generated text. ρ_m, ρ_{cue} and ρ_{ans} are computed for every (press, budget, depth, protocol) on the same items, per layer and KV head. Every contrast is computed on the same items and carries a 95% item-level paired bootstrap interval (2000 resamples); a pair is *sign-changing* if its intervals exclude zero with opposite signs at two depths, and *certified* if a max-t band over all its depths, Bonferroni-corrected over the 45 pairs, does (Appendix D). The prior family $\mathcal{D}_K$ has $K + 7$ members (Appendix D); rankings are compared by Kendall's τ. The link g is fitted by isotonic regression (Barlow et al., 1972) on nine presses and evaluated on the tenth by R^2 against the training mean, against the grid-averaged $\bar{\rho}(B)$ and against the pooled budget curve alone. We locate the gold evidence in 20,081 items of 15 tasks by string match, giving an empirical prior per task (Appendix L).

4 RESULTS

Cells H1–H4 (task H) are the primary evidence; E1–E5 (task E) replicate the measurement on the easier format and contain the nearest case to the regime of Proposition 1(i). Numbers are under P_{late} unless stated and ranges are over a cell's six budgets. Table 3 indexes every control, including the 18 of 48 with negative results, which scope Section 8.

4.1 RETENTION PROFILES ARE STRUCTURAL

Fig. 1a and Fig. 4 show $\rho_m(p; B)$ for every press. The template presses are the step functions of Proposition 3(b): ρ is exactly 0 or 1 in every (budget, depth) cell but the one needle straddling the smallest window's boundary under P_{late} (0.26 for StreamingLLM at $B = 64$). Accumulated attention

is early-biased everywhere: H_2O-acc's profile falls with depth (Spearman -1.00 to -0.98) and so does c_i in every layer (Fig. 7), as Proposition 3(a) predicts, though not in closed form (Appendix X.1).

4.2 RANKINGS ARE DEPTH-CONTINGENT

Fig. 1b is the accuracy counterpart of Fig. 1a and has broadly the same shape: the RII of (SnapKV, StreamingLLM) there is $[-0.13, +0.73]$ with $\Phi = 0.91$. Over all budgets, 12–24 of the 45 pairs are sign-changing in H1 and 2–25 across H1–H4, with 1–11 among the 15 published pairs; the 45-pair counts rise under P_{early} in every cell, the published-pair counts in all but H4 (Tables 1 and 2). A bootstrap band over a pair's whole depth curve, Bonferroni-corrected over all 45 pairs, still certifies 7–16 in H1 and 1–8 published pairs across H1–H4, though the upper ends are maxima over budgets (Appendix D); at half the band's level every per-cell maximum is unchanged (Appendix X.4). Most of these reversals do not need teacher forcing: re-run on a separate machine with the model's greedy output scored by its digit edit distance to the gold value, H1–H4 certify 98 published-pair reversals under P_{late} and 135 under P_{early}, summed over budgets, against 96 and 122 teacher-forced on the same records, 78 and 97 of them the same (Appendix W.4). Keeping the question whole in the four presses that lose the most of it under P_{late} moves no published count by more than 1 at any budget, and re-run on a third machine, H1, H3 and H4 re-certify at least 95% of each run's certified pairs (94% of published ones), and 91% on the free-decoding machine, where many individual records differ (Appendices W.1 and W.5). Under P_{late} the instability concentrates at the boundary a template press creates, as Corollary 2 predicts (H_2O counts as one: half its budget is pure recency, and its retention at $p = 1$ is 1.00 in H1 at $B = 256$). So counted, a published pair involving a template is certified in 26.1% of (pair, budget, cell) observations against 1.5% for a pair of content-driven presses, an enrichment of 16.9× (exact 95% interval 7.1–53; 1.2× under P_{early}; Appendix X.8). It is not confined there: 2 content pairs are certified under P_{late} in 3 of the nine cells and 6 under P_{early}. The two groups reverse in different places, as Proposition 3 predicts: withholding both grid endpoints still leaves 20 template certifications over 3 pairs (most, 96 of 127, use $p = 1$), while no content pair survives withholding $p = 0$ alone. Nor does it need the window's leak: on a cell run with both arms, replacing StreamingLLM by its leak-free recomputation leaves all 8 tested (published pair, budget) observations certified and adds 2 (Appendix X.8).

Those counts measure the estimator as much as the geometry. Over all 12 cells, how many pairs a cell *certifies* tracks answer length (Spearman 0.93) more than full-KV exact match (-0.32), while how many are *depth-contingent* tracks saturation (-0.57), the direction Proposition 1(i) predicts (Appendix X.10). A certificate is evidence about sample size; a RII read off point estimates describes the geometry, and fires on 25% of depth-permuted contrasts (Appendix X.4). Exact match certifies at most 4 of 15 published pairs in any cell, even in one built at 0.53 full-KV exact match (Appendix P).

Fig. 10 applies Proposition 1 directly. Under the 18 priors of $\mathcal{D}_{11}$ the ten presses receive 16–18 distinct full rankings in H1, minimum Kendall τ -0.24. Per-cell counts across H1–H4 run from 8 of 18 (H3) to 24 of the 28 priors of $\mathcal{D}_{21}$ in H4, at $\tau_{\min} \geq -0.24$ (Table 2), above a matched item-noise null in 139 of 144 (cell, protocol, budget) combinations (Appendix X.2). Most of $\mathcal{D}_K$ is point masses; over the 7 members that are not, the count is still 3–7, above a matched item-noise null in 73 of 108 combinations (107 of 108 for the six published presses under $\mathcal{D}_K$; Appendix X.2), and the first-ranked press changes in 9 of 9 cells, in E4 and E5 only among presses tied at ceiling (Appendix X.7). In H1 at $B = 256$ a query-anchored press leads under every prior but the point mass at $p = 1$, where the tail-keeping presses take the top three; against the random null the template presses gain 0.75 there and nothing at $p = 0.5$.

4.3 RETENTION PREDICTS ACCURACY, AND THE LINK TRANSFERS

Proposition 2 says one monotone function of ρ should reproduce the surface when α is flat, as it is here (Appendix J). We fit g on nine presses and predict the $6 \times K$ (budget, depth) cells of the tenth (Fig. 11); averaged over held-out presses, R^2 is 0.47–0.77 across H1–H4, against 0.41–0.58 for the depth-averaged $\bar{\rho}(B)$, the statistic a position-agnostic account would use, and 0.01–0.05 for the budget alone. With the held-out press as the unit, the P_{late} gap is carried by the window templates: ρ is ahead for 4–6 of the ten presses and a press-level interval includes zero in every cell, whereas under P_{early} it is ahead for 8–9 of ten in every cell (Appendix X.6); depth resolution loses in 16 of 48 combinations (Appendix X.1). Holding out a press's whole family lowers R^2 by at most 0.04;

the link's depth shape added to a held-out press's own budget curve beats that curve in 4 of 4 cells; and, fitted on the ten, the link predicts Expected Attention, a press from none of their families, at R^2 0.79–0.88 under P_{early} and 0.48–0.75 under P_{late}, where $\bar{\rho}$ does better in 2 of 3 cells (Appendices W.2 and W.3). Under P_{early}, ρ gives 0.25–0.84 against 0.25–0.38 for $\bar{\rho}$, which wins in H3 and loses by 0.15–0.58 elsewhere. A depth-permutation null, which keeps every marginal and $\bar{\rho}$ and destroys only the alignment with depth, drops the fit to 0.04–0.31 across H1–H4, and the observed value beats all 200 permutations in each of the 16 cells of Appendix D and both protocols. A third family, Mistral-7B-Instruct-v0.3 on H1's task at 8,187 tokens (cell H5), reproduces the ranking claims, and links fitted in one of five cells predict every other with no refitting at positive R^2 (in-sample fits; Table 7, Appendix S). An item-level bootstrap puts the gap at 0.19 [0.18, 0.20] in H1 and above zero in all 4 of 4 cells, with the presses held fixed (Appendix D). Without a link, Proposition 2 predicts the sign of every contrast from the retention difference, and at the 5,726 significant (published pair, budget, depth) points of H1–H4 the two agree in 88.7% (Appendix H; per-depth orderings in Table 2).

4.4 Where the framework predicts stability, and what retention alone cannot explain

Luo et al. (2026) show on RULER that query-agnostic compression re-orders the published presses; the grid reproduces the re-ordering and locates it in ρ_m: the protocol moves the profiles, Proposition 1 applies within each, and protocol and prior are two coordinates of one object (Appendix T).

Proposition 1(i) says a press that dominates at every depth is ranked first under every prior, and Llama-3.1-8B on task E comes closest (Fig. 6): SnapKV and PyramidKV sit at ceiling at every depth, E4–E5 show 0–2 sign-changing published pairs (Table 2), and first place moves only among presses tied at ceiling (Appendix X.7). The mediation fit is poor there (R^2_ρ 0.01–0.05) because the shared-β half of Assumption 1 fails: at zero retention StreamingLLM still scores 0.45 of the two answer tokens against 0.00 for LOCAL, because the retained query window was prefilled with the evidence in view and the sink tokens gate its readout. Varying only the sink size makes the floor a gate: 0.000 wherever either block is missing, 0.26–0.48 with both (Appendix M). An intervention tests the mechanism. Recomputing the retained tail against the already-pruned cache, so it cannot have attended to anything the press evicted, leaves prompt, keep set, positions and budget identical: the floor falls from 0.43–0.45 to 0.03–0.08 over 4 templates, while where the evidence is retained the same recomputation costs 0.000 (Appendix Q). Task H lowers the floor and shrinks the gap without closing it (H3: 0.21 against 0.03), and R^2_ρ is 0.47. Depth-contingency belongs to a (press pair, model, task, protocol, budget), and the certified signs of the RII's endpoints say which regime holds.

5 Beyond the controlled grid

To test whether the account survives real text we rebuild the same design from SQuAD paragraphs (Rajpurkar et al., 2016): 16,335 tokens (median), $K = 11$ gold paragraphs with their own questions and answer spans, placed on an even depth grid and sharing one prefill so depths stay paired, scored by RULER's own contains-match on 16 greedily decoded tokens (Appendix K). Qwen2.5-7B, 160 items, ten presses at 6 matched budgets, two protocols, 214,720 scored generations.

Rebuilt from that data, Fig. 1 has the same two shapes (Fig. 14): StreamingLLM retains ρ 0.00 at interior depths and 0.95 at the tail, and answers 0.16 against 0.70 at $p = 0.9$, with the uncompressed model at 0.73. The central claim survives the benchmark's own metric under the same band: at $B = 2048$ it certifies 2 of the 15 published pairs under P_{late} and 7 under P_{early}, 3 of the latter between two content-driven presses, and none at any smaller budget under P_{late} (Table 11); 160 further items run on another machine certify the same 2 pairs at $B = 2048$, none below, and 6 under P_{early} (Appendix W.5). Mediation carries over under contains-match, at R^2 0.58 against 0.46 for $\bar{\rho}$ under P_{late}, though H1's link does not transfer here (Appendix K).

An answer that cannot be copied. Every cell so far is extractive, confounding retention with copyability; cell S separates them. The needle stores two four-digit addends and the question asks for their sum, so the answer occurs nowhere in the context (40 items; Appendix R). The framework carries over: 52% of (published pair, budget) observations are depth-contingent under P_{late} against 17–67% on the extractive cells, 0–2 of 15 pairs are certified, and ρ still beats $\bar{\rho}$ (0.43 against 0.31).

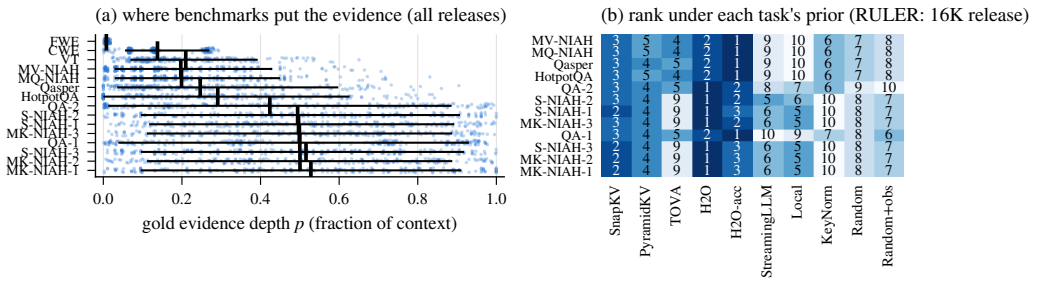

Figure 2: (a) Gold evidence depth of 20,081 items in 15 benchmark tasks (every RULER row at 4K, 8K and 16K, pooled; up to 220 points drawn per task); ticks mark means, bars the 10th–90th percentiles; both panels are ordered by the 16K mean. (b) Rank of each press in H1 re-weighted by the prior of each of the 12 localisable tasks ($B = 512$, $\mathrm{P}_{\text{early}}$): the winner splits between H2O-acc and H2O, and mean depth does not separate the two groups.

The benchmark itself. We also ran the ten presses on RULER's own `qa_1` and `qa_2` rows unchanged, depth observed rather than controlled: every located row at 16K, 996 of 1,000, at three budgets, 61,752 scored generations (Appendix K). In aggregate the presses separate easily (13–14 of the 15 published pairs differ under P_{late}), but no sign change is certified at any budget and at most 1 is pointwise, for three reasons the design fixes and this benchmark cannot: depths are unpaired groups of 70–341 rows, the uncompressed model moves from 0.24 to 0.61 across them, and few rows place their evidence inside a window's retained tail (38 of 996 at the largest budget, 25 at $B = 512$). LongBench's own QA rows give the same answer under its own F1 (593 located rows of four subsets; Appendix K.1): no published pair is certified, while under $\mathrm{P}_{\text{early}}$ the subsets order the six published presses in 3–4 ways per budget, H2O and KeyNorm each first on some subset, and retention predicts F1 poorly on these mostly multi-hop rows. RULER's rows also test the forecast Section 6 declines to make, and do not support it: re-weighting H1 by a task's prior predicts that task's ranking not reliably better than not re-weighting it (τ changes by -0.18 under P_{late} and $+0.09$ under $\mathrm{P}_{\text{early}}$; Appendix X.3).

6 Where benchmarks put the evidence, and what to report

Proposition 1 matters only if real benchmarks differ in D. They do: for 15 tasks a string match located the answer in 99.9% of 20,081 items (Fig. 2a, Table 6). Mean evidence depth runs from 0.008 to 0.53: HotpotQA-distractor (Yang et al., 2018) and Qasper (Dasigi et al., 2021) sit at 0.29 and 0.25. The prior is not even a property of a sub-task: QA-1 sits at mean depth 0.72 in RULER's 4096-token release, 0.28 at 8192 and 0.50 at 16384 — a swing of 0.44 (0.43 over every answer occurrence, 0.55 at the gold passage; Appendix U), wider than the 0.20–0.53 spread of first-span means across the 12 localisable tasks. One number per method is not comparable across a benchmark's configurations. The re-weightings below use RULER's 16,384-token rows and the other tasks' native rows, with mean depth 0.19–0.51 over the 12 tasks (Appendix U).

Restricted to the 12 tasks that neither aggregate over the context nor follow a chain (Appendix U), Fig. 2b re-weights the measurements of H1 by those priors at $B = 512$ under $\mathrm{P}_{\text{early}}$, and the winner splits in two: H2O (7), H2O-acc (5), the second a diagnostic press, so among the published six $\mathrm{H_2O}$ wins under every prior there. Over those tasks there are 7 distinct orderings of the ten presses (5 under P_{late} at that budget, where SnapKV wins throughout; at $B = 256$ two published presses split the tasks, Table 10). The panel is Corollary 1 made concrete, not a forecast, and the reversals it implies are not confined to point masses: of the 219 certified published-pair observations of H1–H4, 92 are reversed by at least one audited task prior (14 of 95 under P_{late}; Appendix X.12).

What to report. For any evaluation with localisable evidence, alongside the mean report the position-resolved accuracy $a_m(p; B)$ on a stated grid with its item count, the worst-depth score W_m, Φ and RII with intervals for each pair of interest (and the certified sign changes, Appendix D), the protocol, and the per-depth random-at-budget null, which no KV-compression paper we know of

reports; for a benchmark, report $\hat{D}$ per configuration (Fig. 2a). One uncompressed prefill returns $\rho_m(\cdot; B)$ for any prefill-time press and budget, hence its retention blind spots and crossings (which locate accuracy crossings to within a grid step in four of nine cells, Appendix H), without decoding, on a corpus whose evidence positions are known; a diagnostic, not a substitute for the benchmark.

7 Related work

Dehghani et al. (2021) show that leaderboard orderings are artefacts of which tasks and instances a benchmark weights, and re-weighting reverses them. Proposition 1 is that argument on one axis, evidence depth, where the weighting is a measurable by-product of how the instances were built; Corollary 2 adds that on this axis some crossings between the standard press families are structural.

Four 2026 papers supply pieces of this. Qiu et al. (2026) and Khodabandehlou & Krishnamachari (2026) measure the fraction of the gold evidence a press retains as a scalar per instance; neither indexes it by position or compares presses as functions of it. Luo et al. (2026) report that query visibility re-orders the published presses, which Section 4.4 reproduces on controlled depths and reads as a change in ρ_m; KVzip (Kim et al., 2025) and SCBench (Li et al., 2025b) had already built for that query-agnostic setting. Shi & Gu (2026) show that optimal token selection closes only 2–5% of the gap to full attention and attribute published margins to memory accounting; our budgets are materialised (Section 3), and Proposition 3 names a second source. None has the position argument: ρ as a function of depth, the crossing it forces, and the prior that turns it into a ranking.

That models fail in the middle of long inputs is established (Liu et al., 2024) and now derived from causal masking (Chowdhury, 2026; Duerinckx et al., 2026); that is the model's profile, while Proposition 3 describes the operator a press applies to the cache, and on the controlled grid Qwen2.5-7B's own profile is flat (Appendix J). That accumulated-attention scores are biased and random eviction offsets it (Chen et al., 2024), and that eviction bias tracks structural role (Li et al., 2025a; Steele, 2026; Mandal, 2026), are reported; none measures a profile. Aggregates conceal non-uniform damage (Chen et al., 2026; Ai et al., 2026; Wu et al., 2026), with position uncontrolled. Appendix O adds the leak's mechanism (Fesser et al., 2026) and the alternative it rules out (Niu et al., 2025).

8 Limitations and conclusion

The effect needs resolution the field's metrics often lack: exact match certifies at most 4 of 15 published pairs in every cell, and on the released rows of RULER (996) and LongBench (593), each under its own metric, none is certified, against 2 (P_{late}) and 7 (P_{early}) on 160 natural-document items. Re-weighting a grid by a task's prior does not forecast that task's ranking (Appendix X.3); what stands is non-identifiability, mediation within a cell, and transfer across synthetic cells, not to natural text (Appendix K). The primary metric is teacher-forced, though the model's own greedy output certifies about as many published reversals, most of them the same (Appendix W.4); the shared floor of Assumption 1 fails where a retained window carries the answer (Section 4.4); accuracy levels need not transfer across hardware, while contrasts re-checked on two other machines transferred (Appendices N and W.5). Most certified reversals use the first or last grid depth, and under P_{late} interior ones appear only at the largest budget (Table 2 and Appendix X.8); the flat uncompressed profile behind Proposition 4 is measured for Qwen2.5-7B on task H only (Appendix J); the family-wise band is anti-conservative by about a factor of two in its least favourable configuration (Appendix X.4); and the equal per-head budget of Section 2 leaves head-adaptive presses out. 1 labelled item scored over the grid predicts better than the probe's link (Appendix X.11). The audit is a string match, which for a multi-answer task records the first span (Appendix U). A position-indifferent press is not the way out: a positional quota flattens ρ_m at a cost in accuracy, and pairs mixing it with a deployed press still reverse (Appendix X.9). Coverage stops at three model families, 32K tokens, 12.6% of the cache (25% in H5, 50% on the 4K rows) and prefill-time eviction, mostly on a synthetic task. Within those limits the reported order of KV-cache compressors is a functional of a benchmark's evidence placement and the task's position-resolved accuracy, which $a_m(p; B)$, W_m and Φ report.

ETHICS STATEMENT

This work evaluates inference-time cache compression on synthetic retrieval tasks, on items built from SQuAD v1.1 paragraphs and on the released QA rows of RULER and LongBench, and audits evidence position in RULER, HotpotQA and Qasper without running the presses on the original releases of the latter two; it involves no human subjects, collects no personal data and introduces no capability beyond what the evaluated methods already provide. Its findings bear on how methods are compared and could change which method a practitioner deploys; the one route to harm we see is the one the paper documents: a window template silently drops everything in the middle of a context, and the retention profile detects it before deployment. Models (Qwen2.5-7B/14B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3) and datasets (RULER, SQuAD, HotpotQA, Qasper, LongBench and the 2WikiMultihopQA and MuSiQue data it redistributes) are used under their published licences.

REPRODUCIBILITY STATEMENT

Every experimental number in the paper is generated from the released records by a script that writes the macros the manuscript uses (`emit_paper_numbers.py`), and every figure by a second script from the same files; the drivers that rebuild every analysis from the records run on CPU, and regenerating the records themselves needs a GPU (Appendix N). The supplementary material contains the engine, every press implementation (the ten of Section 3, the sink, leak-free, quota and question-protected variants and Expected Attention), the one-prefill probe, the haystack generator with seeds, the analysis code, the tool-verification suite, the benchmark-audit script and every scored record the paper uses as compressed JSON lines: the 692,960 answers of the nine primary cells, the 40,260 of the single-needle control, the 360,352 records of the natural-document and RULER-verbatim cells, the records of cells M, S and H5 and of every control in Table 3, and 20,081 audit rows; the manifest lists each file. Proofs are in Appendix A, press definitions and budget matching in Appendix B, task construction and prompts in Appendix C, statistical procedures in Appendix D, per-cell tables for both protocols in Appendix E, the audit procedure in Appendix L and its limits in Appendix U. Hardware, running times and what a replicator on other hardware should expect (contrasts, not levels) are in Appendix N; the model identifiers and the software environment are in the release's README, and each record's header carries its model, context length, seed and budget.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data (the synthetic corpora this paper evaluates on were produced by the released deterministic generator, not by a generative model), to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Nikita Agrawal and Ruben Mayer. Benchmarking KV-Cache Optimizations across Task Quality and System Performance for Long-Context Serving. *arXiv preprint arXiv:2607.05399*, 2026.

Mengting Ai, Jingrui He, and Yue Guo. Does Accuracy Equal Evidence? Reasoning Faithfulness under KV Cache Compression. *arXiv preprint arXiv:2608.01631*, 2026.

Samhruth Ananthanarayanan, Ayan Sengupta, and Tanmoy Chakraborty. Understanding the Physics of Key-Value Cache Compression for LLMs through Attention Dynamics. *arXiv preprint arXiv:2603.01426*, 2026.

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 *Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2308.14508.

Richard E. Barlow, David J. Bartholomew, J. M. Bremner, and H. Daniel Brunk. *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, 1972.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Alex Chen, Renato Geh, Aditya Grover, Guy Van den Broeck, and Daniel Israel. The Pitfalls of KV Cache Compression. In *Annual Meeting of the Association for Computational Linguistics*, 2026. arXiv:2510.00231.

Yilong Chen, Guoxia Wang, Junyuan Shang, Shiyao Cui, Zhenyu Zhang, Tingwen Liu, Shuohuan Wang, Yu Sun, Dianhai Yu, and Hua Wu. NACL: A general and effective KV cache eviction framework for LLMs at inference time. In *Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2408.03675.

