PAGE: PARTITION-AWARE GATED KV-CACHE EVICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

KV-cache eviction can do more than compress. In long-context LLMs, keeping only some cached tokens sometimes matches or exceeds full-cache accuracy, because many redundant prefill tokens otherwise dilute attention away from the tokens that carry the answer. This benefit is not uniform, and evicting the wrong tokens drops accuracy to zero on tasks that need precise retrieval, so the useful question is not only which tokens to keep but whether to evict this input at all. We show that one label-free number computed from the prefill attention, the drop between early and late layers in how much attention heads agree on which tokens to read, predicts per input, before any decoding, which of the two cases an input falls under. We build this into PAGE (Partition-Aware Gated Eviction), a wrapper that runs any SnapKV-style evictor when the drop is large and keeps the full cache when it is small, with no training, labels, or fine-tuning. PAGE is a safety mechanism rather than a compressor, so we measure it by the failures it prevents. It cuts the harm rate on capacity-bound inputs from 0.75 to 0.026, and on multi-key retrieval with Mistral-7B plain SnapKV falls from 99% to 0% as the budget shrinks while PAGE holds it at 89%. Elsewhere it passes the base evictor through unchanged, which is the intended behaviour and is what we observe in 8 of 16 cells. The prediction is one-sided, identifying evictable inputs reliably and resolving the single-near-tie-distractor boundary only with per-input information beyond the mean drop. The head-agreement drop orders tasks the same way across four model families, but the threshold does not transfer, and an unlabeled roughly 100-input pilot per model is needed outside Qwen2.5 and Mistral. Realized compression is 1.8× to 3.4× against a nominal 16× budget and decays toward unity by batch 16, and at matched memory a trained evictor outperforms PAGE. Code is available at `https://anonymous.4open.science/r/PAGE-018239`.

1 INTRODUCTION

Large language models (LLMs) are increasingly used for long-context tasks such as document understanding, retrieval, and extended reasoning. As context length grows, the KV-cache grows linearly with the input and becomes a memory bottleneck during inference. KV-cache eviction addresses this bottleneck by retaining only a subset of the states under a fixed memory budget (Zhang et al., 2023; Xiao et al., 2024; Li et al., 2024; Cai et al., 2025). Existing methods largely share the same formulation: given a fixed budget, determine *which* KV states to retain. When discarded states are redundant, this can reduce attention dilution and even improve long-context performance (Bui et al., 2026; Yang et al., 2026a;c); however, when the answer depends on a small set of critical tokens, aggressive eviction can instead destroy essential information and severely degrade retrieval.

Recent work has therefore focused on improving which states are retained through stronger importance signals, learned predictors, and information-theoretic objectives (Li et al., 2024; Cai et al., 2025; Yang et al., 2026a; Bui et al., 2026; Moschella et al., 2026). While these methods can improve the accuracy–memory trade-off and even outperform full-cache inference by mitigating attention dilution (Bui et al., 2026), they still assume that eviction is appropriate for every input. In particular, adaptive approaches either learn a scorer to estimate which states should be retained (Bui et al., 2026; Yang et al., 2026c; Dong et al., 2026; Moschella et al., 2026) or derive an unconditional capacity measure from query statistics (Yang et al., 2026a). This leaves two fundamental questions open: **(Q1)** *For which inputs does KV-cache eviction help or hurt?* and **(Q2)** *Can this distinction be*

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.