Borun D Chowdhury. Lost in the Middle at Birth: An Exact Theory of Transformer Position Bias. *arXiv preprint arXiv:2603.10123*, 2026.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2021. arXiv:2105.03011.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The Benchmark Lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Alessio Devoto, Yu Zhao, Simone Scardapane, and Pasquale Minervini. A Simple and Effective L_2 Norm-Based Strategy for KV Cache Compression. In *Conference on Empirical Methods in Natural Language Processing*, 2024. arXiv:2406.11430.

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.

Mitia Duerinckx, Borjan Geshkovski, and Stefano Rossi. Kinetic theory for Transformers and the lost-in-the-middle phenomenon. *arXiv preprint arXiv:2605.09213*, 2026.

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*, 2025. arXiv:2407.11550.

Lukas Fesser, Mozes Jacobs, Thomas Fel, Andy Keller, and Sham Kakade. A Unifying View of Attention Sinks: Two Algorithms, Two Solutions. *arXiv preprint arXiv:2606.08105*, 2026.

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 *International Conference on Learning Representations*, 2024. arXiv:2310.01801.

Stephen Gould and Anton van den Hengel. Back from the Future: Key-Value Cache Management by Counter-Causal Surprise. *arXiv preprint arXiv:2607.27600*, 2026.

Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *Conference on Language Modeling (COLM)*, 2024a. arXiv:2404.06654.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the Middle: Calibrating Positional Attention Bias Improves Long Context Utilization. In *Findings of the Association for Computational Linguistics: ACL*, 2024b. arXiv:2406.16008.

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

Greg Kamradt. Needle in a haystack: Pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.

Mohammadali Khodabandehlou and Bhaskar Krishnamachari. Compression-Aware Abstention: Teaching LLMs to Refuse When KV-Compression Masks Remove Answer Evidence. *arXiv preprint arXiv:2608.29934*, 2026.

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*, 2025. arXiv:2505.23416.

Runchao Li, Yao Fu, Mu Sheng, Xianxuan Long, Haotian Yu, and Pan Li. FAEDKV: Infinite-Window Fourier Transform for Unbiased KV Cache Compression. *arXiv preprint arXiv:2507.20030*, 2025a.

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 *International Conference on Learning Representations*, 2025b. 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*, 2024. arXiv:2404.14469.

Minghui Liu, Aadi Palnitkar, Tahseen Rabbani, Hyunwoo Jae, Kyle Rui Sang, Dixi Yao, Shayan Shabihi, Fuheng Zhao, Tian Li, Ce Zhang, Furong Huang, and Kunpeng Zhang. Hold Onto That Thought: Assessing KV Cache Compression On Reasoning. *arXiv preprint arXiv:2512.12008*, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. arXiv:2307.03172.

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*, 2023. arXiv:2305.17118.

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.

Soumil Mandal. Adaptive Filtering of the KV Cache: Diagnosing and Correcting Structural-Role Bias in LLM Inference. *arXiv preprint arXiv:2607.13205*, 2026.

Piotr Nawrot, Robert Li, Renjie Huang, Sebastian Ruder, Kelly Marchisio, and Edoardo M. Ponti. The Sparse Frontier: Sparse Attention Trade-offs in Transformer LLMs. In *Findings of the Association for Computational Linguistics: ACL*, 2026. arXiv:2504.17768.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H. Abdi. Llama See, Llama Do: A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs. In *Annual Meeting of the Association for Computational Linguistics*, 2025. arXiv:2505.09338.

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

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are Multi-State RNNs. In *Conference on Empirical Methods in Natural Language Processing*, 2024. arXiv:2401.06104.

Chen Qiu, Ziwu Liu, Chao Fei, Guozhong Li, and Panos Kalnis. KVDiagnosis: A Diagnostic Benchmark for KV-Cache Compression in Long-Context Language Models. *arXiv preprint arXiv:2608.09412*, 2026.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Conference on Empirical Methods in Natural Language Processing*, 2016. arXiv:1606.05250.

Xingyu Ren, Youran Sun, Chugang Yi, and Haizhao Yang. Understanding Sparse Attention Selectivity in Long-Context Foundation Models via Counterfactual Evaluation. *arXiv preprint arXiv:2608.01676*, 2026.

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

Brady Steele. On the Limits of Learned Importance Scoring for KV Cache Compression. *arXiv preprint arXiv:2601.14279*, 2026.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive Activations in Large Language Models. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2402.17762.

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

Blerta Veseli, Julian Chibane, Mariya Toneva, and Alexander Koller. Positional Biases Shift as Inputs Approach Context Window Limits. *arXiv preprint arXiv:2508.07479*, 2025.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating Position Bias of Language Models: A Mechanistic Approach. In *International Conference on Learning Representations*, 2025. arXiv:2407.01100.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval Head Mechanistically Explains Long-Context Factuality. In *International Conference on Learning Representations*, 2025. arXiv:2404.15574.

Yuan Wu, Mairui Li, Lesia Semenova, and Chudi Zhong. The Asymmetric Harms of LLM Compression. *arXiv preprint arXiv:2608.19670*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. In *International Conference on Learning Representations*, 2024. 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 *International Conference on Learning Representations*, 2025. arXiv:2410.10819.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Conference on Empirical Methods in Natural Language Processing*, 2018. arXiv:1809.09600.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, Zirui Liu, and Xia Hu. KV Cache Compression, But What Must We Give in Return? A Comprehensive Benchmark of Long Context Capable Approaches. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. arXiv:2407.01527.

Chuyifei Zhang, Hongyu Cui, Xiaowen Huang, and Jitao Sang. Positional Failures in Long-Context LLMs: A Blind Spot in Reasoning Benchmarks. *arXiv preprint arXiv:2605.23170*, 2026a.

Meiru Zhang, Zaiqiao Meng, and Nigel Collier. Failure Modes in Multi-Hop QA: The Weakest Link Effect and the Recognition Bottleneck. In *Annual Meeting of the Association for Computational Linguistics*, 2026b. arXiv:2601.12499.

Ruijie Zhang, Haozhe Liang, Da Chang, Li Hu, Fanqi Kong, Huaxiao Yin, and Yu Li. When Does Value-Aware KV Eviction Help? A Fixed-Contract Diagnostic for Non-Monotone Cache Compression. *arXiv preprint arXiv:2605.08234*, 2026c.

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. H_2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2306.14048.

A PROOFS

Proof of Proposition 1. $A_{m_1}(D) - A_{m_2}(D) = \int (a_{m_1} - a_{m_2})\, dD = \int \delta\, dD$ by linearity of the integral, which gives the first statement. The map $\Lambda : D \mapsto \int \delta\, dD$ is affine on $\mathcal{D}$, which is convex, and $\mathcal{D}$ is the closed convex hull of its extreme points, the point masses (Dirac measures) $\{\varepsilon_p : p \in [0,1]\}$ (on a finite grid, $\mathcal{D} = \Delta^{|\mathcal{P}|-1}$ and the extreme points are its vertices). For every probability measure $\int \delta\, dD \le \sup_p \delta$, and $\Lambda(\varepsilon_p) = \delta(p)$, so $\sup_D \Lambda(D) = \sup_p \delta(p)$; likewise $\inf_D \Lambda(D) = \inf_p \delta(p)$. Since Λ is affine and $\mathcal{D}$ is convex, $\Lambda(\mathcal{D})$ is convex, hence an interval, and it contains an endpoint exactly when that endpoint is attained by some $\delta(p)$ (if $\delta < \sup_p \delta$ everywhere then $\int (\sup_p \delta - \delta)\, dD > 0$ for every D, so the supremum is not attained; it is attained always, on a finite grid; for a continuous δ on $[0,1]$ by compactness; not, for example, for a δ whose supremum is approached but never reached). This gives the range. Statement (i): if $\delta \ge 0$ then $\Lambda(D) \ge 0$ for every D, so no D ranks m_2 strictly above m_1, and the ranking is strict whenever D puts mass where $\delta > 0$. Statement (ii): if $\delta(p_1) > 0 > \delta(p_2)$ then the point masses ε_{p_1} and ε_{p_2} order the pair oppositely, and so does any finite grid with uniform weights all of whose points lie in $\{\delta > 0\}$, respectively $\{\delta < 0\}$; such grids exist near p_1 and p_2 whenever δ keeps its sign on a neighbourhood of each, which on a finite depth grid is the point itself. Evenly spaced grids over all of $[0,1]$ need not reverse a given pair; the statement is about the existence of grids, not about every grid. □

Proof of Corollary 1. The first part restates the range: two methods with $0 \in \operatorname{int} \mathrm{RII}$ are ordered either way by some D, so a scalar $A_m(D)$ with D unstated does not determine the order; for a pair with $\delta \ge 0$ everywhere it does, which is statement (i). For the second part, $\min_p a_m(p)$ and $|\{p : \delta(p) > 0\}|/|\mathcal{P}|$ are functions of the profile on the grid alone and involve no weighting over depths, so re-weighting depths leaves them unchanged; both do depend on the grid $\mathcal{P}$, which is why the grid must be stated. □

Proof of Proposition 2. By the two-regime part of Assumption 1, conditioning on availability, $a_m(p; B) = \Pr[\text{avail}]\, \alpha(p) + (1 - \Pr[\text{avail}])\, \beta$, and by the monotone-availability part $\Pr[\text{avail}] = g(\rho_m(p; B))$, which gives the first display. Subtracting the expressions for m_1 and m_2 gives the second. Since $\alpha(p) - \beta > 0$ and g is non-decreasing, $g(\rho_{m_1}) - g(\rho_{m_2})$ is zero when g is constant on the closed interval between the two retention values and otherwise has the sign of $\rho_{m_1} - \rho_{m_2}$, which is the statement; in particular it has that sign whenever g is strictly increasing between them. With a graded metric the two regimes are expected scores, and the same conditioning applies to them by linearity. (A g that is flat on part of the interval and increasing elsewhere still gives a non-zero difference of the stated sign; the fitted isotonic link of Section 4.3 is piecewise constant, so the flat case is not hypothetical.) □

Proof of Proposition 3. (a) Under the uniform null, $c_i = \sum_{q=i}^{T-1} \frac{1}{q+1} = H_T - H_i$ where H_n is the nth harmonic number; $H_T - H_i$ is strictly decreasing in i and $H_T - H_i = \ln(T/i) + O(1/i)$. For the general statement, $\mathbb{E}[c_i] = \sum_{q \geq i} \mathbb{E}[A_{q,i}]$; if for each query q the entries $(A_{q,j})_{j \leq q}$ are exchangeable then $\mathbb{E}[A_{q,j}] = 1/(q+1)$ for all $j \leq q$, so $\mathbb{E}[c_i] = H_T - H_i$ exactly, the uniform-null value; a measured c_i that departs from that shape therefore also rejects exchangeability, while any row law whose expected weights are non-increasing in position within each row still gives a non-increasing $\mathbb{E}[c_i]$. (b) The keep set is $\{0, \ldots, s-1\} \cup \{T-(B-s), \ldots, T-1\}$ for every layer and head, so the retention of an evidence set E is $|E \cap [0, s)|/|E| + |E \cap [T-(B-s), T)|/|E|$; a needle whose tokens lie entirely in $[s, T-(B-s))$ has retention 0, one that lies entirely in the tail has retention 1, and one at the head retains only the part inside the s sink tokens, which is nothing if it starts at or after position s; the middle set is non-empty whenever $B < T$, and holds a whole needle of $|E|$ tokens whenever $B \leq T - |E|$. (c) is the definition of the presses. (d) A uniform random B-subset of $[T]$ contains any fixed position with probability B/T, so the expected retention of any evidence set is B/T regardless of position. Conversely, a press whose expected retention of a single token does not depend on the token's position has the same marginal inclusion probability at every position, and since the marginals sum to B that probability is B/T; random eviction achieves this with no dependence on content (as would any other position-unbiased content-blind rule; only the marginal is pinned down), which is what makes it the null against which a position-specific claim is tested. □

Proof of Corollary 2. Fix $s < B < T$ and a span of $|E| \leq B - s$ tokens. By (d), $\rho_{\text{rand}} \equiv B/T \in (0, 1)$. By (b), a span starting at s has $\rho_{\text{tmpl}} = 0$ when it fits in the middle ($B \leq T - |E|$), and otherwise covers the middle and has $\rho_{\text{tmpl}} = 1 - (T-B)/|E| < B/T$ (since $|E| < T$); either way $\rho_{\text{tmpl}} < \rho_{\text{rand}}$ there, while on the tail $\rho_{\text{tmpl}} = 1 > B/T$, so the two profiles cross. The two conditions give (b) its clean two-level form and are sufficient, not necessary: with $s = B$ the template keeps the first B positions and crosses the null at the head boundary instead, and a span longer than $B - s$ that ends at T keeps $(B-s)/|E|$ of itself, which still exceeds B/T whenever $|E| < T(B-s)/B$; under P_{late} the tail must also hold the $P - T$ query tokens, so the condition is $|E| \leq B - s - (P - T)$. For the accumulated press, (a) orders the expected scores, not the keep set, so its profile is non-increasing in p only under the deterministic null of (a), where the press keeps exactly the first B positions; then a span starting at s and ending before B has $\rho_{\text{acc}}(s/T) = 1 > \rho_{\text{tmpl}}(s/T)$, while at the last depth $\rho_{\text{acc}} \leq \rho_{\text{acc}}(s/T)$ and $\rho_{\text{tmpl}} = 1$, so the two profiles cross unless $\rho_{\text{acc}} \equiv 1$, which the condition $B < T$ excludes for a span in the tail. Empirically the profile is non-increasing in every cell (Section 4.1), which is the form in which the corollary is applied. By Proposition 2, at a crossing where g takes different values on the two retention levels on both sides the accuracy contrast changes sign, and by Proposition 1(ii) the ranking is then depth-contingent; where g is flat over both levels, as it may be near $\rho = 0$ at the smallest budgets, that crossing produces no accuracy reversal of its own. Whether the pair is depth-contingent in accuracy is decided over the whole grid, not crossing by crossing: it is, exactly when $g(\rho_{m_1}) - g(\rho_{m_2})$ takes both signs over the depths, which can happen through depths on either side of a crossing at which g is flat (the template-versus-random case of Section 4.2 is one where it does not). □

Section 2 states the iso-accuracy result in words; formally:

Proposition 4. *Under Assumption 1 with α constant in p, the iso-accuracy set $\{(p, B) : a_m(p; B) = \kappa\}$ equals $\{(p, B) : \rho_m(p; B) \in g^{-1}(\{(\kappa - \beta)/(\alpha - \beta)\})\}$, the preimage under g of one level (a set of retention values, not one value, where g is flat). For a template press and evidence starting at or after position s, the superlevel set $\{a_m \geq \kappa\}$ for $\beta + (\alpha - \beta)g(0) < \kappa \leq \beta + (\alpha - \beta)g(1)$ is bounded, up to the straddle zone, by the tail boundary $p = 1 - (B - s)/T$ (under P_{late} $1 - (B - s - (P - T))/T$): budget moves that boundary linearly in B/T and buys nothing at any depth below it. For a press with a smooth profile and a strictly increasing g, a level set has slope $dp/dB = -(\partial\rho/\partial B)/(\partial\rho/\partial p)$ wherever $\partial\rho/\partial p \neq 0$ (Fig. 13).*

Proof of Proposition 4. With α constant, $a_m = \kappa$ iff $g(\rho) = (\kappa - \beta)/(\alpha - \beta)$, which is the preimage statement. For a template press, Proposition 3(b) gives $\rho(p; B) = \mathbf{1}[p \geq 1 - (B - s)/T]$ away from the sink and away from the straddle zone, so a_m takes two values, $\beta + (\alpha - \beta)g(0)$ below the line $p = 1 - (B - s)/T$ and $\beta + (\alpha - \beta)g(1)$ on and above it; the superlevel set of any κ between them is the tail, bounded by that line, and at a depth below the line ρ, hence a_m, does not change with B. For a smooth profile and a strictly increasing g, $a_m = \kappa$ iff $\rho = r$ for one r, and the level set $\rho(p, B) = r$ has the stated slope by the implicit function theorem. If α varies with p the level sets of

a_m and of ρ differ, and the slope acquires the term $\alpha'(p)g(\rho)$ in its denominator; the controlled grid has a flat uncompressed profile (Appendix J), which is why the constant-α form is the one used. □

B PRESSES AND BUDGET MATCHING

Every press returns, for every layer ℓ and KV head h, an ascending index set of exactly B_ℓ positions of the pressed prefix of length P, with $\sum_\ell B_\ell = LB$. All presses except PyramidKV use $B_\ell = B$; PyramidKV uses an arithmetic schedule from $1.8B$ at the first layer to $0.2B$ at the last, rounded and then adjusted by single tokens so that the total is exactly LB. Scores are computed per KV head from the eager attention of the uncompressed prefill; for grouped-query models the attention of the query heads sharing a KV head is averaged. Accumulated attention $c_i = \sum_{q \geq i} A_{q,i}$ is summed in float32 by a forward hook during a chunked prefill (chunk 256) so that no full attention matrix is materialised. The observation window is the last $w = 32$ positions of the pressed prefix; under P_{late} it contains the query, under $\mathrm{P}_{\text{early}}$ it is the end of the context. SnapKV and PyramidKV sum the attention of the w window queries to every earlier key, max-pool the result with kernel 7 (stride 1, same padding), exclude the window from the vote and keep the top $B_\ell - w$ keys plus the window. The window is $w = \min(32,\ P/4,\ B_\ell - 1)$, so for PyramidKV it follows the per-layer budget rather than staying fixed as in Cai et al.: at $B = 64$ the deepest layers of H1 keep 12 window positions and choose 1 by the vote, which makes PyramidKV behave like LOCAL there. The alternative, a fixed $w = 32$ with $B_\ell < 32$, would have to break the exact $\sum_\ell B_\ell = LB$ matching that every comparison in this paper rests on. TOVA scores by the attention of the last window query only. $\mathrm{H_2O}$ assigns $\lfloor B/2 \rfloor$ positions to the most recent tokens and the rest to the largest c_i among earlier positions; $\mathrm{H_2O}$-acc assigns all of B by c_i. KeyNorm keeps the B keys of smallest L_2 norm. StreamingLLM keeps positions $0..3$ and the last $B - 4$; LOCAL the last B. RANDOM draws a uniform B-subset per layer with a seed fixed by the item; RANDOM+OBS keeps the window and draws the remaining $B - w$ positions uniformly from the rest. Positions are not renumbered after eviction, following the KV-compression rather than the streaming use of StreamingLLM's pattern: the retained keys keep their RoPE phases and the cache index is local. Appendix V measures what the other convention would change.

The engine's verification suite (73 checks in 19 groups labelled T1–T18, released and runnable on CPU; the implementation follows the press interface of NVIDIA (2024)) asserts among other things that pruning with the full index set is the identity, that every press meets its budget exactly per layer and head, that the chunked prefill equals a single-shot prefill to a relative tolerance (10^{-3} in float32; chunking does change bfloat16 arithmetic, Appendix N), that the hooked attention statistics equal those from the reference `output_attentions` path, that a pruned cache leaks no future token into earlier predictions, that teacher-forced scoring on a pruned cache equals greedy decoding, that batched scoring of several presses equals scoring them one at a time, that the random press has $\rho = B/T$ in expectation, and that retention is computed per head rather than as a union over heads.

C TASK CONSTRUCTION AND PROMPTS

The preamble is RULER's: "Some special magic numbers are hidden within the following text. Make sure to memorize them. I will quiz you about the numbers afterwards." Filler sentences and needles share the template "One of the special magic numbers for *key* is: *value*." Needle keys are drawn without replacement from a list of 36 real words; filler keys are random pronounceable pseudo-words that never coincide with a needle key. Filler values are drawn from the same distribution as needle values. Needle j is inserted after $\lfloor p_j(n-1) \rceil$ of the n filler sentences, with slots made strictly increasing, so its relative depth is $p_j = j/(K-1)$ up to one sentence; n is calibrated per model tokenizer so that the prefilled context has the target length. The query is "What is the special magic number for *key* mentioned in the text above? Answer with the number only." The context and query are rendered with the model's own chat template; the split between the pressed prefix and the scored suffix is placed at a token boundary (a layout is chosen automatically if the default one merges tokens across the split). Task E values are 5-digit integers, which Qwen2.5 tokenizes as 5 tokens and Llama-3.1 as 2; task H values are three 4-digit groups joined by hyphens, tokenized as 14 and 8 tokens respectively. The two hyphens are thus 2/14 of a Qwen answer and 2/8 of a Llama one; with the evidence gone Qwen still emits them, which floors its graded metric near 0.14, while Llama, whose 4-digit groups split three plus one, puts its hyphens after the first three digits and scores near zero at zero retention (Appendix X.1). The evidence set E of a needle is its whole sentence, and

ρ_{cue} and ρ_{ans} restrict it to the key and the value span. The 32K cells run to about 32,970 positions including query and answer, slightly past the 32,768 of Qwen2.5-7B-Instruct's configuration, without RoPE scaling; uncompressed accuracy at H4's deepest needle is 0.95 against a cell mean of 0.89, so the overrun does not show in the reference. Task H was introduced because task E leaves strong retrievers at ceiling under $\mathrm{P}_{\mathrm{late}}$ (SnapKV, PyramidKV and TOVA at 0.93, 0.92 and 0.91 at $B = 256$ on Qwen2.5-7B, against 0.69, 0.55 and 0.70 on task H) and gives the two tokenizer families different token granularity. Evidence spans are located by character offsets and mapped to tokens; the key span and the value span of each needle are tracked separately for ρ_{cue} and ρ_{ans}. The uncompressed control of Appendix J builds genuine haystacks of 5 lengths with the same construction and 30 items each.

D STATISTICAL PROCEDURES

All curves and contrasts in a cell are computed on the items that are complete in every (protocol, press, budget, depth) cell; no item was dropped in any cell. The item-level bootstrap resamples items with replacement (2000 resamples) and recomputes every profile, so contrasts between presses and between depths are paired. A pair is sign-changing when the 95% interval of $\delta(p)$ excludes zero from above at one depth and from below at another. The two summaries the paper asks others to report are resampled the same way: each bootstrap replicate of δ gives one Φ and one pair of RII endpoints, and the reported intervals are their percentiles. For (SnapKV, StreamingLLM) at $B = 256$ in H1 the RII endpoints are [−0.16, −0.10] and [+0.69, +0.76] at 95%. Those are pointwise intervals, so we also report the same count under a simultaneous band: for each pair, the bootstrap max-t statistic over the K depths (each depth's replicate deviation from the replicates' mean divided by its bootstrap standard error, a depth with zero spread contributing zero) gives one band centred on that mean and covering the whole curve, and the Bonferroni variant takes that band at $0.05/45$ so the statement is family-wise over every depth of every pair (Table 2). The leave-one-press-out R^2 scores every predictor against the same baseline, the training mean, which is the only constant available before the held-out press is seen; the held-out press's own mean is not available at prediction time and is degenerate for the presses whose accuracy is nearly flat (LOCAL, RANDOM), for which it makes R^2 arbitrarily negative. Because the baseline does not depend on the predictor, the bake-off between ρ, ρ_{ans}, $\bar{\rho}$ and the budget is unaffected by the choice. The choice is in any case not what carries the claim. A depth permutation, shuffling ρ across the depths of each (press, budget), which leaves each press's marginal retention, its budget curve and $\bar{\rho}$ exactly as measured, gives the R^2 that the same numbers reach with the depth alignment destroyed, and it is the same fitting procedure, the same folds and the same baseline on both sides. 200 permutations per cell: in every one of the 16 cells for which the null was run (the 12 controlled-grid cells of Table 4 plus the single-needle control and the three natural-document cells of Appendix K, whose depths are placed on the same grid), and under both protocols, the observed R^2 exceeds all 200 (`analyze_nulls.py`). The per-depth ordering test is checked the same way, by permuting the press labels of ρ within each (budget, depth): the observed Kendall τ between the ρ and accuracy orderings of the ten presses exceeds its permuted 95th percentile in 82–96% of (budget, depth) cells under $\mathrm{P}_{\mathrm{late}}$ and 47–61% under $\mathrm{P}_{\mathrm{early}}$ across H1–H4. One property of the band needs stating rather than leaving to be discovered. A depth at which every item takes the same value for both presses has no item-level spread, so every bootstrap replicate of δ there is identical and the band collapses to a point, which declares significance whenever $\delta \neq 0$. That is what the bootstrap says, and at a saturating depth where one press scores every item and the other scores none it is not wrong, but it is a certificate with no width. We therefore checked all 58,860 (pair, depth) differences of the nine primary cells, over all 45 pairs, for constancy: 1,567 are constant and non-zero, every one of them in the 2 cells where the task saturates (E4, E5) and none in the 4 primary task-H cells that carry the headline counts. 28 of the 1,055 family-wise certifications over all pairs rest on such a depth, 13 of them published-pair certifications, all in those two cells (67 published certifications there in all), and those are the cells Appendix X.1 already reports the link failing on. A second property of the band is that its critical value at $0.05/45$ is the 99.89th percentile of 2,000 max-t draws, interpolated between the fourth and third largest by NumPy's linear rule (which by itself puts the expected per-pair non-coverage at 0.0016 rather than 0.0011), so it is itself a noisy quantity and a replicator with another seed will not reproduce a count exactly. We measured that too, re-running H1–H4 under both protocols with 5 bootstrap seeds: the largest published-pair count over budgets is identical across seeds in 6 of the 8 (cell, protocol) combinations and moves by at most 1 in the others, and by at most 2 for the count over all 45 pairs. The headline value, 8 published pairs in H4 under $\mathrm{P}_{\mathrm{late}}$, is one of the stable ones. At H1's largest budget the count is also unchanged when

the resamples are raised from 2,000 to 20,000 (spread 0 over the same 5 seeds), so the thin tail is not distorting it. Finally, the band's guarantee is family-wise over the depths and pairs of one (cell, protocol, budget), and every count this paper quotes as "up to" is a maximum over the six budgets of a cell and, where stated, over cells as well. That selection is not corrected for: Table 2 gives the minimum beside the maximum for every cell, and the maximum should be read as the largest family-wise statement the grid supports, not as a five-percent test in its own right. The prior family $\mathcal{D}_K$ consists of the K point masses on the grid; the uniform prior; $D \propto e^{-3p}$ (front-loaded), $D \propto e^{3p}$ (back-loaded) and $D \propto e^{-8(p-1/2)^2}$ (middle-loaded), all normalised on the grid; and uniform priors on the sub-grids $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$, $\{0, \frac{1}{2}, 1\}$ and $\{0, 1\}$, each target snapped to the nearest grid point, so on the 11-point grid the first is uniform on $\{0, 0.2, 0.5, 0.7, 1\}$ (on the 21-point grid the point masses are 21 and the family correspondingly larger; Table 2 reports distinct rankings out of the family size). Rankings are by mean accuracy with exact ties broken by press name (a stable sort), and Kendall's τ is computed on the prior scores themselves, a tied pair counting zero (τ_a). Empirical priors $\hat{D}$ are histograms of the located evidence depths of a task on the cell's grid, requiring at least 50 located items per task; every audited task clears that by a wide margin (Table 6). The link g is fitted by isotonic regression (pool-adjacent-violators, increasing, clipped outside the training range) on the mean accuracy and mean retention of every (press, budget, depth) cell; leave-one-press-out fits on nine presses and evaluates R^2 on the tenth relative to the training mean, averaged over the ten held-out presses. The gap between the R^2 of ρ and of $\bar{\rho}$ is bootstrapped over items: each of 200 resamples recomputes every cell mean and refits both links (`boot_r2.sh`). That interval treats the ten presses as fixed; Appendix X.6 gives the press-level view, which is the stricter one. The crossing depth of a pair is the first grid interval in which $\rho_{m_1} - \rho_{m_2}$ changes sign and the flip depth the first interval in which the mean contrast δ does; both are located by linear interpolation of the sign change inside that interval, which is why the median errors of Table 5 are not multiples of the grid step. The shape distance of Appendix J is the mean absolute difference between the compressed profile and a genuine short-context profile after each is divided by its own mean (a monotone reparametrisation of a flat profile is flat, so any hole is a distance from every reparametrisation); the reported range is over the best-matching short context per budget.

E FULL TABLES

Table 2: Ranking instability and mediation in every cell and protocol. "sign": sign-changing pairs of 45 (range over budgets), and "fw": the same under a bootstrap max-t band covering a pair's whole depth curve at a level Bonferroni-corrected over all 45 pairs; "fw int.": the family-wise count restricted to reversals both of whose significant depths lie strictly inside the grid (the published-pair count reaches 3 under P_{late} and 7 under P_{early}) — it is 0 at every budget but the largest in the task-H cells under P_{late}, and Appendix X.8 separates the two mechanisms that produce it; "pub.": sign-changing pairs among the 15 pairs of the six published methods of Section 1, pointwise and family-wise; "dc": depth-contingent pairs ($0 \in \operatorname{int} \mathrm{RII}$); "rank": distinct full rankings under $\mathcal{D}_K$ / family size; $\tau_{\min}$: minimum Kendall τ between two rankings in $\mathcal{D}_K$, over all budgets; "orders" and "winners": distinct orderings and distinct first-ranked presses under the empirical priors of the 12 audited tasks with a localisable answer span; R^2: leave-one-press-out mean R^2 for $\rho(p, B)$, $\rho_{\mathrm{ans}}(p, B)$, $\bar{\rho}(B)$ and budget only; "resid": largest absolute mean residual of the ρ link at any single depth; $\tau_{\rho a}$: median over (budget, depth) of Kendall τ between the presses ranked by retention and by accuracy.

cell	protocol	sign	fw	fw int.	pub.	dc	rank	$\tau_{\min}$	orders	winners	R^2_ρ	$R^2_{\rho_{\mathrm{ans}}}$	$R^2_{\bar{\rho}}$	R^2_B	resid	$\tau_{\rho a}$
H1	late	12–24	7–16	0–13	3–8 (1–7)	20–30	16–18/18	−0.24	3–8	1–2	0.77	0.85	0.58	0.05	0.063	0.67
H1	early	16–31	12–22	0–16	4–10 (3–6)	27–37	17–18/18	−0.38	7–9	2–3	0.84	0.83	0.26	0.10	0.025	0.36
H2	late	3–10	2–8	0–4	1–7 (1–6)	9–14	10–14/18	0.53	2–4	1–2	0.62	0.69	0.55	0.01	0.084	0.58
H2	early	8–19	7–12	0–10	4–11 (3–7)	18–24	13–17/18	0.16	3–6	1–2	0.45	0.36	0.30	0.09	0.063	0.52
H3	late	2–6	2–5	0–4	1–5 (1–4)	6–12	8–16/18	0.53	1–6	1–2	0.47	0.65	0.41	0.03	0.083	0.70
H3	early	14–19	6–18	0–5	5–12 (0–11)	20–29	14–18/18	−0.11	4–8	1–2	0.25	0.10	0.38	0.02	0.052	0.50
H4	late	17–25	13–17	0–13	7–11 (4–8)	21–32	19–24/28	−0.07	3–9	1–2	0.75	0.83	0.55	0.04	0.054	0.68
H4	early	21–29	12–22	0–17	5–9 (3–7)	34–39	25–28/28	−0.20	7–9	1–3	0.83	0.80	0.25	0.08	0.026	0.40
E1	late	4–20	0–16	0–10	2–7 (0–4)	25–31	16–18/18	0.00	3–9	1–1	0.59	0.82	0.41	0.03	0.097	0.36
E1	early	21–31	7–25	0–14	5–10 (2–5)	33–45	17–18/18	−0.47	7–11	1–3	0.79	0.87	0.25	0.07	0.025	0.31
E2	late	12–20	6–16	0–12	2–7 (1–4)	25–29	16–18/18	−0.09	4–8	1–2	0.64	0.82	0.46	0.02	0.097	0.40
E2	early	17–25	13–22	0–14	4–9 (3–7)	30–44	17–18/18	−0.38	5–10	1–4	0.81	0.88	0.27	0.03	0.039	0.48
E3	late	6–20	2–15	0–9	0–4 (0–2)	17–29	14–18/18	−0.16	4–7	1–1	0.69	0.85	0.48	0.02	0.078	0.50
E3	early	15–24	9–19	0–10	3–11 (2–5)	31–43	16–18/18	−0.33	5–8	1–3	0.80	0.89	0.26	0.02	0.030	0.45
E4	late	1–4	0–2	0–2	1–2 (0–2)	4–8	5–10/18	0.42	1–4	1–1	0.01	0.17	−0.08	0.00	0.105	0.54
E4	early	11–17	6–13	0–3	6–11 (2–10)	14–23	11–16/18	−0.13	4–8	1–3	0.19	0.35	0.37	−0.06	0.066	0.63
E5	late	0–3	0–2	0–2	0–2 (0–2)	3–8	4–14/18	0.40	1–5	1–2	0.05	0.19	−0.05	0.00	0.102	0.55
E5	early	10–15	3–12	0–4	5–10 (0–8)	12–19	10–15/18	−0.11	5–9	1–3	0.29	0.45	0.40	−0.05	0.074	0.62

Table 3: Every control in the paper and its outcome. The 18 italic rows of 48 are negative results; Section 8 scopes the claims by them.

control	outcome
Random eviction at matched budget	$\rho_{\mathrm{rand}}/(B/T)$ in 0.98–1.01; the position-unbiased null behaves (Proposition 3d)
Depth-permutation null on the link	observed beats all 200 permutations in 16 cells and both protocols
Label-permutation null on the ordering	above the permuted 95th percentile in 82–96% of (budget, depth) cells under P_{late} and 47–61% under P_{early} (Appendix D)
Noise null on the ranking counts	observed exceeds it in 139 of 144 combinations (Appendix X.2)
$K = 1$ unpaired layout	H1's counts reproduce to within a few pairs (family-wise 6–14 against 7–16; Appendix G); H2–H4 were not repeated
Multiplicity correction over 45 pairs	1–8 of 15 survive
Budget breakdown	12 of 12 cells certify a pair at the largest budget
Third model family (H5, Mistral-7B-Instruct-v0.3, 8K)	ranking claims reproduce; the link fitted there transfers out at R^2 0.64–0.82, links fitted elsewhere predict it at 0.36–0.48 (Appendix S)
Non-copyable answer	framework carries over (Appendix R)
RoPE position convention	shifts window templates by at most 0.05 and content-driven presses by up to 0.43 (Appendix V)
Leak-free intervention	floor 0.43–0.45 $\rightarrow$ 0.03–0.08; costs 0.000 where evidence is kept
Negative control for that intervention	finds nothing where the floor is the answer format (Appendix Q)
Grid density, 21 depths against 11	same certified and depth-contingent counts on the same records (Appendix X.4)
Removing the window's leak from the certified pairs	all 8 tested observations survive (3 more are unaffected by construction) and 2 appear (Appendix X.8)
False-positive rate of the certification	0 of 240 depth-permuted contrasts certify (Appendix X.4)
A press whose budget is allocated by a positional quota	0 and 1 of 6 flat pairs certified in the two cells against 5 and 7 of 15 published ones (Appendix X.9)
Zero-width bands in the simultaneous band	none in the 4 task-H cells; in E4–E5, 13 of 67 certified published pairs rest on one (Appendix D)
Bootstrap seed of the family-wise band	headline count identical over 5 seeds, as are 6 of 8 (cell, protocol) counts, the rest within 1 (Appendix D)
Run-to-run determinism on one machine	132 of 132 scored answers bit-identical (Appendix N)
Question kept whole in the four presses that lose most of it (P_{late})	published counts move by at most 1 per budget in H1, H3, H4; H_2O-acc's held-out R^2 rises from 0.11 to 0.79 in H1 (Appendix W.1)
A 2025 press from none of the ten's families	the link fitted on the ten predicts Expected Attention at R^2 0.79–0.88 under P_{early} with no refitting, against 0.24–0.45 for $\bar{\rho}$ (Appendix W.2)
Leave one family out; the press's own budget curve	R^2 falls by at most 0.04; the link's depth shape improves the own curve in 4 of 4 cells (Appendix W.3)
160 further natural-document items, another machine	2 published pairs certified under P_{late} at $B = 2048$ and none below, as on the released items; 6 under P_{early} (Appendix W.5)
Every press re-run on another machine (H1, H3, H4)	δ correlates at 0.999 or more; at least 95% of each run's certified pairs (94% of published ones) are certified in the other (Appendix W.5)
Proposition 2's sign at the significant depths	the accuracy contrast has the sign of the retention difference at 88.7% of 5,726 significant points in H1–H4 (Appendix H)
Audited task priors against the certified reversals	92 of 219 certified observations reversed by at least one (14 of 95 under P_{late}; Appendix X.12)
The audit on every row, with all spans and gold passages	the QA-1 swing is 0.43 over all occurrences and 0.55 at the gold passage (Appendix U)
Ranking-count null for the shaped priors and the published presses	exceeded in 73 of 108 and 107 of 108 combinations (Appendix X.2)
The model's own greedy output (H1–H4)	98 (P_{late}) and 135 (P_{early}) published reversals against 96 and 122 teacher-forced, 78 and 97 shared; all 45 pairs, 116 against 206 under P_{late} (Appendix W.4)
A fourth machine whose records differ (H1–H4)	41–91% of scores identical, yet δ correlates at 0.996 or more and 91% of certified pairs recur (Appendix W.5)
Exact match, and a cell built to have resolution	certifies at most 4 of 15 in every cell (Appendix P)
A released benchmark's own rows, 996 items	certifies 0 pairs: depth is observed in unpaired bands, accuracy is low and band-dependent, and few rows sit in a window's tail (Section 5)
LongBench's own rows, 593 of 800	its F1 certifies no published pair (contains-match: one, at $B = 2048$); retention predicts F1 poorly (Appendix K.1)
Re-weighting against the un-reweighted baseline	gains -0.18 under P_{late} on RULER's 996 rows (-0.09 on the pilots); we make no constructive claim (Appendix X.3)
Depth-averaged retention as a predictor	beats depth-resolved in 16 of 48 combinations (Appendix X.1)
The uniform average over layers and heads of Definition 1	weighting heads by their attention on the evidence predicts better, 0.92 against 0.77 (Appendix X.5)
Depth-contingency against a matched null	fires on 25% of depth-permuted contrasts, so it describes rather than tests (Appendix X.4)
The probe against simply decoding labelled items	1 labelled item over the grid beats it on R^2 (0.88 against 0.77; Appendix X.11)
Distinct-ranking counts for the flat presses	lower than the published presses' in one cell and not in the other, because near-tied means reshuffle for free (Appendix X.9)
Expected Attention under P_{late}	$\bar{\rho}$ predicts it better than ρ in 2 of 3 cells, and, on the fourth machine with its question protected, in 2 of 3 (Appendix W.2)
The first answer token alone, which teacher forcing cannot help	the P_{early} count holds (87 against 86), the P_{late} count falls to 14 of 84 (Appendix W.4)
The same cell measured on other hardware	no log-probability agrees and mean token accuracy shifts 0.039 one way, so levels need not transfer; contrasts do (Appendices N and W.5)
The audit's depth for a multi-answer task	the first of several gold spans, which moves 3 tasks by up to 0.32 (Appendix U)
The closed form $c_i \propto \ln(T/i)$ of Proposition 3(a)	fits the accumulated attention at $R^2 = 0.005$; only the monotone statement is used (Appendix X.1)
The retention crossing as a locator of the accuracy flip	Spearman 0.17–0.61 with median errors 0.064–0.350 depth units across H1–H4; claimed only where the error is under a grid step (Appendix H)
The band in its least favourable configuration	false-certification rate 0.0022 against the nominal 0.0011; at half the level no per-cell maximum changes (Appendix X.4)
The held-out press as the unit, under P_{late}	ρ ahead of $\bar{\rho}$ for 4–6 of ten presses; the advantage is the window templates' (Appendix X.6)
The shared floor β of Assumption 1	fails where a retained window carries the answer: E4, E5 and, more weakly, H3 (Appendix X.1)

Table 4: Every cell with a controlled depth grid. The nine primary cells carry Tables 1 and 2 and every range quoted as "across H1–H4" or "across the nine cells"; M, S and H5 are added in Appendices P, R and S and join the nine wherever a count is quoted over 12 cells. The largest budget is at most 12.6% of the cache in every cell but H5 (12.4–12.6%, the spread being the few hundred tokens by which context lengths differ); H5 reuses H1's ladder at half the context and so reaches 25%. Controls without a full depth grid — the single-needle cell (Appendix G), the short-context control (Appendix J), the sink sweep (Appendix M), the leak-free control (Appendix Q) and the RoPE ablation (Appendix V) — are described where they are used. The natural-document and released-benchmark cells of Section 5 are separate again, with their own metric.

cell	model	T	K	items	answer tok.	role
H1	Qwen2.5-7B	16,454	11	50	14	primary
H2	Qwen2.5-14B	16,452	11	40	14	primary
H3	Llama-3.1-8B	16,371	11	50	8	primary
H4	Qwen2.5-7B	32,891	21	40	14	primary
E1	Qwen2.5-7B	8,158	11	60	5	primary
E2	Qwen2.5-7B	16,368	11	100	5	primary
E3	Qwen2.5-7B	32,717	11	40	5	primary
E4	Llama-3.1-8B	16,332	11	60	2	primary
E5	Llama-3.1-8B	32,680	11	40	2	primary
M	Qwen2.5-7B	16,367	11	50	9	metric resolution
S	Qwen2.5-7B	16,397	11	40	4	non-extractive
H5	Mistral-7B	8,187	11	50	15	third family

Table 5: Crossing-predicts-flip (Proposition 2), key/value dissociation and protocol sensitivity, per cell. n: sign-changing (press pair, budget, protocol) triples with a located crossing; ρ_S, ρ_P: Spearman and Pearson correlation between crossing depth and flip depth, quoted without p-values because the triples share items and presses (Appendix H); err: median $|\text{crossing} - \text{flip}|$ in depth units; gap: largest depth-averaged $\rho_{\text{cue}} - \rho_{\text{ans}}$ over presses and budgets, with the press and budget attaining it; τ: Kendall τ between the uniform-prior rankings under P_{late} and P_{early} (range over budgets).

cell	n	ρ_S	ρ_P	err	gap	attained by	τ P_{late} vs P_{early}
H1	92	0.46	0.33	0.064	0.41	KeyNorm, B=2048	0.24–0.78
H2	88	0.25	0.20	0.295	0.48	TOVA, B=2048	0.38–0.69
H3	82	0.61	0.57	0.088	0.64	KeyNorm, B=2048	0.24–0.82
H4	114	0.17	0.10	0.350	0.44	H2O-acc, B=4096	0.20–0.60
E1	100	0.34	0.33	0.203	0.37	H2O-acc, B=1024	−0.02–0.56
E2	91	0.35	0.28	0.107	0.38	H2O-acc, B=2048	0.27–0.69
E3	81	0.21	0.14	0.187	0.38	H2O-acc, B=4096	0.49–0.69
E4	72	0.49	0.61	0.066	0.57	KeyNorm, B=2048	0.55–0.91
E5	47	0.33	0.39	0.096	0.57	KeyNorm, B=4096	0.64–0.87

Table 6: Gold evidence depth per benchmark task and context length: items sampled, items in which the evidence was located, and the mean and median relative depth among located items. RULER configurations are its 4096-, 8192- and 16384-token releases.

task	configuration	n	located	mean depth	median depth
Qasper	Qasper	281	276	0.25	0.18
HotpotQA	HotpotQA distractor	300	300	0.29	0.26
CWE	RULER 16K	500	500	0.06	0.07
FWE	RULER 16K	500	500	0.00	0.00
MK-NIAH-1	RULER 16K	500	500	0.51	0.51
MK-NIAH-2	RULER 16K	500	500	0.50	0.49
MK-NIAH-3	RULER 16K	500	500	0.50	0.48
MQ-NIAH	RULER 16K	500	500	0.19	0.16
MV-NIAH	RULER 16K	500	500	0.19	0.16
S-NIAH-1	RULER 16K	500	500	0.49	0.47
S-NIAH-2	RULER 16K	500	500	0.47	0.43
S-NIAH-3	RULER 16K	500	500	0.50	0.48
QA-1	RULER 16K	500	500	0.50	0.68
QA-2	RULER 16K	500	496	0.35	0.25
VT	RULER 16K	500	500	0.19	0.15
CWE	RULER 4K	500	500	0.23	0.26
FWE	RULER 4K	500	500	0.01	0.01
MK-NIAH-1	RULER 4K	500	500	0.54	0.57
MK-NIAH-2	RULER 4K	500	500	0.51	0.50
MK-NIAH-3	RULER 4K	500	500	0.50	0.49
MQ-NIAH	RULER 4K	500	500	0.20	0.15
MV-NIAH	RULER 4K	500	500	0.20	0.15
S-NIAH-1	RULER 4K	500	500	0.50	0.50
S-NIAH-2	RULER 4K	500	500	0.51	0.49
S-NIAH-3	RULER 4K	500	500	0.53	0.54
QA-1	RULER 4K	500	500	0.72	0.90
QA-2	RULER 4K	500	489	0.45	0.36
VT	RULER 4K	500	500	0.24	0.21
CWE	RULER 8K	500	500	0.12	0.13
FWE	RULER 8K	500	500	0.01	0.01
MK-NIAH-1	RULER 8K	500	500	0.53	0.56
MK-NIAH-2	RULER 8K	500	500	0.50	0.52
MK-NIAH-3	RULER 8K	500	500	0.50	0.51
MQ-NIAH	RULER 8K	500	500	0.20	0.16
MV-NIAH	RULER 8K	500	500	0.20	0.16
S-NIAH-1	RULER 8K	500	500	0.49	0.47
S-NIAH-2	RULER 8K	500	500	0.50	0.47
S-NIAH-3	RULER 8K	500	500	0.52	0.53
QA-1	RULER 8K	500	500	0.28	0.34
QA-2	RULER 8K	500	490	0.47	0.36
VT	RULER 8K	500	500	0.20	0.17

Table 7: Transfer of the retention link: R^2 when g fitted on the row cell (all ten presses) predicts every (press, budget, depth) cell of the column cell from its own retention profile, with no refitting. Diagonal: in-sample.