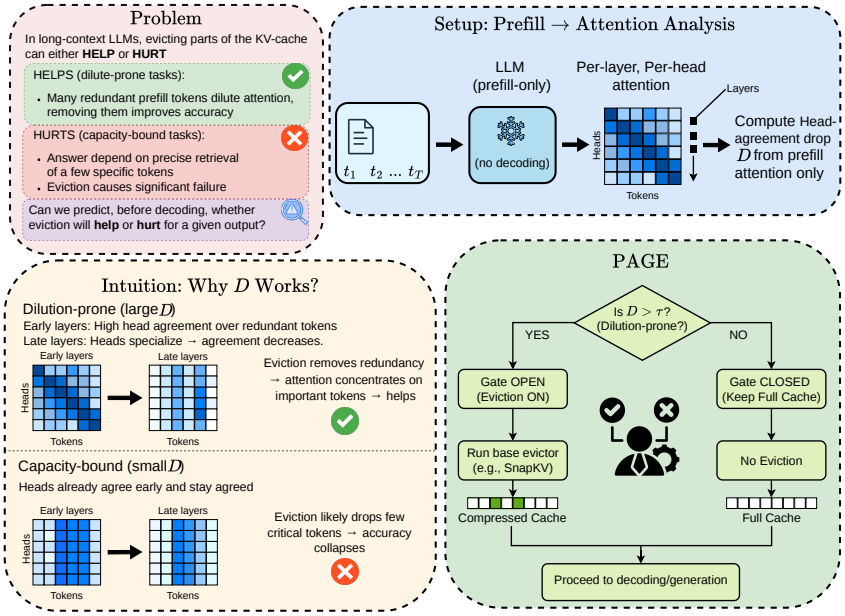


Figure 1: PAGE uses prefill-only attention to compute the head-agreement drop D and decide *whether* to evict. A large D indicates dilution-prone inputs, for which PAGE opens the gate and applies a base evictor, and a small D indicates capacity-bound inputs, for which the gate remains closed and the full cache is retained.

predicted a priori from a cheap inference-time signal before decoding? These questions are important because a benchmark-level accuracy–memory trade-off can mask a subset of inputs for which eviction is catastrophic.

We answer both affirmatively by reframing KV-cache eviction as an *input-dependent admission decision*: rather than assuming that eviction should be applied to every input, we ask whether the model's prefill attention contains a signal indicating when eviction is likely to be beneficial or harmful. We find that the early-to-late drop in mean pairwise top-k head agreement, denoted by D, provides a useful label-free signal that partitions inputs into a dilution-prone class $\mathcal{D}$ and a capacity-bound class $\mathcal{C}$ before any decoding (Figure 1). Based on this observation, we introduce **PAGE** (**P**artition-**A**ware **G**ated **E**viction), a training-free wrapper that applies an existing eviction method when D exceeds a threshold and otherwise retains the full cache. Since a benchmark mean can hide exactly this failure mode, we evaluate PAGE by what it prevents on the capacity-bound regime rather than by an average over tasks. It cuts the rate at which eviction destroys a correct answer by $29\times$ at $16\times$ compression, and at matched *achieved* cache it lifts the accuracy-memory frontier above a crossover near $1/3$ kept cache, losing to plain eviction below it (Figure 4). The same selectivity has a converse. The gate is inactive wherever eviction is already safe, so on inputs that eviction does not harm the gated and plain systems become identical and their measured difference is exactly zero. PAGE is designed to prevent harmful eviction, not to maximize average compression.

Contributions

1. **An empirical partition** of inputs into $\mathcal{C}$ and $\mathcal{D}$, demonstrated across eight (model, context) cells on RULER (Hsieh et al., 2024) and confirmed by a controlled distractor sweep. The early-to-late head-agreement drop $D := \mathrm{agr}_{\mathrm{early}} - \mathrm{agr}_{\mathrm{late}}$ orders tasks by eviction safety and separates the two classes at the task level (Section 4.1).
2. **PAGE, a training-free gating wrapper** that applies any base evictor when $D \geq \tau$ and retains the full cache otherwise, with a z-score calibration recipe that transfers across four architecture families (Section 3).