	P_{late}					P_{early}				
fitted on	H1	H2	H3	H4	H5	H1	H2	H3	H4	H5
H1	0.87	0.75	0.67	0.86	0.48	0.93	0.69	0.62	0.89	0.53
H2	0.83	0.79	0.73	0.81	0.43	0.87	0.75	0.68	0.84	0.49
H3	0.75	0.75	0.76	0.70	0.36	0.76	0.67	0.74	0.67	0.32
H4	0.86	0.75	0.66	0.88	0.48	0.90	0.67	0.59	0.92	0.53
H5	0.82	0.71	0.64	0.81	0.59	0.62	0.43	0.27	0.52	0.74

Table 8: Per-press view of cell H1 at $B = 256$, P_{late}: uniform-prior mean, worst-depth score W_m with its item-level bootstrap interval (a minimum over K noisy estimates is downward-biased, and the bias grows with K, so W_m is comparable only within a stated grid), lead fraction Φ against SnapKV (fraction of depths at which the press beats SnapKV), both with item-level bootstrap intervals, rank under the uniform prior and under the point mass at $p = 1$, and the leave-one-press-out R^2 obtained when this press is the held-out one, over all budgets of the cell (one column per protocol).

press	mean	W_m [95% CI]	Φ vs SnapKV	rank (unif.)	rank (p=1)	R^2 P_{late}	R^2 P_{early}
SnapKV	0.69	0.60 [0.56, 0.63]	–	2	5	0.94	0.94
PyramidKV	0.55	0.44 [0.40, 0.48]	0.00 [0.00, 0.00]	3	6	0.90	0.92
TOVA	0.70	0.62 [0.59, 0.64]	0.64 [0.45, 0.91]	1	4	0.75	0.35
H2O	0.33	0.21 [0.19, 0.21]	0.09 [0.09, 0.09]	4	1	0.90	0.93
H2O-acc	0.26	0.21 [0.19, 0.21]	0.00 [0.00, 0.00]	5	7	0.11	0.82
StreamingLLM	0.21	0.14 [0.11, 0.14]	0.09 [0.09, 0.09]	6	2	0.97	0.98
Local	0.19	0.09 [0.07, 0.10]	0.09 [0.09, 0.09]	7	3	0.98	0.99
KeyNorm	0.17	0.14 [0.12, 0.15]	0.00 [0.00, 0.00]	8	9	0.23	0.76
Random	0.14	0.13 [0.10, 0.13]	0.00 [0.00, 0.00]	10	10	0.93	0.90
Random+obs	0.14	0.12 [0.10, 0.13]	0.00 [0.00, 0.00]	9	8	0.96	0.86

Table 9: Metric robustness on Qwen2.5-7B at 16K. Task E (5-digit answers) analysed with the graded token metric and with exact match; task H (12-digit grouped answers) with exact match, on which even the uncompressed model is near floor. Entries as in Table 2. The uncompressed rows of the two protocols differ only in where the prompt is split between prefill and scored suffix, so their small differences are the arithmetic spread of Appendix N.

cell	metric	protocol	uncompressed	sign	rank	R^2_ρ	R^2_ρ / R^2_B
E2	token	late	0.95	12–20	16–18	0.64	0.46 / 0.02
E2	token	early	0.95	17–25	17–18	0.81	0.27 / 0.03
E2	exact match	late	0.76	1–8	5–12	0.51	0.46 / −0.04
E2	exact match	early	0.76	0–13	5–12	0.58	0.23 / −0.03
H1	token	late	0.89	12–24	16–18	0.77	0.58 / 0.05
H1	token	early	0.89	16–31	17–18	0.84	0.26 / 0.10
H1	exact match	late	0.19	0–2	1–11	0.19	0.49 / −0.40
H1	exact match	early	0.16	0–6	2–11	0.51	−0.10 / −0.19

Table 10: First-ranked press under the empirical priors of the 12 audited tasks with a localisable answer span (Appendix U), cell H1, by budget and protocol (number of tasks crowning each press in parentheses).

B	P_{late}	P_{early}
64	TOVA (12)	H2O-acc (10), H2O (2)
128	TOVA (12)	H2O-acc (6), H2O (6)
256	SnapKV (6), TOVA (6)	H2O-acc (6), H2O (6)
512	SnapKV (12)	H2O (7), H2O-acc (5)
1024	SnapKV (12)	SnapKV (6), H2O-acc (5), H2O (1)
2048	SnapKV (12)	SnapKV (8), H2O-acc (4)

Table 11: The 160-item natural-document cell (Section 5), by budget: published pairs certified under the Bonferroni max-t band (fw) and sign-changing under pointwise 95% intervals (pw), out of 15, for RULER's contains-match and for the graded metric, under both protocols. The 30-item pilot of the same design was analysed with pointwise intervals only (Appendix K).

	contains-match				graded			
	P_{late}		P_{early}		P_{late}		P_{early}	
B	fw	pw	fw	pw	fw	pw	fw	pw
64	0	0	0	0	0	0	0	0
128	0	1	0	2	0	1	0	5
256	0	1	0	6	0	1	2	6
512	0	0	0	5	0	2	3	5
1024	0	1	1	7	0	2	4	7
2048	2	2	7	10	2	4	8	10

F ADDITIONAL FIGURES

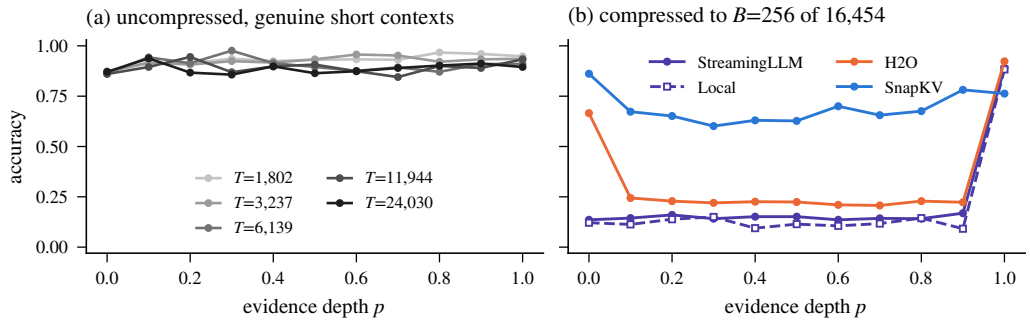

Figure 3: (a) The uncompressed model on genuine short contexts of five lengths: flat, with headroom. (b) The same model at 16K compressed to $B = 256$: the template presses have a hole that no short context has. Qwen2.5-7B, task H, P_{late} (Appendix J).

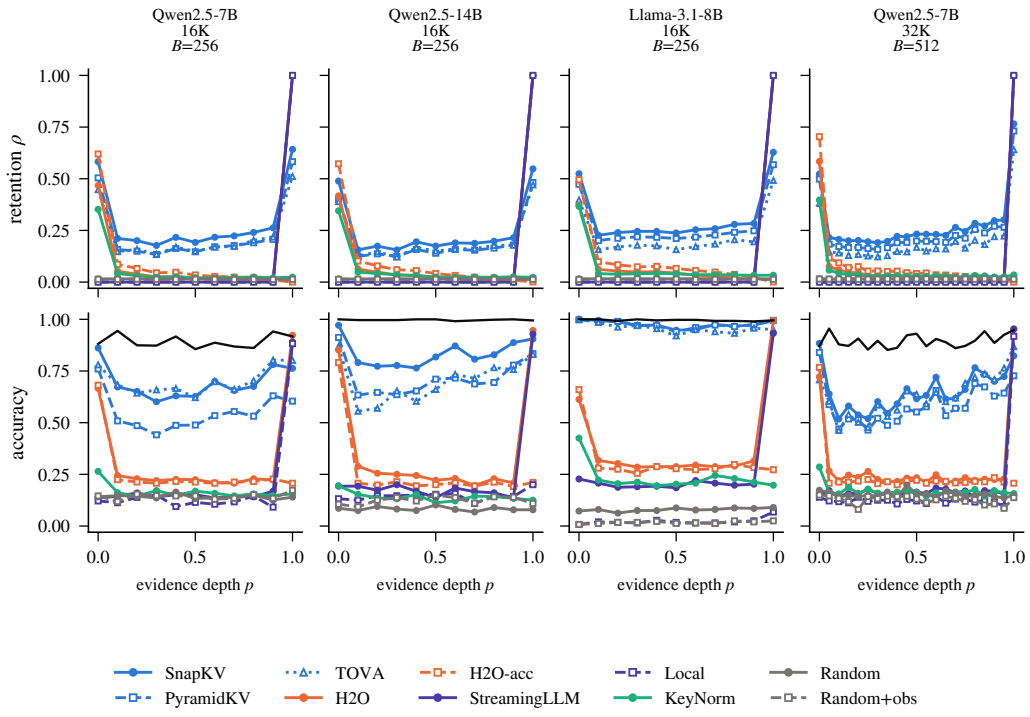

Figure 4: Retention profiles (top) and position-resolved accuracy (bottom) at the third budget of each task-H cell, P_{late}; the black line is the uncompressed model. Legend as in Fig. 1.

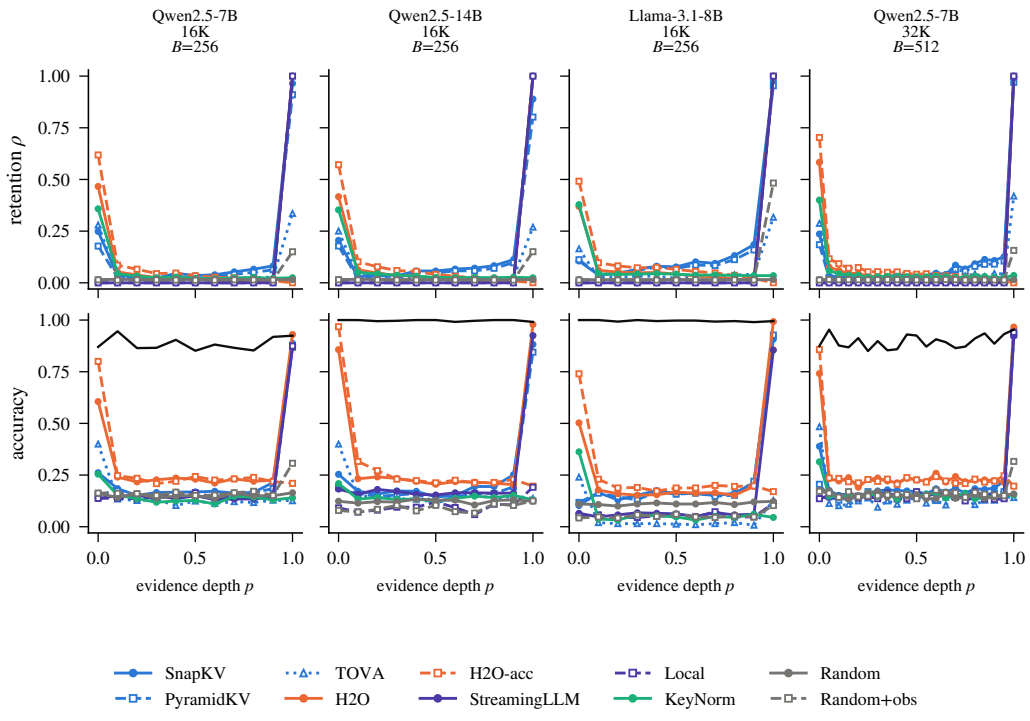

Figure 5: As Fig. 4, under P_{early}.

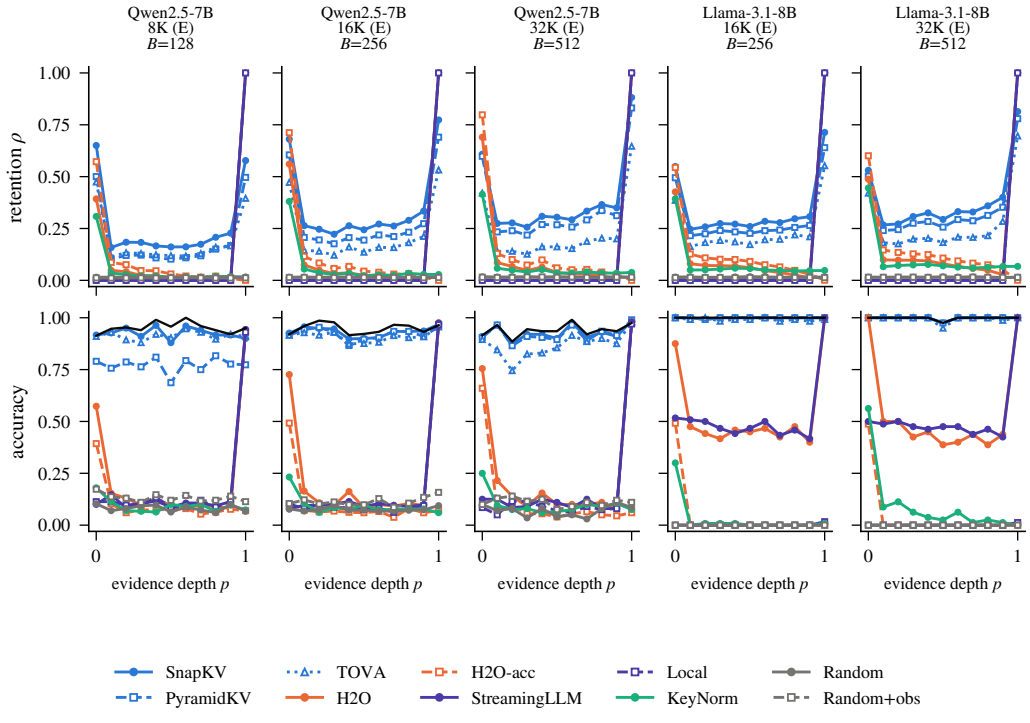

Figure 6: As Fig. 4, for the five task-E cells. Llama-3.1-8B at both lengths is the saturated regime of Section 4.4.

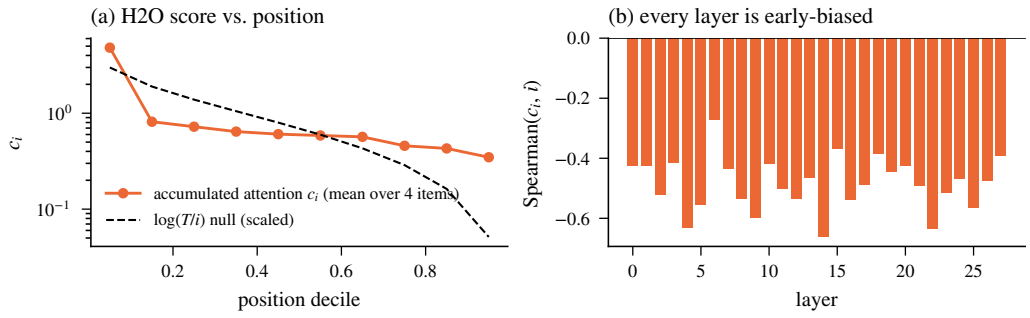

Figure 7: The accumulated-attention score of Proposition 3(a) on Qwen2.5-7B at 16K, averaged over 4 items. (a) Mean score by position decile against the $\ln(T/i)$ null (log scale): the first decile carries the attention sink, the rest decays more slowly than the null. (b) Spearman correlation between score and position in each of the 28 layers, over positions within an item; all negative, though far from -1 (the -1.00 to -0.98 of Section 4.1 is the correlation of H_2O-acc's *retention profile* with depth over the grid, a different statistic). Legend: the null is $\ln(T/i)$, rescaled.

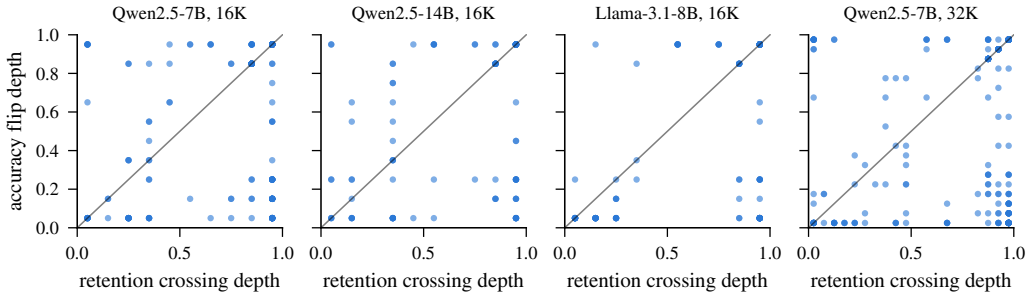

Figure 8: Accuracy flip depth (y) against retention crossing depth (x) for every (pair, budget, protocol) triple of the 45 pairs whose pointwise intervals make it sign-changing, in the task-H cells, both depths placed at the midpoint of the grid interval in which the sign changes; Table 5 and Appendix H use the interpolated locations of the published pairs instead. Many triples coincide on the grid, so darker points are several triples and the cloud understates multiplicity.

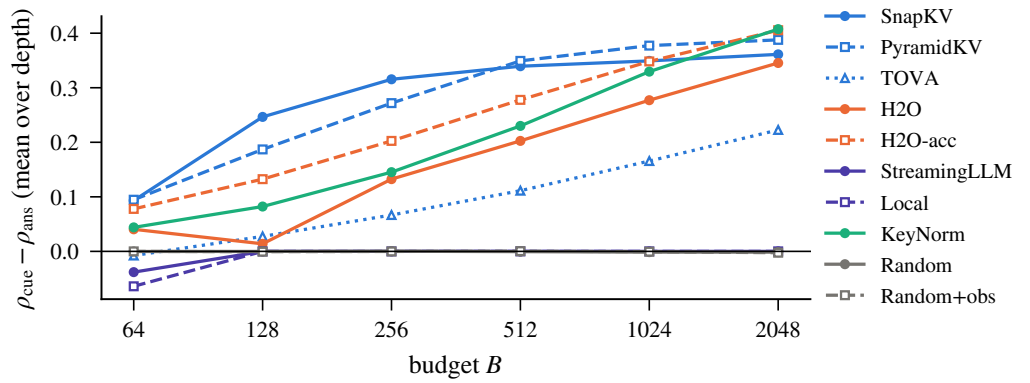

Figure 9: Cue/answer dissociation in H1 under P_{late}: the depth-averaged gap $\rho_{cue} - \rho_{ans}$ by budget. Every content-scoring press keeps more of the matchable cue than of the emittable answer as the budget grows; KeyNorm and H_2O-acc most at $B = 2048$, with SnapKV and PyramidKV close behind. RANDOM and, beyond $B = 64$, StreamingLLM sit on the zero line.

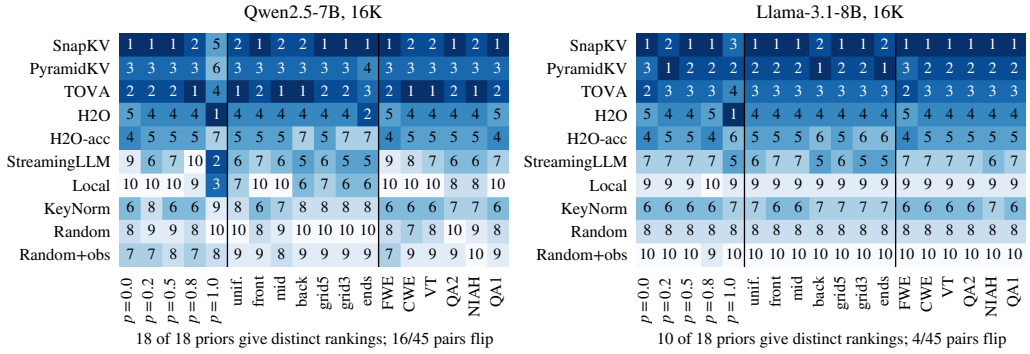

Figure 10: Rank of each press (1 = best) at $B = 256$, P_{late}, under 18 columns: 5 of the 11 point masses of $\mathcal{D}_{11}$ and its 7 shaped and coarse-grid priors (left of the second rule), then the empirical priors of six RULER sub-tasks (Section 6; FWE, CWE and VT are the three Appendix U excludes from the re-weighting claims and are shown for completeness, NIAH is `niah_single_1`), which are not in $\mathcal{D}_{11}$. The panel axis label counts distinct rankings over all of $\mathcal{D}_{11}$, not over the columns shown, and "pairs flip" counts pairs whose order differs between two members of the family. The same measurements, re-weighted, place StreamingLLM anywhere from 2nd to 10th in the Qwen2.5-7B panel.

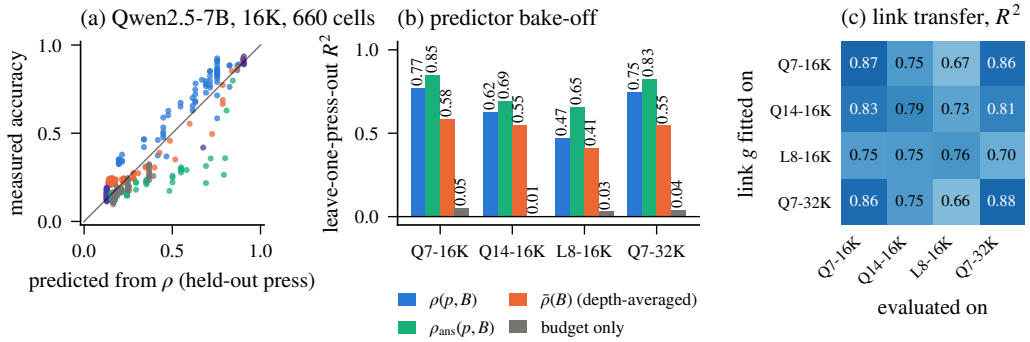

Figure 11: Retention mediation, P_{late}. (a) Measured against predicted accuracy for every (press, budget, depth) cell of H1, each press predicted by a link fitted on the other nine. (b) Leave-one-press-out R^2 of four predictors in the task-H cells; the budget-only bar is the pooled accuracy-versus-budget curve of the other nine presses, which carries no press identity, not the held-out press's own curve. (c) R^2 when the link fitted on the row cell, on all ten presses, predicts every (press, budget, depth) cell of the column cell from its retention profile alone; the diagonal is in-sample, and these values are not comparable to the held-out-press values of (b). Cells are H1 = Q7-16K, H2 = Q14-16K, H3 = L8-16K, H4 = Q7-32K.

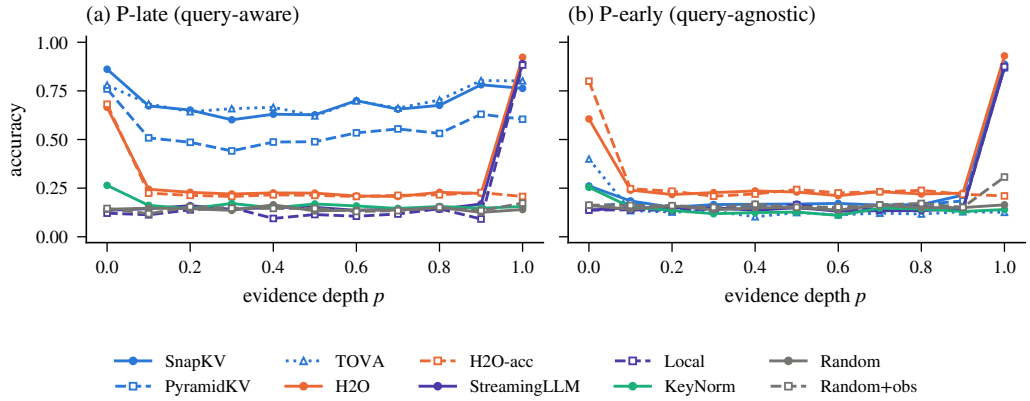

Figure 12: The ten presses of Fig. 1 at $B = 256$ in H1 under the two protocols (a: as Fig. 1b, without the uncompressed line). Under P_{early} the query is outside the observation window and the query-anchored presses lose most of their advantage.

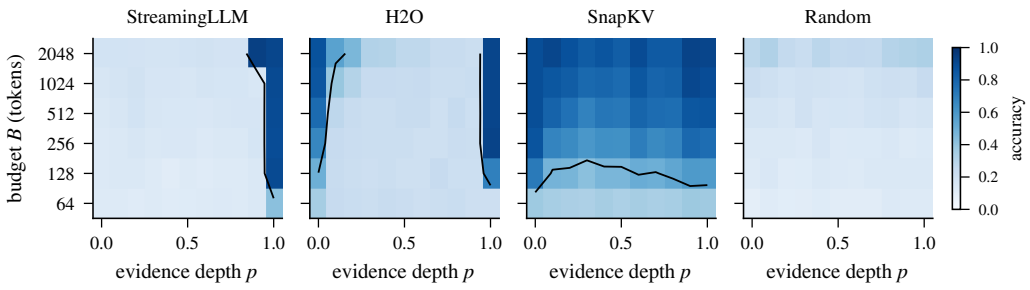

Figure 13: Accuracy surfaces $a_m(p, B)$ in H1 under P_{late} with the 0.5 iso-accuracy contour (Proposition 4). StreamingLLM's boundary sits at the window's edge and moves only as $(B-s-(P-T))/T$, with $P - T = 26$ query tokens, from $p \approx 0.998$ at $B = 64$ to 0.88 at $B = 2048$ (Proposition 4): budget buys no middle-depth accuracy. H_2O has an early boundary that moves with budget and a late, window-like one. SnapKV's is graded. RANDOM never reaches 0.5 at these budgets.

G THE SINGLE-NEEDLE CONTROL

The nine primary cells place $K = 11$ or 21 needles in one haystack so that every depth comparison is paired within an item. That design controls variance, but it also means every measurement is made with K competing evidence spans present, which is not the layout the NIAH literature uses. The control repeats cell H1 (same model, length, presses, budgets and protocols) with $K = 1$: each item carries one needle, its depth is drawn from the same 11-point grid, and depth therefore varies across items rather than within them, so the comparison across depth is unpaired. 330 items, 16,381 tokens, 40,260 scored answers.

Table 12: The single-needle control against the paired multi-key cell it repeats. Ranges are over the six budgets. "sign" and "fw": sign-changing pairs of 45, pointwise and under the family-wise band; "pub. fw": the same among the 15 pairs of the six published methods; "rank": distinct full rankings of 18; R^2: leave-one-press-out for ρ / $\bar{\rho}$.

design	protocol	sign	fw	pub. fw	rank	R^2_ρ / $R^2_{\bar{\rho}}$
K=1, unpaired	late	13–21	6–14	1–7	16–17	0.76 / 0.58
K=11, paired (H1)	late	12–24	7–16	1–7	16–18	0.77 / 0.58
K=1, unpaired	early	19–30	8–21	2–5	17–18	0.86 / 0.23
K=11, paired (H1)	early	16–31	12–22	3–6	17–18	0.84 / 0.26

The per-depth ordering test agrees as well: Kendall τ between the retention and accuracy orderings runs 0.33–0.85 under P_{late} in the control against 0.36–0.90 in H1. Nothing in Section 4 depends on the multi-key layout; the layout buys the paired bootstrap, which is why the primary cells use it.

H PROFILE CROSSINGS AND ACCURACY FLIPS

For the 15 published pairs whose mean contrast changes sign with depth (the point-estimate criterion, which Appendix X.4 shows fires on a quarter of depth-free contrasts, so this is a description of where flips sit, not a test that they exist), we compare the first depth at which it does with the first depth at which the retention profiles cross, both interpolated within the grid interval (Appendix D). Proposition 2 predicts only that a flip lies at a crossing, not that the first crossing is a flip, so this is a test of the stronger reading. Over 92 such (pair, budget, protocol) triples in H1 the two agree with Spearman 0.46 and a median absolute error of 0.064 depth units, under one grid step. We quote no p-value for it: those triples are not independent: they come from the same 50 items, each press appears in five pairs, and each pair recurs at twelve (budget, protocol) settings, so the effective sample size is a small fraction of 92 and any nominal p would be far too small. Read the correlation as descriptive; across H1–H4 it is 0.17–0.61 with median errors 0.064–0.350 (Fig. 8, Table 5). By rank correlation it is weakest at 32K in both tasks, E3 (Spearman 0.21, median error 0.187) and H4 (0.17, median error 0.350, seven steps of its finer grid), so the finer grid does not rescue it; by median error the worst cells are H4, H2 (0.295) and E1 (0.203). We claim the location only where the median error is under one grid step, which is H1, H3, E4 and E5; in H2, H4 and E1–E3 it is not, and we do not.

The direct test. Location is the strong reading; Proposition 2 itself makes a weaker and sharper prediction, that a certified reversal cannot occur without a retention crossing, and that at each significant depth the accuracy contrast has the sign of the retention difference (or the difference is zero). Over the 219 certified published-pair observations of H1–H4 under both protocols, 213 have a retention crossing; the rest are 6 P_{late} observations of (H_2O, SnapKV) and (PyramidKV, TOVA), in each of which the answer-span retention ρ_{ans} does cross with the accuracy contrast's sign. A crossing is also present in 70% of the observations that are not certified, so its presence alone is weak evidence; the sign test is sharper. At the 5,726 Bonferroni-significant (published pair, budget, depth) points of H1–H4 the accuracy contrast has the sign of the retention difference in 88.7% of cases and the opposite sign in 10.8%, where the retention difference is small (median $|\Delta\rho|$ 0.013); 41 of the 219 certified observations contain at least one such depth, concentrated in H3 under P_{early} and H4 under P_{late}, and many of them are a window template at zero retention beating a press that keeps a little of the evidence, the failure of the shared-β half of Assumption 1 that Appendix X.1 reports. We read

this as support for the mediation account in the direction it is stated, and not as a test with a stated error rate: the observations share items and recur across budgets.

I CUE AND ANSWER TOKENS DISSOCIATE

A press can keep the cue the query matches without the answer the model must emit. The depth-averaged gap $\rho_{\text{cue}} - \rho_{\text{ans}}$ reaches 0.41–0.64 across the task-H cells, and in H1 at $B = 2048$ KeyNorm keeps 0.57 of the cue tokens against 0.16 of the answer tokens under either protocol (Fig. 9 shows P_{late}; Table 5); SnapKV and PyramidKV show gaps nearly as large at that budget. Accordingly ρ_{ans} is the best single predictor in every task-H cell under P_{late} (Table 2).

J COMPRESSION IS NOT CONTEXT SHORTENING

Definition 2 (Equivalence to shortening)**.** Compression to budget B is equivalent to shortening if $a_m(p; B) = a_B^{\text{full}}(\varphi(p))$ for a monotone φ, where a_B^{full} is the uncompressed profile on a genuine length-B context.

We measured the uncompressed model (Qwen2.5-7B) on genuine contexts of 5 lengths, 1,802 to 24,030 tokens, on task H (Fig. 3). The profiles are flat to within 0.112 at every length at a mean accuracy of 0.91: this model has no intrinsic lost-in-the-middle deficit on this task, the flatness is not a ceiling effect, and on this model and task every position dependence measured under compression is induced by the press (the natural-document cells, whose uncompressed profiles do vary, are treated separately in Appendix K). The comparator is not a context of length B, which an 11-needle haystack cannot fit in at $B \leq 1024$, but the shortest genuine contexts the design admits; since every genuine profile is flat, and a monotone reparametrisation of a flat profile is flat, the test reduces to whether the compressed profile is flat to the same tolerance. Against that control StreamingLLM produces an interior hole at an L_1 shape distance of 0.30–0.68 from the closest genuine short context, against a flatness tolerance of 0.112; the decision is for StreamingLLM alone. For the query-anchored presses the profile is within 0.02–0.12 of some short context and we make no claim.

K NATURAL-DOCUMENT CONSTRUCTION AND RESULTS

The pool is SQuAD v1.1 validation (Rajpurkar et al., 2016): paragraphs of 300–1800 characters with exactly one question whose gold answer occurs exactly once in the paragraph, one question per paragraph, 1,200 paragraphs from 29 articles. An item draws K gold paragraphs from distinct articles and fills the rest with paragraphs from articles that contribute no gold; a gold candidate whose answer string appears in any filler or in another gold's paragraph is rejected before assembly whenever an unambiguous candidate exists, and otherwise the least ambiguous one is kept: in the 160-item cell 2 of 1,760 needles have their answer string elsewhere in the context. Gold paragraph j is inserted at the filler index whose cumulative character position is closest to $j/(K-1)$, which places the gold sentences on the target grid to within 0.03 in depth (checked on the 220 needles of 20 rebuilt items). The context is the concatenation "`Document` i`:`" followed by each paragraph, under RULER's qa_1 instruction; each question is asked separately through the model's chat template. The evidence span is the sentence containing the answer, the answer span is the answer string, and the cue span is the rest of that sentence (the remainder after the answer when the answer opens the sentence). Accuracy is a contains-match of any accepted answer in 16 greedily decoded tokens, the rule RULER itself uses, computed by the same batched decoder as the graded metric.

The controlled natural design is also run on Llama-3.1-8B at 7,936 tokens with four budgets, $B \in \{32, 128, 512, 1024\}$ (30 items, 27,060 generations, on the 24 GB card of Appendix N, where 16K did not fit for this model at the batch size the design needs; it is run as a floor check and enters no count). That cell is where the press-dependent floor of Section 4.4 shows up most clearly on natural text (the Qwen cell shows it too, 0.16 against 0.08 in the interior at zero retention): at the 9 depths at which StreamingLLM and LOCAL both retain none of the evidence, the first answers 0.41 of the time and the second 0.00, while RANDOM, which keeps B/T of the evidence at every depth, answers 0.00. The sink tokens, not the retained evidence, carry that difference.

The RULER-verbatim cells take the released `qa_1` and `qa_2` rows of the 4096- and 16384-token sets unchanged, one question per row, and locate the evidence as the sentence containing the answer string in the released context. They run the ten presses at 3 budgets ($B \in \{128, 512, 2048\}$, which at 4K reach half the cache) under both protocols rather than the six of the controlled cells, because every located row is scored rather than a sample of them: the 996-item 16K cell of Section 5 is 61,752 generations at that design, and Appendix X.3 uses 120-row pilots of each task at each length. Depth is observed, not set, so items are grouped into 5 equal-width depth bands rather than placed on a grid. Equal-width bands are used because the point of the grouping is depth contrast: the observed distribution is bimodal (Fig. 2a), and equal-count bands over it can be a few hundredths wide. `qa_2` is two-hop, so the located span is the second hop and its retention understates the evidence the model needs; the cell is reported as an external-validity check, not as a mediation test. Uncompressed accuracy itself varies with the observed depth (0.24–0.61 across the five bands of 70–341 items), which the controlled design removes and this one cannot; in aggregate 13–14 of the 15 published pairs differ under a paired 95% bootstrap on the uniform mean, but no sign change is certified at any budget and at most 1 is pointwise.

Under the graded metric of Section 3 the 160-item SQuAD cell gives 1–9 pointwise sign-changing pairs of 45 under P_{late} and 2–23 under P_{early}, certifies 2 published pairs at its largest budget (Table 11), and gives leave-one-press-out R^2 0.17 and 0.35 for ρ against 0.04 and 0.00 for the budget. The graded metric's mediation fit is weaker here than the contains-match's because a natural answer shares a prefix with the question's own wording, which gives it a high content-dependent floor; that is why the benchmark's own rule is primary in Section 5.

The 160-item cell carries Section 5. Its depth-permutation null is run under the benchmark's own contains-match rather than the graded metric: the observed leave-one-press-out R^2 is 0.58 against a null of 0.14 (upper bound 0.23) under P_{late} and 0.59 against -0.03 (0.01) under P_{early}, beating all 200 permutations in both. The 30-item pilot of the same cell (four budgets, pointwise intervals only, at most 16 sign-changing pairs of 45 under contains-match) is retained in the released records because Section 5's point about sample size is read off the two together: nothing about its construction differs, and its items are not among the 160.

The link fitted on H1 does not transfer to the 160-item cell: R^2 0.32 under P_{late}, far below every synthetic-to-synthetic entry of Table 7, and -4.16 under P_{early}. The two tasks differ in floor and ceiling (the natural task's uncompressed accuracy ranges 0.66–0.78 across depths against near-ceiling on the synthetic grid, and the link was fitted on teacher-forced token accuracy and applied to contains-match) so a link calibrated on one is not expected to be calibrated on the other. What does carry over is the within-cell fit and the per-depth ordering: the label-permutation test of Appendix D run on this cell puts the ρ-versus-accuracy ordering above its permuted 95th percentile in 18% of (budget, depth) cells under P_{late} (median τ 0.32).

K.1 LongBench's own rows

LongBench (Bai et al., 2024) is where many KV-compression papers report. We ran the ten presses on its four English QA subsets whose gold answers are mostly spans of the context, HotpotQA, 2WikiMultihopQA, MuSiQue and Qasper (MultiFieldQA-en's are mostly free-form), with LongBench's own prompt templates split at the context: the instruction and the context form the prefix that P_{early} presses, and the closing instruction and the question follow it, pressed too under P_{late} as on the grid (`fetch_longbench.py`). The evidence is located as in the RULER-verbatim cells, at the first case-insensitive whole-word occurrence of any accepted answer in the context, extended to its sentence; a row whose answers do not occur in the context (yes/no, "unanswerable", abstractive answers) carries no locatable evidence and is dropped, leaving 593 of 800 rows (HotpotQA 161, 2WikiMQA 178, MuSiQue 182, Qasper 72). Qwen2.5-7B decodes 32 greedy tokens, LongBench's own limit for three of the subsets (Qasper's is 128), and each output is scored by LongBench's QA-F1 over the whole output, the rule its evaluation applies to these subsets; budgets $B \in \{128, 512, 2048\}$ under both protocols, 36,766 generations on the fourth machine of Appendix W.5. Contexts run from 1,707 to 20,118 tokens (median 13,087), so, unlike on the grid, one budget is a different share of each row's cache, and for 4 rows the largest budget compresses nothing. Mean located depth is 0.29 (HotpotQA), 0.38 (2WikiMQA), 0.36 (MuSiQue) and 0.45 (Qasper; the audit, which reads Qasper's annotated evidence rather than the answer string, puts it at 0.25), and the uncompressed F1 of a subset is 0.32–0.60.

The subsets do not share an ordering. At every (protocol, budget) the four subsets order the ten presses in four different ways. Among the six published presses they give 2–3 orderings per budget under P_{late}, with H2O, PyramidKV and SnapKV each first somewhere, and 3–4 under P_{early}, with H2O and KeyNorm first somewhere and Kendall τ between two subsets' orderings as low as -0.53. Pooled over the subsets, 13–14 of the 15 published pairs differ under a paired item bootstrap under P_{late} and 7–13 under P_{early}, yet no published pair is ordered oppositely by two subsets with both subsets' 95% intervals excluding zero. The subsets differ in task as well as in evidence placement, so their disagreement cannot be put down to D; it says only that LongBench's average hides it.

Their priors are not a reliable forecast. Re-weighting H1's accuracies by a subset's measured prior, as in Appendix X.3, changes the Kendall τ with that subset's measured ordering by -0.16 on average under P_{late} (better in 0 of 12 (subset, budget) combinations) and by $+0.09$ under P_{early} (better in 8 of 12), against mean baselines of 0.67 and 0.40. The query-agnostic protocol is again the more favourable case, as on RULER's rows (Appendix X.3); with 12 dependent combinations and a gain of that size we make no claim for it.

Depth within the benchmark. Grouping all rows into five equal-width bands of observed depth (69–235 rows) and running the procedure of the RULER-verbatim cells unchanged (`analyze_natural.py`), LongBench's own F1 certifies no published pair at any budget under either protocol (pointwise, at most 1 under P_{late} and 4 under P_{early}). Under RULER's contains-match one is certified, (H2O, StreamingLLM) at $B = 2048$, under both protocols. The reasons are those of Section 5: depths are unpaired groups of rows that also mix subsets of different difficulty, the uncompressed F1 falls from 0.48 in the first band to 0.28 in the last, and only 94 of the 593 rows put their answer inside StreamingLLM's retained tail under P_{late} at $B = 2048$ (11 at $B = 512$). Retention predicts F1 poorly on these rows: leave-one-press-out R^2 is 0.12 for ρ against 0.26 for $\bar{\rho}$ under P_{late}, and 0.28 against 0.17 under P_{early}. On a multi-hop row the located span is one answer string, not all the passages the model needs, a limit of locating evidence by string match that Appendix U states for the audit.

L BENCHMARK AUDIT PROCEDURE

For RULER we sweep every row of the public 4096-, 8192- and 16384-token releases of its thirteen sub-tasks (19,500 rows, 500 per sub-task and length) and locate the gold evidence as the first character offset of the answer string in the context (for the frequent-words and common-words tasks the first occurrence of the first answer word; for variable tracking the first assignment in the chain), divided by the context length. For HotpotQA-distractor (300 items) the evidence is the first sentence listed in `supporting_facts`, located in the concatenation of the ten paragraphs in their released order; for Qasper (281 items, 276 located) it is the first annotated evidence span of the first answer, located in the full text. Items whose evidence could not be located by string match are reported as not located and excluded (Table 6). Depths are relative character positions, which differ from token positions by tokenizer-dependent factors that are nearly constant within a task. The audit is a measurement of where these benchmarks place their evidence, not a claim about the difficulty of the tasks.

M WHAT THE FLOOR TURNS ON, MEASURED

Section 4.4 attributes the press-dependent floor to a retained query window that was prefilled with the evidence in view, and to sink tokens that gate its readout. That is a falsifiable claim about two quantities the same uncompressed prefill already gives: $\sigma_m(B)$, the fraction of the first 4 positions the press keeps, and $\omega_m(B)$, the fraction of the last 32. Neither needs decoding, labels or accuracy, so measuring them costs nothing beyond the retention probe itself (`probe_sink.py`, 6 items per cell).

The prediction holds, on the presses where it can be tested. A floor is an accuracy at zero evidence retention, so only the 7 of E4's 10 presses that ever reach $\rho_m = 0$ have one to measure; the other 3 (PYRAMIDKV, SNAPKV, TOVA) never do, and are not counterexamples but absentees. Among the 7, the 2 that keep at least half of both blocks (budget-averaged; the threshold is descriptive, and the sweep below is what tests the mechanism) are StreamingLLM ($\sigma = 1.00$, $\omega = 1.00$) and H_2O (0.53, 1.00), and they are exactly the ones whose zero-retention accuracy is non-zero, at 0.45 and

0.46. The 5 that miss at least one block, LOCAL (0.00, 1.00), RANDOM+OBS (0.04, 1.00), H_2O-acc (0.62, 0.01), KeyNorm (0.66, 0.47) and RANDOM (0.04, 0.04), all score 0.000 to three decimals. Under P_{early}, where the retained window is the end of the context and not the query, every floor is at most 0.003 including StreamingLLM's, whose two retentions are unchanged. Keeping the window is not enough and keeping the sinks is not enough; the readout needs both, and it needs the window to have seen the query. What the sweep below cannot vary is the window: KeyNorm keeps 0.66 of the sinks and about half the window and has no floor, so the gate is on a near-complete window with at least one sink, not on the product κ as such.

A sweep settles the shape of the dependence, which the two presses of E4 cannot. Write $\kappa_m(B) = \sigma_m(B)\,\omega_m(B)$ under P_{late} and 0 under P_{early}, the product of the two block retentions the probe returns. The sink-sweep cell repeats the E4 design at 8K (120 items at 8,079 tokens, the leak-free cell of Appendix Q, P_{late} only) with 5 window templates that differ only in how many sink tokens they keep ($s \in \{1, 4, 16, 64\}$ against LOCAL's $s = 0$), so every one of them has $\rho = 0$ on the open middle by Proposition 3(b) and they contribute zero-retention rows at different σ. Across 7 (press, budget) points at $\kappa = 0$ the floor is 0.000 to three decimals and across the 24 at $\kappa > 0$ it is 0.26–0.48 (1 of those 24 is a content-driven press carried through the same sweep, at 0.45). The two ways of reaching $\kappa = 0$ both occur and both give zero: 6 points keep the observation window with no sink block at all (LOCAL at every budget), and 1 keeps a 64-token sink block at $B = 64$, which leaves no window. The second arm is a single (press, budget) point, so it is one observation consistent with the mechanism rather than a second dissociation with its own weight; the causal control below is what carries the claim.

What the sweep also shows is that the dependence is a gate rather than a scaling. A press keeping 0.25 of the sink block (one position of four) already gets the full floor, and among the presses that clear the gate the floor varies with the budget (from 0.48 at $B = 64$ down to 0.26 at $B = 2048$) rather than with κ. Fitting a scaled variant of Assumption 1's floor, $\beta\kappa_m(B)$, accordingly does not help on this cell, and we do not adopt it; the regression of the floor on κ over E4's (press, budget) points with a measured floor has slope 0.44 at $r = 0.93$ (15 dependent points, so we quote no p-value), but that linear appearance is an artefact of κ being near-binary across those presses. The claim we make is the gate and the dissociation, both of which are read from the probe alone.

N COMPUTE

The nine primary cells, cells M, S and H5 (Appendices P, R and S), the single-needle control, the short-context control and the RoPE ablation ran on one 48 GB GPU in bfloat16; the later controls named below ran on a 24 GB GPU, and the natural-document and RULER-verbatim cells were run on whichever of the two was available, which is why no cross-cell level is compared between them and a primary cell except with the caveat below. Per item, with K depths, ten presses, six budgets and two protocols, scoring costs one uncompressed prefill per protocol plus $K \times (60 + 1)$ short teacher-forced forwards on pruned caches, batched by press; wall time was 34–79 s per item for the 7B/8B models between 8K and 32K and about twice that for the 14B model, for roughly 12 GPU-hours for the nine primary cells; the controls, the decoded natural-document and RULER-verbatim cells and the transfer cells roughly double that. Peak memory was 18.9 GB for Qwen2.5-7B at 16K, 21.0 GB at 32K with 21 depths and 46.8 GB for Qwen2.5-14B at 16K. The benchmark audit and all analyses run on CPU in minutes. The cells of Appendix W ran on a third machine, a 24 GB RTX 4090, in about 11 GPU-hours, and the free-decoding, protected Expected Attention and LongBench runs (Appendices K.1, W.2 and W.4) on a fourth, with one 48 GB GPU, in about 21 hours of wall time, two runs sharing the GPU for part of it. To fit the 8B cell at 16K and the 32K cell there, the runner scores the P_{late} full-KV reference on a view of the extended cache instead of a truncated copy (`--low-mem-ref`) and now builds one pruned stack at a time; on CPU the new and the original paths write identical records. The later controls (Appendices M, Q, X.8 and X.9 and the Llama natural cell) ran on a 24 GB GPU, which is why the flat-press cells are Llama-3.1-8B at 8K and Qwen2.5-7B at 16K rather than Llama at 16K: that cell's cache did not fit that GPU with the runner used.