3. **A substrate-dependent capacity limit**: the three scorer families we test (attention-based, geometric, attention-free) all collapse on capacity-bound inputs at aggressive budgets, evidence of an information floor that no scorer among these families removes (Section 4.1).
4. **A scaling relation** $\rho \approx (1-A_{\text{full}})\cdot p_{\text{recov}}$, validated by head-to-heads with CapKV and DBTrimKV and a pre-registered 32K test that preserves ordinal content but falsifies constant-ratio proportionality (Sections 4.1, 5).

We adopt the dilution mechanism of Bui et al. (2026) (Proposition 3.1, Corollary 3.2) and repurpose a known head-agreement primitive as a gating signal; Section 2 positions these choices. PAGE is a safeguard, not a compressor: at matched memory a trained evictor wins (Section 6), it is inactive wherever eviction is already safe ($\Delta = 0$ in half the matrix cells), and the $+22.9$pp headline is driven by one task family (excluding NIAH-MK3: $+4.5$pp; task-label oracle: $+20.1$pp). Realized compression is 1.8–$3.4\times$ (mean $2.9\times$) and decays toward $1\times$ by batch 16 under static provisioning.

2 Related Work

KV-cache eviction methods primarily differ in how they identify states to retain under a fixed budget. H2O (Zhang et al., 2023), StreamingLLM (Xiao et al., 2024), SnapKV (Li et al., 2024), and PyramidKV (Cai et al., 2025) use attention, recency, or layer-dependent statistics for retention. Recent methods develop more specialized selection criteria: DBTrimKV (Bui et al., 2026) characterizes attention dilution from near-tie distractors and uses selective retention to mitigate it, while CapKV (Yang et al., 2026a) derives an information-theoretic eviction criterion. IndexMem (Yang et al., 2026c) combines learned indexing with latent memory, and ForesightKV (Dong et al., 2026) and Learning-to-Evict (Moschella et al., 2026) predict future KV utility. VECTOR (Lin et al., 2026b) and CriticalKV (Feng et al., 2026b) instead provide plug-and-play mechanisms over existing eviction policies. These methods retain the conventional formulation of selecting which states to preserve once eviction is applied.

Adaptive eviction methods further condition retention or budget allocation on the input or task. DynamicKV (Zhou et al., 2025) uses task-dependent attention statistics to allocate layer-wise KV budgets, while EvicPress (Feng et al., 2025a) incorporates a context-level eviction decision into a profiled quality-latency optimization. Attention-Gate (Zeng et al., 2025) and Fast KVzip (Kim et al., 2026) learn adaptive gating or budget policies. SAGE-KV (Wang et al., 2025) is particularly close in its use of post-prefill attention, using a one-shot attention signal to select tokens and heads for retention. PAGE instead uses the prefill signal to characterize whether the eviction operation should be applied, while the underlying retention policy remains unchanged.

A complementary line of work studies the internal attention structure associated with retrieval and compression. Retrieval-head analyses identify sparse subsets of heads that drive long-context retrieval (Wu et al., 2025), while Ananthanarayanan et al. (2026) analyze how head consensus evolves across layers and relate these profiles to compression tolerance. Chen et al. (Chen et al., 2026) document task-dependent degradation under eviction, and Zhang et al. (Zhang et al., 2026) study when value-aware eviction is beneficial. VaSE (Chang et al., 2026) attributes some eviction failures to high-magnitude value states, whereas ManifoldKV (Datta et al., 2026) considers geometric structure in key representations. PAGE draws on this diagnostic perspective by using the early-to-late change in head agreement as an input-level prefill statistic rather than as a retention score or an offline architectural descriptor.

3 The Gating Method: PAGE

Problem Formulation For task T, model M, context length T, let $A(b, x)$ denote the accuracy on input x at cache budget b, with $b_{\max} = 1$ denoting the full, unevicted cache. We call an input *capacity-bound* if eviction at every tested budget degrades its accuracy relative to the full cache, and *dilution-prone* if there exists some budget at which eviction improves accuracy. The key evaluation criterion is the per-input recovery rate:

$$\rho_{\text{KV}}(\mathsf{T}, M, T) := \Pr_{x}\big[A(b^*(x), x) > A(b_{\max}, x)\big], \qquad b^*(x) := \arg\max_{b} A(b, x). \tag{1}$$

This per-input strict-Pareto improvement frequency is a stronger statement than "mean accuracy is higher at some b." We pre-registered $\rho \geq 0.05$ as "dilution-prone" definition before any sweep.

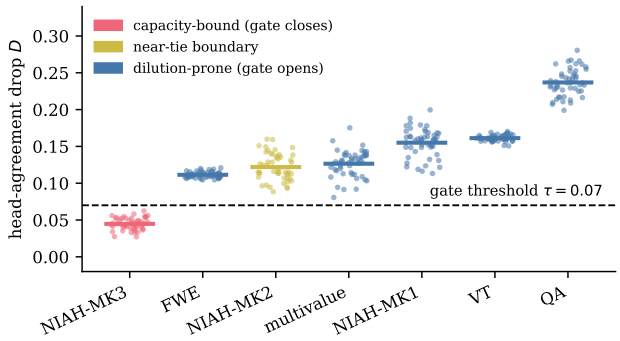


Figure 2: D separates the two task classes (Qwen2.5-1.5B, RULER 4K, $N = 50$ per task; dots are inputs, bars per-task means). NIAH-MK3 sits below $\tau = 0.07$ and the dilution-prone tasks above it. The near-tie MK1/MK2 tasks fall on the boundary, tracking the distractor gradient (App. A).

Algorithm PAGE wraps any SnapKV-style evictor in one conditional. Given a prompt x, the base evictor's budget $b \in (0, 1]$ (the fraction of prompt positions, the evictor is allowed to retain), a sink-and-recency protection (n_{sink}, w) that always preserves the first n_{sink} and last w positions, and a gate threshold τ on the head-agreement drop:

$$PAGE(x;\, \text{base}, b, \tau) = \begin{cases} \text{base.evict(cache}, b) & \text{if } D(\text{cache.attentions}) \geq \tau \\ \text{cache} & \text{otherwise} \end{cases} \tag{2}$$

where the prefill produces `cache` together with its prefill attentions, from which we compute D as in Eq. (3). The gate fires once per input, at prefill, before any decode step, and involves no backward pass, no training, and no per-token decision.

Head-Agreement Drop D Take the prefill attention tensor $[L, H, T, T]$, for L layers, H heads per layer and prompt length T. Project onto the last $w = 32$ rows, the observation queries, and average across them for a per-(layer, head, key) attention vector. For each layer and head pair we take the Jaccard similarity of the two heads' top-k key sets ($k = 32$) and average over pairs, giving a per-layer agreement a_ℓ. Binning the L layers into thirds and writing a_{early}, a_{late} for the mean agreement in the first and last bins, we set

$$D := a_{\text{early}} - a_{\text{late}}. \tag{3}$$

D is one scalar per input, costing $O(LH^2k)$: 17ms on Qwen2.5-1.5B and 89ms on Mistral-7B at 4K (Table 28). It is not free even when the base evictor already scores by attention, since that scoring uses head-*averaged* attention whereas D needs the per-head pattern, which requires materializing attention that fused FlashAttention (Dao et al., 2022) does not produce. A same-inputs ablation against five other cheap prefill statistics justifies the cost: D is the only one that separates the partition, every alternative failing task-level separation on the clear-cut NIAH-MK3 anchor. Table 3 also tracks how that separation narrows as the anchor moves toward the near-tie MK2 boundary.

Calibration We distinguish two senses of label-free. PAGE is accuracy-label-free because the gate never observes correctness. A calibration variant is additionally task-label-free when it also requires no task identities to set τ. The two senses come apart across the variants below.