What is reproducible, and what is not. Because the later controls ran on different hardware, we measured what that changes rather than assuming it changes nothing (`analyze_repro.py`). We ran three arms, all on items of H1. Two runs of the same command on the same machine with the same prefill chunk agree exactly: all 132 of 132 scored answers are identical, so the pipeline

itself is deterministic. Changing the prefill chunk from 256 to 512 on that machine leaves 0 of 132 log-probabilities identical, but the effect on accuracy is unbiased (86 of 132 scored answers are unchanged, the mean shift is 0.0005 and the largest 0.14) so chunking is numerical noise rather than a direction. Re-measuring the released cell's own full-KV reference, on the same items with the same code, tokens and chunk but on the 24 GB GPU of the later controls, is different in kind: 0 of 1,100 log-probabilities agree, 452 of 1,100 scored answers are unchanged, and the mean token accuracy moves by 0.039 in the *same direction* at every one of the 22 (protocol, depth) means, by up to 0.083. No press is involved in that arm, so the whole difference is the forward pass.

The consequence is a constraint on how these numbers may be combined, and every contrast, band, ranking and leave-one-press-out R^2 this paper reports satisfies it: each cell is one run on one machine. The cross-cell quantities (the transfer matrix of Table 7, the natural-document transfer of Appendix K and the re-weighting test of Appendix X.3) compare levels across runs, and where those runs sit on different machines part of the gap may be hardware; the transfer matrix's five cells all ran on the 48 GB card, the natural-document transfer may not have. What a replicator on other hardware should expect to reproduce is the *contrasts*, not necessarily the levels; three cells re-run with every press on a third machine reproduce the contrasts, and there the levels as well (Appendix W.5), while on the machine where the levels moved no press was re-run, so contrast agreement there is untested; a fourth machine whose records agree as poorly, though its levels hold, reproduces the contrasts too. It is also why the flat-press cells of Appendix X.9 were measured self-contained, with all twelve presses in one run, instead of adding two presses to a released cell: the shift above is the size of some of the contrasts that comparison would have rested on. The runner recomputes the full-KV reference, which made this check free.

O Extended related work

Position without compression. Hsieh et al. (2024b) and Wang et al. (2025) attribute the middle failure to positional attention bias and Veseli et al. (2025) to how full the window is, which is why we hold T fixed; Zhang et al. (2026a) find no reasoning benchmark controlling position, filler and length jointly, and audit benchmarks for whether they control position at all — the audit of Section 6 instead locates the gold evidence in each item, which is what yields a distribution $\hat{D}$ rather than a yes/no. Zhang et al. (2026b) report that multi-hop accuracy is bounded by its weakest hop rather than by an average over hops, which is W_m arrived at from another direction.

Compression evaluation. The published methods (Xiao et al., 2024; Zhang et al., 2023; Li et al., 2024; Cai et al., 2024; Oren et al., 2024; Devoto et al., 2024; Ge et al., 2024; Feng et al., 2025) and their benchmarks (Yuan et al., 2024; Li et al., 2025b; Bai et al., 2024; Agrawal & Mayer, 2026; Liu et al., 2025) report aggregates. The closest published crossings are qualitative: passkey heatmaps over length and depth (Yuan et al., 2024; Devoto et al., 2025), and H_2O's appendix experiment moving the gold document out of StreamingLLM's window (Zhang et al., 2023). Ananthanarayanan et al. (2026) predict a compression cliff from a global eviction ratio, which is the budget-only predictor of Fig. 11b, and Zhang et al. (2026c) name the query-domain mismatch that Proposition 3(a) formalises under a uniform-attention null, and their Appendix F derives the exposure bias of accumulated attention under the same uniform baseline, whose closed form our measurements reject (Appendix X.1); Chen et al. (2024) counter the same bias by mixing random eviction into accumulated-attention eviction, and Liu et al. (2023) rank by accumulated importance as H_2O does. Gould & van den Hengel (2026) score eviction under a counter-causal mask, which is the asymmetry of Proposition 3(a) engineered away rather than measured. Sparse attention (Nawrot et al., 2026; Ren et al., 2026) has been studied without a position axis.

The window leak. Wu et al. (2025); Tang et al. (2025); Xiao et al. (2025) on retrieval heads and Gu et al. (2025); Sun et al. (2024) on attention sinks are the mechanisms behind the profiles. For Section 4.4 the relevant distinction is Fesser et al.'s (2026), between a sink that absorbs attention and one that broadcasts aggregated context: only the second can carry answer information, and the floor we measure appears only for the presses that keep it. Niu et al. (2025) supply the null (a model raises the logits of any token it has seen) which the leak-free control of Section 4.4 rules out, since that control leaves the prompt unchanged and alters only which positions the retained tail attended to.

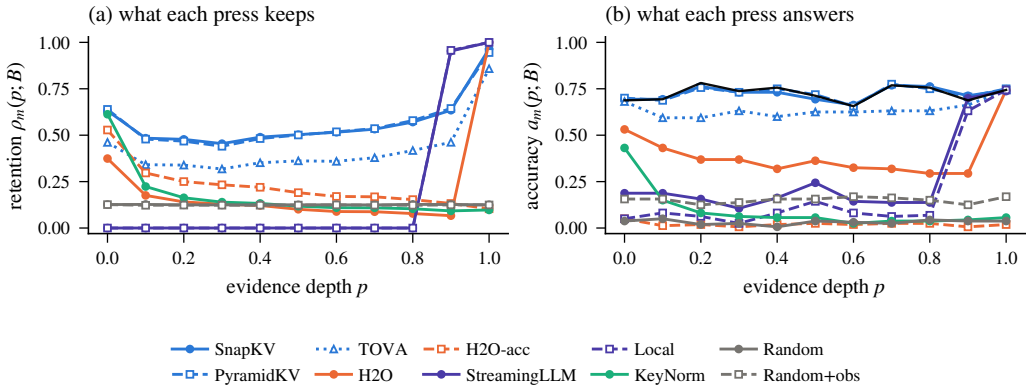

Figure 14: Fig. 1 rebuilt from natural documents: Qwen2.5-7B on 16,335 tokens (median) of SQuAD paragraphs, the 160-item cell, $K = 11$ of them gold at controlled depths, each with its own question, scored by the benchmark's contains-match on greedily decoded text. $B = 2048$, $\mathrm{P}_{\mathrm{late}}$. (a) retention, (b) accuracy; the black line is the uncompressed model. The two panels have broadly the same shape, as on the synthetic grid; at zero retention StreamingLLM sits above LOCAL, the floor of Section 4.4, and RANDOM keeps B/T of the evidence at every depth yet answers near zero, with $\mathrm{H_2O}$-acc, because keeping an eighth of a span is not keeping it.

P CELL M: A DIFFICULTY AT WHICH EXACT MATCH HAS RESOLUTION

The graded metric of Section 3 is the fraction of answer tokens predicted correctly under teacher forcing; exact match is that measurement thresholded. Across the nine primary cells the uncompressed exact-match rate under $\mathrm{P}_{\mathrm{late}}$ runs from 0.19 to 1.00, with 4 at or above 0.90 and 2 at or below 0.25, so the obvious reading of Section 4.2's "at most 4" (per cell: H1 0, H2 0–1, H3 0–2, H4 0, E1 0–2, E2 0–4, E3 0–1, E4 0–2, E5 0–2) is that exact match certifies little because the cells sit at one end or the other. Cell M tests that reading. It fixes only the answer format (two four-digit groups rather than one five-digit value or three groups, 9 tokens for Qwen2.5-7B) and holds the model, context length, haystack, depth grid, budgets, protocols and item count of H1. Uncompressed exact match is 0.53 and the graded metric 0.93, so both have headroom in both directions; the predictable hyphen floors the graded metric at 0.11 for this format.

The reading is wrong. At 0.53 exact match still certifies 0–2 of the 15 published pairs under $\mathrm{P}_{\mathrm{late}}$ and 0–2 under $\mathrm{P}_{\mathrm{early}}$, against 1–7 and 2–5 for the graded metric on the same records. What separates the two is resolution per item, not the level of the mean. The median RII width halves, from 0.73 to 0.38 under $\mathrm{P}_{\mathrm{late}}$, and the mean largest $|\delta|$ over published pairs from 0.58 to 0.33: the effect is smaller under the coarse metric, not absent, and it is the width of the corrected band that removes the certificates. The summaries that need no test move least under $\mathrm{P}_{\mathrm{late}}$, the depth-contingent share of (published pair, budget) observations from 46% to 40%, though under $\mathrm{P}_{\mathrm{early}}$ it falls from 58% to 21%, below the 25% null rate of Appendix X.4, and the ordering test from 0.67 to 0.68 and 0.48 to 0.55.

The mediation fit is the one quantity that reverses: R^2_ρ is 0.31 under exact match against 0.72 under the graded metric, and falls below $R^2_{\bar\rho}$, which is the failure Appendix X.1 reports for H1 and for the same reason — an isotonic link fitted to a thresholded surface is fitting mostly the threshold. Under $\mathrm{P}_{\mathrm{early}}$, where the presses separate further, ρ leads $\bar\rho$ under both metrics (0.54 against 0.13 under exact match, 0.81 against 0.25 graded).

Q THE LEAK-FREE CONTROL

The gate says which presses have a floor; it does not show that the floor is the window's trace of the evidence rather than something else about those presses. The control intervenes on exactly that. For one (item, press, budget) under $\mathrm{P}_{\mathrm{late}}$ we score two arms. The *normal* arm is the standard path: prefill the whole prompt, let the press choose its keep set S, prune, score. The *leak-free* arm splits S at the

start of the retained tail into S_{head} and S_{tail}, prunes the cache to S_{head} alone, re-runs the tail block of the prompt through *that* cache so those positions can no longer have attended to anything the press evicted, and takes S_{tail} from the recomputed states. Everything else (items, questions, gold tokens, keep sets, budgets, and the absolute RoPE positions every retained token carries) is identical, and the arms are paired within an item. With a budget large enough that nothing is evicted the two arms agree bit for bit, which is asserted in the released verification suite.

Two readings are needed, because recomputing the tail also removes its access to everything else the press evicted, not only to the evidence. At depths where the press retains no evidence the floor falls from 0.43–0.45 to 0.03–0.08 (a residual, not zero: the recomputed tail still holds the sinks and the question). At depths where it retains the evidence the same recomputation changes the mean accuracy by less than 0.0005 in absolute value (0.000 to three decimals), so the drop is not a generic cost of recomputing. The difference in differences is 0.37–0.40 across the 4 templates that have a floor at all, on 120 items of cell E4's design (Llama-3.1-8B, task E) at 8,079 tokens, 26,400 scored answers. LOCAL, which has no floor to remove, shows none. A second cell says what the intervention does when the floor is not a window leak. On Qwen2.5-7B's task H at 16,087 tokens (60 items) every press has a zero-retention floor of 0.12–0.22, LOCAL included at 0.12, and the intervention changes it by -0.01 to 0.03. That is the right answer: at zero retention this model emits 2.11 of the 14 answer tokens, and the modal value is exactly two, the hyphens of `dddd-dddd-dddd` that Section 3 floors the graded metric at 0.14 for. There is no sink-gated readout to remove, no dissociation between StreamingLLM and LOCAL, and the control correctly finds nothing. The intervention removes the floor where the floor is the window's trace and leaves it where it is a property of the answer format. This also rules out the reading on which the effect is contextual entrainment (Niu et al., 2025) (the model raising the logits of any token it has seen) since the prompt is unchanged between the arms and only the provenance of the retained tail's keys and values differs. The release carries 4 leak-free runs on this design of which 2 are quoted here, besides the ten-press cell of Appendix X.8; the other 2 are earlier pilots, one of which needs a note because it looks like a contradiction. It is a 20-item run at 15,908 tokens with a fixed 2048-token window, and StreamingLLM's zero-retention accuracy in it is 0.00 where cell E4 reports 0.45. The two are not the same quantity: that pilot builds its own haystacks and shares no item with E4, and its window is a constant rather than the press's own budget-dependent tail, so its zero-retention rows are not E4's. No number in this paper is taken from it.

R CELL S: AN ANSWER THAT IS NOT IN THE CONTEXT

Every other cell is extractive: the gold string appears in the context, so a press that retains the evidence lets the model copy it, and retention and copyability are confounded. Cell S separates them. The needle reads "The two special magic numbers for `<key>` are: a and b" with a, b drawn uniformly from $[1000, 4999]$, and the question asks for $a + b$; the gold string is the sum, which occurs nowhere in the context. Everything else is H1's design (Qwen2.5-7B, 16,397 tokens, 11 depths on the same even grid, distractor haystack, ten presses at six matched budgets, both protocols, 40 items, 53,680 scored answers), and the gold is 4 tokens, among the shortest of any cell (E4–E5 have two), which by Section 4.2 is part of why it certifies few pairs (0–2 under P_{late}, tied with E3–E5); a sum that collides with a value elsewhere in the haystack is not rejected, and at four digits such a collision is possible, which is a second reason to read the cell as a check of direction rather than of size.

The task is harder than the extractive ones, at 0.71 uncompressed against 0.93 for cell M, because the model must also add. That is a cost, not a confound: the arithmetic is the same at every depth, so it moves $\alpha(p)$ down uniformly and by Proposition 2 leaves the sign of $\delta(p)$ (a difference between two presses on the same items) unaffected while shrinking its size. The framework's claims carry over. 52% of the (published pair, budget) observations are depth-contingent under P_{late} and 60% under P_{early}, inside the 17–67% range of the extractive cells (a description, given the null rate of Appendix X.4); the prior family gives 14–18 distinct rankings of 18; and the mediation bake-off still favours depth-resolved retention (0.43 against 0.31 under P_{late}, 0.68 against 0.23 under P_{early}). Retention buys access to the evidence, and the framework does not require that access be a copyable span.

S CELL H5: A THIRD MODEL FAMILY

The transfer claim of Section 4.3 rested on the two tokenizer families of Section 3. Cell H5 adds Mistral-7B-Instruct (Jiang et al., 2023) on H1's task at 8,187 tokens (50 items, 67,100 scored answers, 15 answer tokens), which is a third tokenizer and a length task H had not used; the checkpoint is Mistral-7B-Instruct-v0.3, which has no sliding-window attention. Its own engine check is worth stating: no prompt layout the engine had split at a token boundary under this tokenizer, because SentencePiece attaches a word-boundary marker to the first token of a segment, and a boundary that is not exact would silently change the shared context cache from question to question. The released engine walks the split backwards to the nearest safe position, which keeps the whole question out of the compressed prefix; the verification suite asserts the resulting split is exact.

The ranking results reproduce: 4–9 of the 15 published pairs are family-wise sign-changing under P_{late} and 0–9 under $\mathrm{P}_{\text{early}}$, the prior family gives 9–15 distinct rankings, and 60% of (published pair, budget) observations are depth-contingent, near the top of the 17–67% range of the other cells. Under P_{late} the mediation results are among the weakest of the task-H cells (R^2_ρ 0.39 against 0.35 for $\bar{\rho}$ under P_{late}, 0.53 against 0.25 under $\mathrm{P}_{\text{early}}$), and the reason is visible in the uncompressed row: at 8,187 tokens this model answers 0.999 of the task correctly with no compression at all, so $\alpha(p)$ is flat at the ceiling and Assumption 1 leaves g almost nothing to fit. That saturation is why the transfer matrix is asymmetric: the link fitted on H5 predicts the four unsaturated cells at 0.64–0.82 under P_{late} (Table 7 gives $\mathrm{P}_{\text{early}}$, where it is weaker), while links fitted on them predict H5 at 0.36–0.48, below H5's own in-sample 0.59. It is not the depth-robust regime of Proposition 1(i): with 4–9 certified pairs the cell reverses as much as any.

T THE PROTOCOL AS A SECOND COORDINATE

Whether the press sees the query before choosing what to keep is rarely stated and changes the ranking. Luo et al. (2026) report this on RULER at matched budgets: moving to the query-agnostic protocol re-orders the published presses, some falling behind a window template. The controlled grid reproduces the re-ordering but not that last detail. Kendall τ between the uniform-prior rankings under the two protocols is 0.20–0.82 across H1–H4 (Table 5), and in H1 the winner moves from TOVA or SnapKV to $\mathrm{H_2O}$-acc or $\mathrm{H_2O}$ at $B \leq 512$ (Table 10; Fig. 12 shows $B = 256$), while StreamingLLM stays mid-table. The grid also shows where the re-ordering comes from. Under P_{late} the observation window a query-anchored press scores against contains the query, so the positions it rates highly are those the query attends to; under $\mathrm{P}_{\text{early}}$ that window is the end of the context, and the same rule rates a different set. The protocol therefore acts on $\rho_m(\cdot; B)$, and every consequence of a change in ρ in Section 2 follows: the profiles move, so by Proposition 2 the contrasts move, and Proposition 1 then applies to the new contrasts within the new protocol. Protocol and prior are two coordinates of the same object, and a comparison that states neither is identified by neither.

U WHAT THE AUDIT CAN AND CANNOT LOCATE

The procedure records the first character offset at which the answer string occurs, which identifies the evidence only when a single span carries the answer. For 3 of the 15 tasks it does not. Frequent-word and common-word extraction require counting occurrences across the whole context, so no single position is the evidence and the first occurrence of the first answer word necessarily moves to a smaller relative offset as the context grows — which is what the apparent 4K-to-16K drift in those rows of Table 6 is. Variable tracking follows an assignment chain whose whole extent is the evidence. Those three are shown in Fig. 2a, because a reader should see what the procedure returns for them, and are excluded from every re-weighting claim in Section 6. For the remaining 12 tasks the answer is a span and the offset is its position; the `qa_1` and `qa_2` rows are the two we then measure directly (Section 5), and they are also the two where a string match can be checked against the released gold answer (the 99.9% location rate itself is over all 20,081 rows, Table 6). We did not hand-validate a sample, and a stricter audit would use RULER's generator metadata rather than a string match. The shape of what it returns is nonetheless informative about which of the two is moving. A matcher that fails at random, or that drifts as contexts lengthen, gives a smooth distribution that shifts; a generator that inserts one gold passage among a bounded number of distractors gives a small set of attainable slots. What the audit returns is the second: over the 41 (task, configuration) cells the largest decile

band holds up to 100% of a cell's items, and QA-1 occupies 6 of ten bands at 8192 (61% of its items in the largest) and 8 at 16384 with maxima 0.73 and 0.77, so no item there reaches the last fifth of the context at all. Slotting is evidence about the generator only if the match lands in the gold passage; a match in a distractor passage also lands on a slot, and we did not check which, so this remains an inference. The same statistics say how often these benchmarks test the region a window template retains entirely: the median cell places 17% of its items above depth 0.8, and 10 of the 41 place none; 0.8 corresponds to a window of a fifth of the context, more generous than any budget of the primary cells (though not than the largest budgets of H5 and of the 4K rows), so the share inside a real window is smaller still. Measured directly on the 996 RULER-verbatim rows at 16K of Section 5, with each row's own query length, 38 rows put their evidence inside StreamingLLM's retained tail at the largest budget and 25 at $B = 512$.

There is a third thing the procedure records that needs saying plainly: when a task has several gold answers, the first offset is the smallest of several positions, and the minimum of several draws is biased toward the front relative to the spans' mean depth. 3 of the 12 localisable tasks are in that position. We re-ran the audit on a sample of 44–80 items from each of 7 tasks with every located span recorded rather than only the earliest (`analyze_audit_spans.py`; the sample's first-offset means differ from the full-population means of Table 6 for that reason): for the 3 multi-answer localisable tasks (MQ-NIAH, MV-NIAH, QA-1) the mean depth of the gold spans exceeds the first offset by up to 0.32, while for every single-span task the two agree to 0.000. MQ-NIAH moves from 0.24 to 0.56 and MV-NIAH from 0.21 to 0.53, which puts their evidence in the middle of the context rather than at its front — and it is where VT, which we exclude on the same grounds stated differently, also lands (0.24 to 0.55). Those two tasks set the lower end of the 0.20–0.53 range quoted in the abstract and Section 6, so that end is a property of the summary and not only of the benchmark. Restricted to the 9 localisable tasks not measured to carry several spans the range is 0.25–0.53. We report the first-offset range because it is the statistic computed over all 20,081 rows, and we say here what it is: a first-occurrence depth. The re-weightings of Section 6 use each task's whole first-offset histogram, which for those three tasks inherits the same shift, so their rows in Fig. 2b and Table 10 are re-weightings by a front-biased prior; HotpotQA-distractor and `qa_2` are two-hop and use one supporting sentence each, so their priors carry the same caveat.

A sample is the obvious weakness of that check, so the audit was repeated on every row of `qa_1`, `qa_2`, MQ-NIAH and MV-NIAH at all three releases (`analyze_audit_allspan.py`), with the gold passage of each QA row identified from its source dataset. The multi-answer shift holds on the full data, and the swing of Section 6 does not depend on the first-occurrence rule: QA-1 moves by 0.44 across releases on first occurrences, 0.43 on the mean over every occurrence of every answer, and 0.55 at the gold passage, whose positions sit in narrow slots per release, which is the generator slotting inferred above. The first occurrence misses the gold passage in 30–44% of `qa_1` rows, landing in an earlier distractor, which is the matcher error the paragraph above could not check.

Several ranges appear in the text and they are not the same. Pooling every released length gives mean depths 0.008–0.53 across all 15 tasks, which is the spread Fig. 2a shows, and 0.20–0.53 across the 12 localisable ones, which the abstract quotes. The re-weightings use each RULER task's 16,384-token rows alone and the native rows of the other two, where the statistic runs 0.19–0.51 over the 12 tasks; per-length values are in Table 6.

V THE ROPE POSITION CONVENTION

StreamingLLM as published assigns positions *within* the cache; the KV-compression setting this paper measures keeps each retained key's original rotary phase, and the cache index is local. The two differ, and since StreamingLLM carries Fig. 1 and Appendix J the difference should be measured rather than argued. `Engine.renumber_rope` converts one convention into the other exactly: the cached keys are already rotated by $R(p)$, so rotating further by $R(j - p)$ gives the key a prefill at position j would have produced, using the model's own `inv_freq` so Llama-3.1's frequency scaling stays correct. The released suite checks that the rotation preserves key norms and that at layer 0, whose keys depend only on the token embeddings, the renumbered keys match a genuine contiguous prefill of the retained tokens to 10^{-6} relative while the original phases differ by 0.47.

Repeating H1's design under both conventions on the same items and keep sets, renumbering costs every press accuracy, and it costs the window templates least: at most 0.05 for StreamingLLM

and LOCAL (mean −0.04 for StreamingLLM) against up to 0.43 for the content-driven presses (mean −0.19 for SnapKV). The convention this paper uses is therefore not the one that flatters StreamingLLM; renumbering would penalise the query-anchored presses several times more, and would widen the tail gaps and narrow the interior ones that Section 4.2 reports, in neither case removing a crossing. Of 36 (pair, budget, protocol) cells among the template and query-anchored presses the sign of the mean contrast changes in 2, both under P_{early} and both involving LOCAL, which is a null press rather than a published one; that LOCAL, a pure translation under RoPE, moves at all is a measure of the numerical noise of Appendix N. No conclusion in the paper turns on the choice, and a paper reporting StreamingLLM under the streaming convention would be measuring a different operator, whose profile Proposition 3(b) still describes.

W FOUR CONFOUNDS, ONE OUT-OF-FAMILY PRESS AND TWO MORE MACHINES

Three task-H cells (H1, H3 and H4, 306,520 scored answers) were run again on another GPU (an RTX 4090) with the released code, items, budgets and ten presses (the two runner changes of Appendix N write identical records), plus five presses that exist to test a specific objection: four question-protected versions of presses that can evict the question under P_{late}, and Expected Attention (Devoto et al., 2025). Each cell is analysed three ways with the unchanged analysis code: the ten presses (the cross-machine replication), the ten with the four protected versions substituted under P_{late} (under P_{early} there is no question in the prefix, so the two coincide), and the ten plus Expected Attention (Bonferroni over 55 pairs); `analyze_replication.py` does all three. A fourth machine, with a 48 GB GPU, later re-ran H1–H4 with the ten presses and the model's greedy output recorded (Appendices W.4 and W.5), re-ran Expected Attention beside a question-protected version (Appendix W.2), and ran LongBench's own rows (Appendix K.1).