The fixed $\tau = 0.07$ used throughout is a within-family convenience. It mis-classifies Qwen2.5-3B at 16K and fails outright on Qwen3. The recommended variant standardizes D per model over an unlabeled $\sim$100-input pilot, and is task-label-free. It transfers a single global threshold zero-shot to every held-out family tested: Llama-architecture models (Yi-1.5-9B, Llama-3.1-8B) and Qwen3-4B (AI et al., 2025; Grattafiori et al., 2024; Yang et al., 2025). The intermediate 40-input variant of Eq. (5) is accuracy-label-free but not task-label-free, because it uses task identities to form $\bar{D}_{\mathcal{C}}$ and $\bar{D}_{\mathcal{D}}$. We report fixed-τ numbers as the conservative default. Table 8 maps the three variants to their label requirements and to the results that use each, so every gated number can be traced to the variant that produced it. The one number computed at a threshold other than its reported value is the

Table 1: Per-input recovery rate ρ on RULER with Wilson 95% intervals (small type). Bold marks point estimates $\rho \geq 0.05$. NIAH-MK3 (precise multi-key retrieval) is the only task that systematically fails the threshold in every cell tested. $N = 100$ per (task, cell) except Qwen-14B ($N = 50$ at 4K, $N = 30$ at 16K), Mistral 16K VT/FWE/QA_1/NIAH-MK3 ($N = 50$), Qwen-3B 16K NIAH-MK3 ($N = 50$), and Qwen-1.5B 16K NIAH-MK3 ($N = 46$). At these N most single cells straddle the 0.05 threshold and no per-cell call is individually powered; the claim rests on the pooled comparison, which is unambiguous (NIAH-MK3: $2/526$, $\rho = 0.004$, Wilson $[0.001, 0.014]$). Per-cell N varies with compute budget and with the number of examples available at each context length, not with any post-hoc selection.

	Qwen 1.5B		Qwen 3B		Qwen 14B		Mistral	
Task	4K	16K	4K	16K	4K	16K	4K	16K
VT	**0.06** [.03, .13]	**0.19** [.13, .28]	0.00 [.00, .04]	**0.05** [.02, .11]	0.00 [.00, .07]	0.03 [.01, .17]	0.00 [.00, .04]	**0.32** [.21, .46]
FWE	**0.06** [.03, .13]	0.04 [.02, .10]	0.01 [.00, .05]	**0.32** [.24, .42]	0.00 [.00, .07]	**0.20** [.10, .37]	0.02 [.01, .07]	**0.22** [.13, .35]
QA_1	**0.08** [.04, .15]	**0.14** [.09, .22]	0.03 [.01, .09]	**0.08** [.04, .15]	**0.06** [.02, .16]	**0.10** [.03, .26]	0.04 [.02, .10]	0.00 [.00, .07]
QA_2	**0.08** [.04, .15]	**0.05** [.02, .11]	0.04 [.02, .10]	0.02 [.01, .07]	0.02 [.00, .11]	0.03 [.01, .17]	0.01 [.00, .05]	0.00 [.00, .04]
niah_multivalue	**0.07** [.03, .14]	**0.10** [.06, .17]	0.02 [.01, .07]	**0.13** [.08, .21]	**0.08** [.03, .19]	**0.07** [.02, .21]	**0.06** [.03, .13]	**0.21** [.14, .30]
NIAH-MK3	0.01 [.00, .05]	0.00 [.00, .08]	0.00 [.00, .04]	0.02 [.00, .11]	0.00 [.00, .07]	0.00 [.00, .11]	0.00 [.00, .04]	0.00 [.00, .07]

Qwen2.5-3B 4K SnapKV matrix cell, run at $\tau = 0.04$ and reported at $\tau = 0.07$ by exact post-hoc re-evaluation (Appendix G). Further calibration details appear in Table 7 and Figure 6, Appendix A.