W.1 EVICTING THE QUESTION DOES NOT DRIVE THE P_{LATE} REVERSALS

Under P_{late} the pressed prefix is the context followed by the question, so a press can evict the question itself, and a reversal involving that press could then measure which query survived rather than where the evidence sat. Every record of these runs logs ρ_q, the share of the question's positions a press kept. Five of the ten keep all of it at every budget, PyramidKV at least 0.96 (at its smallest per-layer budget) and TOVA at least 0.79; three do not: H_2O-acc keeps at most 0.04 of it, KeyNorm 0.11–0.84 and RANDOM at most 0.125 (its budget share). The protected version of a press keeps every question position and spends the rest of the *same* budget on the context by the press's own rule, exactly B per layer and head, which is the structure SnapKV's forced window already has (`build_keep_protected`; checked in suite group T17, Appendix B). We protect the four that lose the most of it, H_2O-acc, KeyNorm, TOVA and RANDOM, and leave PyramidKV, which keeps at least 0.96; of the four only KeyNorm and TOVA are published, so the published counts below test those two.

Protection leaves the published reversals where they were: summed over budgets, the certified published pairs under P_{late} move from 26 to 27 in H1, 12 to 11 in H3 and 38 to 37 in H4, by at most 1 at any budget. Over all 45 pairs the certified counts rise, from 172 to 197 summed over the three cells, mostly in pairs with H_2O-acc, whose protected version reverses against more presses, not fewer. What it changes most is the link. The two presses the shared link mispredicted worst (Table 8) are two of those three, and protecting the question lifts H_2O-acc's held-out R^2 from 0.11 to 0.79 in H1 and from −0.08 to 0.80 in H4 (KeyNorm's from 0.23 to 0.33 and 0.20 to 0.28), and the mean over presses from 0.47–0.77 to 0.55–0.85 across the three cells; in H3 H_2O-acc stays below zero (−0.86 to −0.21). Part of what Section 4.3 reported as a failure of retention to mediate was question eviction, which ρ_m, a statistic of the evidence span, does not see.

W.2 AN OUT-OF-FAMILY PRESS: EXPECTED ATTENTION

The ten presses come from six families, and the leave-one-family-out fit of Appendix W.3 still draws its test press from them. Expected Attention (Devoto et al., 2025) is from none: it scores each key by the attention a Gaussian model of *future* queries would pay it (the mean μ and covariance Σ of the prefix's pre-RoPE queries, sinks excluded, rotated by the RoPE matrix averaged over the next 512

positions) as $\mathrm{softmax}(\mu^\top k/\sqrt{d} + k^\top \Sigma k/(2d))$ with d the head dimension, averaged over the query heads of a KV group and scaled by $\|v\|$, and keeps the top B per head with four sinks. It reads no attention map and keeps no window; its only positional rule is the four sinks. The engine accumulates the query statistics in a hook during the same prefill; on the query-agnostic path its scores match NVIDIA `kvpress` 0.5.5's own implementation to 3.6×10^{-5} relative error with identical keep sets (`verify_expatt_kvpress.py`, on Qwen2.5-0.5B), and both paths match an independent re-implementation in suite group T18.

Its profile carries a positional prior the method does not state: in H1 at $B = 512$ under $\mathrm{P}_{\mathrm{early}}$ it keeps 0.23 of the evidence at $p = 0$, 0.03 at $p = 0.5$ and 0.49 at $p = 1$, a soft version of the window templates' interior hole (Proposition 3b), and it keeps 0.64–0.97 of the question under $\mathrm{P}_{\mathrm{late}}$. The link fitted on the ten presses measured on the same machine predicts its surface with no refitting at R^2 0.79–0.88 under $\mathrm{P}_{\mathrm{early}}$ across H1, H3, H4, the query-agnostic setting the press was designed for, against 0.24–0.45 for $\bar{\rho}$. Under $\mathrm{P}_{\mathrm{late}}$ it predicts it at 0.48–0.75 against 0.58–0.67, and $\bar{\rho}$ is the better predictor in 2 of the 3 cells: once the question is in the prefix, this press's accuracy follows its budget more closely than where the evidence sits. It enters 30 ($\mathrm{P}_{\mathrm{late}}$) and 40 ($\mathrm{P}_{\mathrm{early}}$) certified reversals with the six published presses over the three cells and six budgets, and ranks 4–9 of eleven by mean accuracy under $\mathrm{P}_{\mathrm{late}}$.

Does question eviction explain this? The fourth machine of Appendix W.5 ran it again under $\mathrm{P}_{\mathrm{late}}$ in H1, H3 and H4 (50, 50, 40 items) next to a protected version, `expatt_q`, that keeps every question position and spends the rest of the same budget by the press's own scores, as in Appendix W.1. The unprotected press keeps 0.83–0.90 of the question on average over its records and at least 0.59; protection changes its mean accuracy by at most 0.007 at any budget, the link fitted on the same machine's ten presses predicts the protected press at R^2 0.52–0.69 (0.44–0.69 unprotected), and $\bar{\rho}$ still does better in 2 of 3 cells (0.59–0.71). Question eviction is not why $\bar{\rho}$ wins under $\mathrm{P}_{\mathrm{late}}$. The link's error has one consistent part: at the first and last depths, where this press keeps the most evidence, the mean residual, measured minus predicted accuracy of the protected press over all budgets, is -0.26 to -0.10, so it answers less there than the ten presses do at the same retention (in the interior it is -0.04 to 0.10). A uniform average over heads (Definition 1) can hide which heads kept the span, and Appendix X.5 finds that this matters; we have not tested that it is the cause here.

W.3 The link without near-twins, and against a press's own budget curve

Leave-one-press-out has two weaknesses (`analyze_mediation_robust.py`). The held-out press usually has a near-twin in the fit (StreamingLLM/LOCAL, SnapKV/PyramidKV, the two random presses, the two $\mathrm{H_2O}$ variants); holding out the whole family instead lowers the mean R^2 by at most 0.04 under $\mathrm{P}_{\mathrm{late}}$ (0.44–0.75 across H1–H4) and 0.02 under $\mathrm{P}_{\mathrm{early}}$. And the budget-only baseline pools nine presses, so it cannot tell presses apart; the curve a paper actually reports is the held-out press's *own* accuracy at each budget, averaged over depth. That curve knows the press's level exactly, which no predictor from retention can, and reaches R^2 0.69–0.76 under $\mathrm{P}_{\mathrm{late}}$. Adding to it the depth shape of the link's prediction (the prediction minus its own depth mean) raises that to 0.80–0.92, better in 4 of 4 cells (0.77 or more when the family is held out); under $\mathrm{P}_{\mathrm{early}}$ it is better in 3 of 4, and worse in H3. On depth-centred values alone, where a depth-averaged predictor scores exactly zero, the pooled R^2 is 0.70 in H1 and 0.76 in H4 but 0.24 in H2 and -0.22 in H3: the link predicts the depth shape where accuracy varies strongly with depth, and adds little beyond the level where it does not.

W.4 Answer tokens teacher forcing cannot help

Token accuracy under teacher forcing supplies the gold prefix, and a supplied prefix could point the model at the rest of a value it could not have found. Two stricter metrics limit that help: whether the first answer token is right, predicted from the question with no gold token supplied, and the fraction right within the first four-token digit group, where a supplied digit can still help inside the group but no later group can. Both read the gold token ids off the answer string, one token per digit, which holds for Qwen, so they are computed on H1, H2 and H4. Summed over those cells and six budgets, the first token certifies 87 published-pair reversals under $\mathrm{P}_{\mathrm{early}}$ and the first group 108, against 86 for the graded metric; under $\mathrm{P}_{\mathrm{late}}$ they certify 14 and 20 against 84. The $\mathrm{P}_{\mathrm{early}}$ count is about the same without teacher forcing (66 of the 86 graded certifications are also first-token certifications); under $\mathrm{P}_{\mathrm{late}}$ most certifications need answer tokens after the first group, which these two metrics do not score,

and the direct test below says whether they need the supplied prefix. The link mostly survives the change of metric: leave-one-press-out R^2 is -0.13–0.64 under the first-token metric and 0.50–0.66 under the first group under P_{late} (the negative value is H2), and 0.31–0.71 and 0.45–0.85 under P_{early}.

The model's own output. The direct test scores what the model writes. H1–H4 were run again on a fourth machine (Appendix W.5) with the released code, items, budgets and ten presses, every record now also carrying the model's greedy continuation of the prompt (16 new tokens, 24 for Llama, whose answers more often open with a sentence). The free score of an output is one minus the edit distance between its digits and the gold digits, over twelve, floored at zero, where its digits are those of its first `dddd-dddd-dddd` group or, failing one, of its first run of digits and hyphens (`free_decode_metrics.py`). No gold token is supplied at any point, and the score does not depend on the tokenizer, so Llama's cell is included. An output with no digits scores zero and a random twelve-digit string 0.15 on average; uncompressed, the free score is 0.91–1.00 against 0.93–1.00 teacher-forced on the same records, and at most 2.5% of a cell's outputs are cut off inside the value by the token limit. With the analysis code unchanged, the free score certifies 98 published-pair reversals under P_{late} and 135 under P_{early}, summed over cells and budgets, against 96 and 122 for the teacher-forced metric on the same records, and 78 and 97 of them are the same (cell, budget, pair) under both. Over all 45 pairs the counts move more, 116 against 206 under P_{late} and 365 against 319 under P_{early}, and what moves them is the floor: where a press keeps none of the evidence the teacher-forced score sits at 0.01–0.26 across presses and protocols and the free score at 0.00–0.32, depending on whether the model guesses digits or says it cannot find the number, and pairs of diagnostic presses, which sit at those floors over most depths, gain or lose certifications with them. Exact match of the free output certifies 26 and 13 published reversals, the loss of resolution of Appendix P once more. The link, refitted on the free score, fits less well than on the teacher-forced score of the same records, as a press-dependent floor, which breaks the shared β of Assumption 1, predicts: leave-one-press-out R^2 is 0.29–0.67 under P_{late} (teacher-forced 0.47–0.74; $\bar{\rho}$ on the free score 0.23–0.53, ahead only in H2), and -0.49 to 0.69 under P_{early} ($\bar{\rho}$ -0.40 to 0.27), where $\bar{\rho}$ does better in H2 and H3: there 47–61% of all outputs contain no digit, most of them the model saying it cannot find the number, so the free floor is zero for some presses and not for others. What the first-token metric lost under P_{late} was resolution (it scores an answer 0 or 1) not the reversals.

W.5 Other machines

Appendix N found that levels need not transfer across hardware and said a replicator should *expect* contrasts to; these runs measure it. On the RTX 4090 every keep set is identical to the released one (largest $|\Delta\rho|$ 0.000), 68–96% of scored answers are identical, and (press, budget, depth, protocol) means correlate at 0.999 or more with a mean shift of at most 0.001 (largest single cell 0.04). The contrasts $\delta(p)$ correlate at 0.999 or more; of the pairs certified in one run at least 95% are certified in the other (94% of published pairs), and leave-one-press-out R^2 moves by at most 0.00. On this machine the levels transfer too; on the 24 GB GPU of the later controls (Appendix N) they did not, which is why each cell remains one run on one machine.

The fourth machine, with a 48 GB GPU, ran H1–H4 again for the free-decoding test of Appendix W.4, the ten presses on the released items with the released code. How closely its records match the released ones depends on the cell: 41–91% of teacher-forced scores are identical (H4 lowest, H3 highest), the low end as poor as on the 24 GB GPU of the later controls (Appendix N; 41% there), and the uncompressed reference moves by up to 0.047 per protocol, as it did there. The keep sets of the score-driven presses differ (largest $|\Delta\rho|$ 0.24), while StreamingLLM, LOCAL and the two random presses keep identical sets. The compressed cells' levels hardly move: (press, budget, depth, protocol) means correlate at 0.997 or more, with a mean shift of at most 0.008. The contrasts transfer: $\delta(p)$ correlates at 0.996 or more, at least 91% of the pairs certified in one run are certified in the other (89% of published pairs), and leave-one-press-out R^2 moves by at most 0.03. Record-level disagreement of the size that accompanied the level shift of Appendix N, with keep sets that differ too, therefore need not move a contrast.

The natural-document cell of Section 5 was also run on 160 items it had not used (seeds 160–319, 235,840 scored generations, the same design, Expected Attention added) and analysed on its own, over the same ten presses. Under contains-match these items certify 2 published pairs under P_{late}, at $B = 2048$ and at no smaller budget, as the released items do, and they are the same two pairs,

(H_2O, StreamingLLM) and (KeyNorm, StreamingLLM), one of them at the band's edge in the fresh run. Under P_{early} they certify 6 (7 on the released items, 6 of them shared at $B = 2048$; 8 against 8 summed over budgets, though the one or two pairs certified at $B = 1024$ differ between the runs, all within 0.03 of the band's edge). The uncompressed model scores 0.74 (0.73 on the released items), and the held-out R^2 of the link is 0.61 under P_{late} (0.58) and 0.54 under P_{early} (0.59). The natural-document result is therefore not a property of the first 160 items drawn.

X CONTROLS AND NEGATIVE RESULTS

Every entry here is a check the paper ran against a claim it wanted to make, and Table 3 indexes them against those claims. The 18 with negative results are here too.

X.1 WHAT DID NOT HOLD

Six things did not hold as predicted. The closed form $c_i \propto \ln(T/i)$ of Proposition 3(a) fits the accumulated attention poorly ($R^2 = 0.005$); c_i falls with position in every layer (Fig. 7), and the paper uses only the monotone statement. The crossing-predicts-flip prediction misses by more than a grid step in five of the nine cells, worst in H4 (Spearman 0.17, median error 0.350) and H2, and the paper claims the location only in the four cells where the median error is under a grid step (Appendix H). The equivalence-to-shortening test is decided only for StreamingLLM; for the query-anchored presses the compressed profile is within L_1 shape distance 0.02–0.12 of some genuine short context and we do not claim a distinction. On task E, Llama-3.1-8B under P_{late} is saturated and has few certified reversals (0–2 certified published pairs, with the leaders tied at ceiling); this is the nearest the grid comes to the regime of Proposition 1(i) and is reported as such in Section 4.4. The shared floor of Assumption 1 fails under P_{late} on task E for Llama-3.1-8B, and more weakly on task H (H3, below): with $\rho = 0$, StreamingLLM scores 0.45 (E4) and 0.45 (E5) of the two answer tokens while LOCAL scores 0.00 and StreamingLLM under P_{early} scores 0.00; the retained query window, prefilled with the evidence in view, carries the first answer token, and the sink tokens gate its readout. The held-out R^2 for StreamingLLM in E4 is -3.5 and for H_2O -0.33, against 0.77 for SnapKV and 0.75 for RANDOM, which is what drives the cell means of Table 1. On task H the zero-retention scores are 0.21 (StreamingLLM) against 0.03 (LOCAL) for Llama and 0.15 against 0.13 for Qwen, the latter at the hyphen floor of 0.14 (Llama does not place its hyphens without the evidence, Appendix C).

Where $\bar{\rho}$ beats ρ. The gap $R^2_\rho - R^2_{\bar{\rho}}$ is 0.18 on task E under the graded metric, 0.05 on the same cell under exact match and 0.11 on the 160 natural-document items under the benchmark's contains-match. Under P_{late} it is positive in every task-H cell in the mean, but not press by press: in H1 under P_{late} it is $+1.02$ for H_2O, $+0.67$ for StreamingLLM and LOCAL, within 0.04 for SnapKV, PyramidKV, TOVA and the two random presses, and -0.43 for KeyNorm, one of the two presses the shared link mispredicts worst (H_2O-acc is the other, Table 8). That pattern is what Proposition 2 implies: a press whose profile is nearly flat has $\rho \approx \bar{\rho}$, so depth resolution has nothing to add, and the gap is carried by the presses whose profiles vary with depth. Enumerated rather than described, $\bar{\rho}$ wins in 16 of the 48 (cell, metric, protocol) combinations reported anywhere in this paper, and 13 of those 16 are under exact match. That is the dominant explanation: an isotonic link fitted to a thresholded surface is fitting mostly the threshold, and in the worst of them R^2_ρ is strongly negative while $R^2_{\bar{\rho}}$ is not, which is what fitting noise looks like. The remaining 3 are under the graded metric, in cells E4, E5, H3, and every one of them is under P_{early}, where the query-anchored presses no longer see the query, their profiles flatten and ρ has less to add over $\bar{\rho}$. Depth-resolved retention is the better predictor in 32 of 48, not in all of them; Table 2 gives both columns for the graded metric in every cell, Table 9 and Appendix P the exact-match cases that are tabulated, and the released macros the rest.

Separately, exact match on task H for Qwen2.5-7B is near floor for every press including the uncompressed model (Table 9; the other task-H cells sit higher, Appendix P), which is why the graded metric is primary; on task E, where exact match has resolution, it gives the same qualitative picture with weaker mediation (R^2 0.51 against 0.64 for the graded metric under P_{late}). The H1 exact-match rows of Table 9 are the clearest case of a metric without resolution rather than a counterexample: with the uncompressed model at 0.19 and almost every press at zero, the surface is nearly constant and the depth-resolved predictor loses to the depth-averaged one (R^2_ρ 0.19 against $R^2_{\bar{\rho}}$ 0.49), which is

what a predictor with nothing to predict does. The benchmark-realistic metric with resolution is the contains-match of Section 5, on natural text rather than this content, and there ρ leads $\bar{\rho}$ again.

X.2 The distinct-ranking counts against an item-noise null

A count of distinct rankings has an obvious failure mode: two presses separated by less than the item-level noise swap order between two priors for free, so a cell can report many rankings without any prior changing an ordering that is actually resolved. The null has to match the observed statistic in sample size. The observed number is *distinct rankings over $|\mathcal{D}_K|$ priors*; the null is therefore *distinct rankings over $|\mathcal{D}_K|$ item-resamples at one fixed prior*, the uniform prior, repeated 400 times to get its distribution; a point-mass prior uses one depth's items and is noisier, so the null is conservative for the uniform member and generous for the point masses, which Appendix X.7 addresses by dropping them. Comparing the observed count against distinct rankings over 400 resamples instead (the naive version) compares a count over $|\mathcal{D}_K|$ draws with a count over 400 and is guaranteed to favour the null.

Over the 144 (cell, protocol, budget) combinations of all twelve cells, the priors give 4–28 distinct rankings and item noise at a fixed prior gives 1.0–15.2 on average; the observed count exceeds the noise-only 97.5th percentile in 139 of them. In H1 the observed counts are 15–18 against a noise-only mean of at most 6.7. The multiplicity is a property of the priors, not of the sample size. Nor is it made by the four diagnostic presses: restricted to the six published ones, the priors give 4–12 distinct orderings per (cell, budget) across H1–H4 under $\mathrm{P}_{\mathrm{late}}$ (4–23 under $\mathrm{P}_{\mathrm{early}}$), and first place changes hands among the six in 88% of those (cell, budget) combinations under $\mathrm{P}_{\mathrm{late}}$ (75% under $\mathrm{P}_{\mathrm{early}}$), between up to 4 presses.

The same null answers the objection that the count is carried by point masses and diagnostic presses. Over the seven members of $\mathcal{D}_K$ that are not point masses, the observed count exceeds the null's 97.5th percentile in 73 of the 108 (cell, protocol, budget) combinations of the nine primary cells (in many of the rest that percentile already equals the family size, seven, so no count could exceed it); restricted to the six published presses it does so in 107 of 108 over all of $\mathcal{D}_K$ and 85 of 108 over the seven shaped members.

X.3 Re-weighting against not re-weighting

Section 5 reports that it does not, and the comparison matters because a coarser version of it can look like a validation. The predicted ranking is H1's measured accuracies weighted by a task's empirical prior; the target is the ranking a run on that task's own rows would report, which weights each depth bin by how many of the task's items fall in it (the bins are far from equal (20, 10, 71 and 19 items in the four quarter-bands of the 120-row `qa_1` pilot at 16K) so uniform-over-bins is not the ranking anyone would publish). The measured rankings come from 120-row pilots of each task at each of the 4096- and 16384-token releases at $B \in \{128, 512, 2048\}$, twelve (task, length, budget) cells per protocol; the 4K prior is the one measured at 4K, since a sub-task's prior moves between releases. The baseline is the same H1 measurements under the uniform prior, that is, the prediction one would make knowing nothing about where the task puts its evidence.

Against $\tau = 0$ the re-weighted prediction looks good (0.38–0.73 under $\mathrm{P}_{\mathrm{late}}$). Against the baseline it is not better: the baseline reaches τ 0.66 under $\mathrm{P}_{\mathrm{late}}$ and re-weighting changes it by -0.09 (paired Wilcoxon over the twelve cells, which share items across budgets and so are not independent, $p = 0.09$; better in 1 of 12 cells, tied in several), and by $+0.04$ under $\mathrm{P}_{\mathrm{early}}$ ($p = 0.29$, 5 of 12). Top-1 agreement is 6 against 6 for the baseline. The reason is not mysterious: the synthetic grid and a RULER QA task differ in document source, question type and answer length as well as in evidence depth, and the prior carries only the last; for `qa_1` the prior is also a first-offset histogram (Appendix U). It is evidence against using the prior as a forecasting device, and no evidence against Proposition 1, which says a scalar under an unstated prior fails to determine an ordering.

The pilots are small, so the comparison was repeated at full size on the RULER-verbatim cells of Section 5, every located `qa_1` and `qa_2` row at 16K (996 of 1,000; $B \in \{128, 512, 2048\}$, six cells per protocol). Under $\mathrm{P}_{\mathrm{late}}$ re-weighting reaches a mean τ of 0.53 (0.47–0.64 over the six cells) against the baseline's 0.71, a change of -0.18 (better in 0 of 6; the six cells share rows across budgets, so the exact signed-rank $p = 0.03$, the smallest attainable at six, overstates the evidence); under $\mathrm{P}_{\mathrm{early}}$ it

changes τ by $+0.09$ ($p = 0.22$, better in 4 of 6). At full size the negative result under P_{late} is firmer, not weaker, and the P_{early} gain stays small and unresolved.

X.4 Two controls on the criterion itself

The counts of Section 4.2 rest on two criteria, and each deserves a control the paper can be held to.

Grid density. RII is a range over K noisy per-depth estimates, so it is biased outward and the bias grows with K; Table 8 states this for W_m and it applies at both ends here. H4 is the only 21-point cell and it reports the largest counts, which is exactly the pattern that confound would produce. H1's grid is every other point of H4's, so subsampling H4 to those 11 depths holds the items, the records, the presses and the budgets fixed and varies only K. It changes nothing: the largest family-wise count is 8 at 21 depths and 8 at 11, and the depth-contingent count averages 10.0 against 10.0. The counts of Table 2 are not an artefact of grid density.

W_m is a different matter, and since Section 6 asks other people to report it the size of its grid dependence should be stated rather than only warned about. On the same records, halving the grid raises W_m by 0.007 on average and by at most 0.064, changing it in 27 of 60 (press, budget) rows. The shift is structured: 0.002 for the window templates, whose profiles are step functions and whose worst case is the interior at any grid, against 0.008 for the content-driven presses, whose profiles are graded and so have a worst depth a coarser grid can miss. A W_m compared across papers that used different grids is therefore comparable to about 0.064 for a factor-of-two difference, which is smaller than most of the contrasts in Table 8 but not smaller than all of them.

The same bias shows in the intervals. A bootstrap percentile interval for a minimum is built from resampled minima, which are systematically at or below the observed one, so the interval can sit just under its own point estimate: this happens in 12 of 1,080 rows (1.1%), always at the upper end and always by less than a rounding step of Table 8. Read those intervals as bounding the sampling spread, not as covering W_m.

False positives. A criterion is only as good as what it does on a contrast with no depth structure. We build one by permuting the depth labels within each (item, press) of a real published pair, which destroys depth structure while leaving every marginal and the item-level pairing exactly as measured, and apply both criteria to the result over 240 (pair, budget, cell) permutations. The family-wise band certifies 0 of them. That is the statistic every certified count in this paper uses, and on this null it never fires; with 240 draws this bounds its false-positive rate at about 1.2% (95%), and a pair whose depth-averaged contrast is far from zero cannot produce a certified sign change under a within-press depth permutation at all, so the null is most informative for near-tied pairs.

The least favourable configuration. That null makes δ nearly constant over depth, where a false certificate needs two false exclusions of opposite sign. The least favourable case for the criterion is a pair with a real effect at one depth and none elsewhere, where one false exclusion suffices. We built it from the records: for each of the 720 (published pair, budget, protocol, cell) combinations of H1–H4 we keep the depth with the largest measured contrast and make the contrast exactly zero at every other depth by swapping the two presses' scores within each item with probability one half, then certify as in Appendix D, over 20 swap seeds. The band falsely certifies 31 of 14,400 constructions, a rate of 0.0022 [0.0015, 0.0031], about twice the nominal per-pair 0.0011: part of that is the percentile rule of Appendix D and the rest is a percentile band from 40–50 items covering less than it should. At half the level the rate is 0.0017, and the counts barely move: summed over H1–H4 under P_{late} the 95 published certifications become 94, and every per-cell maximum is unchanged. Even if every published pair were in this configuration, the expected number of false published certificates per (cell, protocol, budget) would be $15 \times 0.0022 \approx 0.03$, against the 1–8 observed.

Depth-contingency does fire, on 25% of them. It is read off point estimates and clears no band, so this is the expected behaviour of a range over noisy estimates rather than a defect in the implementation — but it means a depth-contingent share is a description of the point estimates and not evidence that a pair reverses. The shares quoted in Sections 4.2 and 5 and Appendices P, R and S should be read against that 25% floor (the lower ends of several of those ranges sit at or below it), and where this paper needs a pair to have reversed it uses the family-wise count instead. An unpaired version of the

same null, splitting one press's items into halves, fires on 119 of 120 and certifies 0, which is what more noise does to each criterion.

X.5 THE UNIFORM AVERAGE OVER LAYERS AND HEADS

Definition 1 averages $|E \cap S|/|E|$ over every layer and KV head, so two presses keeping the evidence in different heads but the same number of head-slots receive the same ρ. The retrieval-head literature of Appendix O says head identity matters, and the two presses the shared link mispredicts worst in H1 are H_2O-acc and KeyNorm (Table 8; the values here, 0.13 and 0.25, are on the recomputed keep sets of the note below and differ from the table's by the residue it reports). Retention needs one prefill and no decoding, so alternative aggregations of the *same* keep sets can be scored against accuracy records that already exist: we recompute H1's keep sets and fit each one under the same leave-one-press-out isotonic procedure and the same baseline (660 cells, 33,000 joined records).

aggregation of $\lvert E \cap S\rvert/\lvert E\rvert$ over (ℓ, h)	mean R^2	median	H_2O-acc / KeyNorm
uniform mean (Definition 1)	0.77	0.91	0.13 / 0.25
min over layers of the per-layer mean	0.23	0.40	−0.77 / −0.49
mean over layers of the max over heads	0.82	0.95	0.45 / 0.36
fraction of (ℓ, h) retaining $\geq$ half	0.88	0.91	0.79 / 0.62
weighted by the head's attention on E	0.92	0.95	0.72 / 0.83

The objection is correct. Weighting each head by the attention mass it puts on the evidence in the uncompressed prefill (still label-free and decode-free, so still within the probe's cost) reaches R^2 0.92 against 0.77 for the uniform mean, and it recovers most of the gap on exactly the two presses the uniform average mispredicts; five aggregations were scored on H1 and this is the best of them, so the gain is an in-sample selection and would need the other task-H cells to be confirmed. The uniform mean of Definition 1 is therefore not the best aggregation of its own ingredients; the framework is if anything stronger than the paper claims once head identity is respected. We keep Definition 1 as stated, and are precise about what that choice does and does not touch. The non-identifiability results (the certified sign changes, the RII, Φ, W_m and the distinct-ranking counts) are accuracy statistics that never read ρ_m, so no aggregation choice can move them. What the choice does govern is the mediation bake-off of Section 4.3 and the probe, and there an attention-weighted profile is better by 0.15 in R^2. We report that rather than redefine the quantity mid-paper, and a head-resolved profile is the obvious next version of it.

A reproducibility note. The probe recomputes the uniform ρ, and it reproduces the released value exactly for the 4 presses whose keep set is a function of position alone (StreamingLLM, LOCAL, RANDOM, RANDOM+OBS) and to a mean absolute difference of 0.005 for the content-driven ones. That residue is not a different definition. A content-driven keep set is chosen by ranking bfloat16 scores, and near the budget boundary many scores tie, so the selected set is not bit-reproducible across arithmetic orderings: two probe runs differing only in prefill chunk size move ρ by up to 0.11 on the content-driven presses and leave all 4 position-based presses identical. Anyone comparing content-driven presses at a fixed budget should expect that much irreducible spread from the arithmetic alone.

X.6 THE BAKE-OFF UNDER A DIFFERENT SUMMARY

The leave-one-press-out R^2 quoted in Tables 1 and 2 is a mean over ten held-out presses, and R^2 against a fixed baseline is unbounded below, so one press can move a cell mean as far as it likes. E4 is the case: its mean of 0.01 is StreamingLLM at −3.5 pulling against a median of 0.53. A reader is entitled to ask whether the comparison between predictors survives a summary that one press cannot dominate, and whether it survives dropping the presses that are not deployed methods (LOCAL, RANDOM, RANDOM+OBS and H_2O-acc) whose surfaces are flat or extreme and might be thought easy to predict (H_2O-acc is in fact the worst predicted, Table 8).

It survives both, and the second makes it stronger. Across H1–H4 under P_{late}, ρ against $\bar{\rho}$ is 0.47–0.77 against 0.41–0.58 by mean, 0.63–0.93 against 0.59–0.75 by median, and 0.58–0.78 against 0.36–0.58 over the six published presses alone. The margin is widest in the last of these, so the reported means are if anything conservative: the position-unbiased and accumulated-only controls

are predicted no better than the methods anyone deploys, and including them narrows the gap rather than manufacturing it.

Across presses. The item bootstrap of Section 4.3 holds the ten presses fixed, and the claim is about a held-out press, so the press is the unit that matters. With the ten held-out presses as the sample, the P_{late} gap $R^2_\rho - R^2_{\bar\rho}$ in H1 is 0.19 with a t interval of $[-0.13, 0.50]$ and a median of -0.01: under P_{late} it is carried by the window templates (H_2O 1.02, StreamingLLM 0.66), ρ is ahead for 4–6 of the ten presses across H1–H4, and the press-level interval excludes zero in 0 of the four cells. Under P_{early} ρ is ahead for 8–9 of ten in every task-H cell and the interval excludes zero in 2 of four (H1: 0.58 [0.35, 0.82]); depth-centred, the pooled P_{early} values are 0.93, 0.67, 0.53 and 0.92 in H1–H4, against the P_{late} values of Appendix W.3. The depth-resolved advantage of ρ is therefore a statement about the window templates under query-aware compression and about most presses under query-agnostic compression, and Section 4.3 states it that way.

X.7 The ranking count without the point masses

$\mathcal{D}_K$ contains the K point masses on the grid alongside the uniform prior, three shaped priors and three coarse grids reported in the NIAH literature, so most of the family is point masses and a count over all of it could be restating that the ranking at one depth differs from the ranking at another. That is the accuracy surface, not a claim about priors a benchmark induces.

Recomputing over the 7 non-degenerate members alone (every one of which spreads mass over at least two depths) leaves 3–7 distinct rankings under P_{late} and 4–7 under P_{early}. The first-ranked press changes in 9 of 9 cells under P_{late} (in E4 and E5 only among presses tied at ceiling within a hundredth, which is how Section 4.4 reads those cells), with up to 4 different presses taking first place across the family, and 5 under P_{early}. Dropping the $\{0, 1\}$ grid as well, the most extreme member, the winner still changes in every cell. The count is not an artefact of the point masses.

This is a statement about a stated family, and Table 10 is the different and weaker statement about the priors we actually measured: under P_{late} the 12 audited task priors crown the same press at most budgets. A family of priors one could defend re-orders the presses; the particular priors the 12 audited tasks happen to induce need not, and Section 5 is why we do not claim they do.

X.8 Template and content pairs

A window template has a structural hole at every budget (Proposition 3b), so a pair containing one is the case where Corollary 2 most obviously bites, and the reading that the whole effect is a sliding window missing the evidence has to be answered with counts rather than an argument. We split the 15 published pairs into the 9 that contain a press giving at least half its budget to a contiguous block of the most recent positions at every budget (StreamingLLM, H_2O) and the 6 whose two presses both rank positions by a content score (SnapKV, PyramidKV, TOVA, KeyNorm; the first two keep a 32-token window, below the threshold at every budget), and count family-wise certifications over every (pair, budget, cell) observation of the nine primary cells.

H_2O is on the template side of that line, and the reason is worth stating because an earlier version of this analysis put it on the other side. Half of its budget goes to the most recent tokens by construction, which is a window whatever the other half does, and its measured profile is the window's: in H1 under P_{late} at $B = 256$ its retention is 1.00 at $p = 1$, identical to StreamingLLM and LOCAL, against at most 0.05 on the open middle and 0.47 at $p = 0$ where its accumulated-attention half concentrates. Counting it as content-driven would inflate the content-vs-content group, and counting it as a template strengthens the enrichment the corollary predicts rather than weakening it; this assignment was made after the counts were first computed: with H_2O on the content side the content-pair rate rises and the ratio falls, and the rows below are the corrected split.

Under P_{late} a template pair is certified in 127 of 486 observations (26.1%) and a content pair in 5 of 324 (1.5%), an enrichment of 16.9× in the direction the corollary predicts (five events on the content side, so the ratio's own uncertainty is wide: an exact conditional 95% interval of 7.1–53 before the dependence across budgets and cells, and with the fifteen pairs as the units a permutation over which six are called content gives $p = 0.0004$). The content group is not empty: 2 distinct content-vs-content pairs are certified, in 3 of the 9 cells. Under P_{early} the enrichment nearly vanishes

(34.4% against 28.4%, 1.2×) and every one of the 6 content pairs is certified somewhere: without the query the content scorers acquire positional biases of their own. Depth-contingency, which needs no test, shows the same pattern at a higher level (58.0% against 22.8% under P_{late}).

The two groups do not merely differ in rate; they reverse at different depths, and the two mechanisms of Proposition 3 say which. A certified reversal survives the withholding of a depth set when a significantly positive and a significantly negative depth both remain outside it. Withholding $p = 1$ leaves 31 of the 127 template certifications and 4 of the 5 content ones; withholding $p = 0$ leaves 101 and 0; withholding both leaves 20 template certifications over 3 distinct pairs and 0 content ones. So a template's hole is interior as Proposition 3(b) describes it, and it is measurable strictly inside the grid, though most template certifications (96 of 127) do use $p = 1$, and the interior ones appear at the largest budgets, where the retained tail reaches the $p = 0.9$ grid point (Table 2, "fw int."); a content scorer's bias sits at the front, and every certified content-vs-content reversal uses $p = 0$ as one of its two depths. Proposition 3(a) explains that bias for accumulated attention, not for the window vote, last-query and key-norm scores of this group, whose early bias is an empirical finding. Both are depths a benchmark can realise, but a mean evidence depth of 0.20–0.53 (Section 6) says little about mass at either endpoint, and Appendix U finds the tail sparsely populated; they are different claims and we separate them.

A second reading of the enrichment has to be excluded, or bounded. StreamingLLM and H_2O are the two presses with a measured floor (0.45–0.46, Appendix M), so template pairs could be enriched because of the floor rather than because of the hole. With H_2O on the template side this alternative no longer touches the content group through a measured floor: KeyNorm was measured and has none, and SnapKV, PyramidKV and TOVA never reach $\rho = 0$ in E4, so their floors are unmeasured rather than measured absent (Appendix M). Of the 5 content-vs-content certifications, 1 is at the smallest budget and 4 above it, so the group does not rest on the budget where a floor is most likely.

That bounds the alternative; an intervention settles it. We rebuilt Llama-3.1-8B's task-H cell at 8,125 tokens with both arms on the same items — the normal arm for all ten presses and, for every press whose keep set admits it, the leak-free arm of Appendix Q, in which the retained tail is recomputed against the already-pruned cache so it cannot have attended to anything the press evicted (50 items, 49,500 scored answers, budgets 32–1024). The intervention does what it should: on the 450 or more rows per cell where StreamingLLM retains none of the evidence its score falls from 0.13–0.26 to 0.03–0.23.

The certified pairs do not need it. Replacing every press that has a leak-free arm by that arm and leaving the rest as measured, all 11 of the 11 certified (published pair, budget) observations (8 distinct pairs, 8 observations whose pair contains StreamingLLM and so are actually tested, 3 pairing H_2O with a content press and unchanged by construction) survive, none is lost, and 2 more appear: removing the leak sharpens a template's interior deficit rather than manufacturing it. The test can only fail on a certification in which the template wins at a zero-retention depth, since removing a floor lowers the template there; that it also gains two says the interior deficit was being masked, not manufactured.

The intervention does not reach every template, and we would rather state the gap than imply it away. The recomputation needs the keep set to split at the window boundary into blocks of the same size in every layer and head, which is true of a press whose pattern is positional and false of one whose scores differ per head. So the arm exists for the 5 presses LOCAL, STREAMINGLLM, STREAMINGLLM-S1, STREAMINGLLM-S16, STREAMINGLLM-S64 and not for H_2O, whose accumulated-attention half is head-specific; H_2O enters the comparison through its normal arm only. Two things bound what that leaves open. H_2O's leak is measured rather than removed: Appendix M puts it at $\sigma = 0.53$ and $\omega = 1.00$ with a zero-retention floor of 0.45–0.46, so it is real and of the same size as StreamingLLM's, which the intervention shows does not manufacture a certificate. And on this cell 0 of the 11 certified observations involve no template at all, so the leak question is not settled here for content-vs-content pairs; those are certified on the nine-cell grid instead (2 under P_{late} and 6 under P_{early} above), where no press with a measured floor appears on either side of the pair. One caveat belongs with this. LOCAL keeps no sink block, so by Appendix M it has no floor to remove, and its leak-free arm is not a floor removal but a fresh prefill of the window it retains, which is why its score moves the other way; it is not one of the 15 published pairs and enters none of the counts above.

Both halves of this matter. The concentration is evidence that the mechanism is the one Proposition 3 names, since the presses with a structural hole are the ones whose contrasts most often reverse. The residue is evidence that the claim is not a restatement of that hole: the content-driven presses carry positional biases too, measured rather than derived, and their profiles cross as well — less sharply under P_{late}, which is why that protocol needs a family-wise band and 324 observations to see it, and far more often under P_{early}.

X.9 A PRESS DESIGNED TO BE POSITION-FLAT

Proposition 3 says the deployed methods carry positional priors, which invites the objection that this paper indicts current practice rather than comparison itself: design a method whose retention does not depend on position and the ordering becomes identifiable again. The objection deserves a press, not an argument, so we built one. QUOTA cuts the pressed prefix into 32 contiguous bins of equal width and gives every bin the same number of slots, filled by SnapKV's pooled vote inside the bin; QUOTA+OBS does the same on the prefix and keeps SnapKV's observation window as well. 32 divides every budget in the ladder and the window, so the allocation is exact. What is flat *by construction* is that allocation (every bin receives $B/32$ positions whatever the scores do, which the released suite verifies holds even under a deliberately early-biased scorer), while the scorer inside a bin stays the one a deployed method uses.

Two cells, both self-contained: LLAMA-3.1-8B at 8K and QWEN2.5-7B on H1's task at 16K, 50 items each, all 12 presses in one run, six budgets, both protocols. They have to be self-contained, because a cross-run merge of two presses into a released cell is not safe at this resolution (Appendix N). The 33 tested pairs of each split three ways under one correction over that family: flat against flat (the two quota presses and the two random ones, 6 pairs), the two quota presses against the six published ones (12), and published against published (15). The numbers below are the 8K cell unless stated; the 16K cell is quoted where the two differ.

Three things follow. A quota over bins does not make the *profile* flat. An evidence span is a handful of tokens and a bin is hundreds, so within its bin the span still competes on score. At $B = 1024$ the measured range of ρ_m over depth is 0.340 for QUOTA, against 1.000 for StreamingLLM and 0.459 for SnapKV (about three-quarters of SnapKV's) but nowhere near the 0.004 of random eviction, which remains the only press whose profile is flat to measurement error. The construction reduces the range; it does not remove it.

Second, that is nearly enough to remove the reversals, with the caveat that four presses near the floor have little resolution to reverse with, so a low certified count is weaker evidence here than a high one would be against. Among the flat presses 0 of 6 pairs are certified as sign-changing at any budget under P_{late}, and only 0–2 are depth-contingent, against 0–5 of 15 certified and 5–10 depth-contingent among the published pairs. The 16K cell agrees: 0–1 of 6 flat pairs certified against 1–7 of 15 published ones, so the result is not one model's, though at one 16K budget one flat pair is certified.

The distinct-ranking counts do *not* replicate, and the reason is one this paper has already had to state. In the 8K cell they point the same way (1–3 orderings over the 18 priors of $\mathcal{D}_{11}$ for the flat presses against 4–7 for the published ones) but in the 16K cell they reverse, 2–10 against 4–9, with the flat set higher at the two smallest budgets. That is exactly the failure mode Appendix X.2 measures for the main analysis: at a budget where all four flat presses sit near the floor their means are within item noise of each other, and a prior then reshuffles them for free, so the count describes ties rather than priors. We therefore rest this section on the certified counts, which are tests, and report the ordering counts only to say that they do not support the claim here.

Third, it is not a recommendation, for two reasons. Flatness is expensive. At $B = 128$ under P_{late} QUOTA reaches a uniform-prior mean of 0.20 against 0.94 for SnapKV, a gap of 0.73 (from the unrounded means), and even at 1024 it is 0.90 against 1.00; in the 16K cell the worst gap is 0.45 at $B = 256$ and 0.77 against 0.91 at the top. A method that spreads its budget evenly over the context cannot concentrate it where the evidence is, which is what the query-aware methods are for. And flatness has to be universal to help: pairs with one flat press and one published press are certified in 0–6 of 12 observations under P_{late} and up to 12 under P_{early} (in the 16K cell, up to 6 and 12), because δ depends on both profiles and one of them is still not flat. Making a single method position-indifferent does not make the comparison between it and a deployed method identifiable; only making every

compared method flat would, at the accuracy cost above. That is an argument for depth-resolved reporting, not for a flat press.

X.10 What decides the certified counts

Section 4.2 reports that a cell's family-wise certified count is a fact about the estimator rather than about the geometry, and this is the analysis behind it. Over the 12 cells with a controlled depth grid we take, per cell, the largest family-wise count over budgets under P_{late}, the depth-contingent share, the answer length in tokens (which runs 2 to 15) and the uncompressed accuracy, and rank-correlate them. Certification tracks answer length at Spearman 0.93 and the uncompressed exact-match rate at -0.32; the depth-contingent share tracks length less (0.49) and saturation more (-0.57 against uncompressed exact match). Answer length is confounded with model and task across these 12 cells (Qwen's task H is 14 tokens, Llama's 8, task E 5 and 2), so the correlation cannot separate tokens per item from tokenizer.

Four rank correlations on 12 points support a qualitative claim and no more, and we quote no p-values for them: the cells share models, tasks and depth grids, answer length takes few distinct values across them, and reading a pattern off which of four tests would clear a threshold is exactly the inference Appendix H declines to make elsewhere. What the numbers are used for is the direction and the contrast between the two rows, which is large and has a mechanical explanation: the graded metric is a per-token rate, so an answer of n tokens carries more information per scored item (less than n times, since the tokens of one retrieved answer are far from independent), the paired bootstrap band narrows accordingly, and a band corrected over 45 pairs is what a certificate must clear. The depth-contingent share is read off point estimates and clears nothing, so it does not move with the band. Table 2 carries both quantities for every cell.

X.11 The probe against measuring accuracy directly

The probe returns $\rho_m(p; B)$ from one prefill, with no decoding and no labels. The alternative a practitioner has is to take a small labelled set, decode it at each depth and read the accuracy off. That alternative is the one the probe has to justify itself against, and on the cell where the link is strongest it wins.

Predicting H1's full-sample per-(press, depth) accuracy under P_{late}: the leave-one-press-out link from ρ reaches R^2 0.77 and a mean per-depth ordering τ of 0.70 (Table 2 reports the median, 0.67) while decoding nothing for the held-out press, though the link itself is fitted on the other nine presses' scored answers. 1 labelled item, scored for every press, budget and depth (660 teacher-forced forwards) reaches 0.88 and 0.70; 20 items (13,200) reach 1.00. The target is the full 50-item mean, which includes the scored items, so the comparison is partly in-sample and flatters the labelled arm; it does not need the help. Measuring accuracy on even a few labelled items predicts better than the link.

The two are not doing the same job, which is what the probe's case rests on. The accuracy probe must *run* the press it reports, on data whose answers are known; the link predicts a press whose accuracy it has not scored, from a quantity that needs no answers for that press. Where a method can be run and labels exist, decode — the probe is not a cheaper substitute for measurement. Its case is the other one: a press or a budget that has not been run, a corpus with evidence positions but no answers, and the structural facts of Proposition 3 (where a profile is flat, where two profiles cross) which are properties of the keep set and need no accuracy at all.

X.12 Prior mass at the reversals, from audited benchmarks

Proposition 1(ii) guarantees that some prior reverses a depth-contingent pair; the fair question is whether the priors benchmarks actually induce are among them, or whether a reversal needs mass at the first or last depth that no real benchmark supplies. For every certified published-pair observation of H1–H4 (219 over both protocols and all budgets) we computed the smallest mass w^* a prior must place on the depths where δ has the minority sign to reverse the uniform-prior order, and the mass each of the 12 audited task priors of Section 6 actually places there (`analyze_prior_reversal.py`). The median w^* is 0.09, and 92 of the 219 certified observations are reversed by at least one audited task prior: 14 of 95 under P_{late} and 78 of 124 under P_{early}. The count falls to 84 when the multi-answer tasks' priors are built from all answer spans instead of first offsets (Appendix U), so the front bias

of that statistic is not what produces it. Real priors therefore do reverse certified pairs, mostly under query-agnostic compression, where the content scorers' profiles tilt toward the front and a front-loaded task prior is enough; under query-aware compression most certified reversals sit at the final depths, where the audited tasks put little mass, which is the same boundary Section 5 meets on RULER's own rows.