4 EXPERIMENTS

Setup We use RULER (Hsieh et al., 2024) via the `simonjegou/ruler` (Jegou, 2024) port distributed with kvpress (Devoto et al., 2025), whose four-evictor lineup our base evictors follow, at context lengths $T \in \{4\text{K}, 16\text{K}\}$ on four models: Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-14B-Instruct (Qwen et al., 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). Eviction is SnapKV-style by default: we score each prompt position by the average attention from the last $w = 32$ queries across heads and layers, then keep the top-b positions with $b \in \{0.0625, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1.0\}$, and always preserve the first $n_{\text{sink}} = 4$ sink positions (Xiao et al., 2024) and the last $w = 32$ recent positions. Cache slicing uses explicit `position_ids` to preserve rotary position embedding (RoPE) alignment (Su et al., 2024), rather than the attention-mask trick that silently shifts RoPE on evicted positions. Task codes are RULER's: VT (variable tracking), FWE (frequent-words extraction), QA_1/QA_2, and NIAH-MKk = `niah_multikey_`k (k near-tie keys, one carrying the answer). The matrix uses RULER 4K with the four-task mixed suite ($N = 100$ each) across all sixteen cells. Gating deltas carry 95% cluster-bootstrap intervals, all 16 positive with smallest lower bound $+9.3$pp, and remain significant under Benjamini-Hochberg correction at $q = 0.05$ across the 16 cells (Appendix A.6.2). Full protocol and deferred tables are in Appendices A and A.6.2.

The partition We define the partition as a pooled comparison: across the eight (model, context) cells, only $2/526$ NIAH-MK3 inputs recover under eviction ($\rho = 0.004$, Wilson $[0.001, 0.014]$), whereas five task families reach $\rho \geq 0.05$ at some context length. Because individual cells at these sample sizes are underpowered, we interpret Table 1 as this pooled comparison rather than as 48 separate calls. We restrict the definition of the partition to 4K and 16K. The single 32K test-case yields $\rho = 0.04$ for NIAH-MK3, just below the 0.05 threshold that defines the class (Section 4.1), and consequently falls outside our fitting range. These below-threshold cells arise from two mechanisms. In saturated cells, $A_{\text{full}} \to 1$ leaves little headroom for eviction. Longer contexts create headroom, as predicted by Eq. (4). The remaining three exceptions are budget-insensitive: headroom exists, yet the same inputs remain incorrect across budgets. These three are all QA cells, consistent with failures arising from parametric knowledge rather than cache allocation (Appendix A).

Boundary and Confound Analysis A single capacity-bound task leaves open whether NIAH-MK3 names a class or an isolated construction, so we tested the boundary on four further RULER tasks chosen by mechanism rather than by outcome (Table 4, Appendix A.2.4). The results confirm a second capacity-bound member (niah_multiquery) and show that neither distractor count nor output length defines the class; what matters is whether distractors are near-tie surface-form variants of the answer. Furthermore, three controls rule out the obvious confounds. First, a distractor sweep

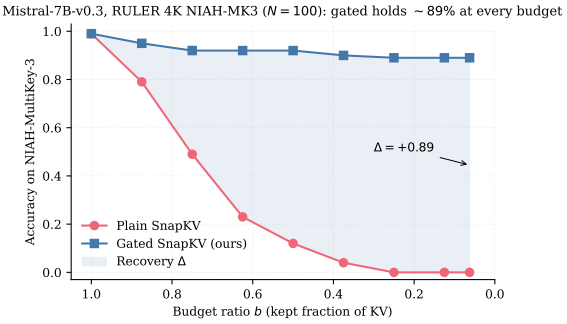


Figure 3: Mistral-7B 4K NIAH-MK3, $N = 100$: accuracy vs. KV budget for plain SnapKV (solid, collapses to 0) and gated SnapKV (dashed, 0.89 flat across budgets). The gate prevents the collapse without retraining the base evictor or seeing labels.

over the RULER NIAH-MultiKey family confirms the dilution mechanism of Bui et al. (2026): mean D shrinks monotonically as near-tie distractors accumulate (Table 5). Second, re-running the partition under the 10% prefix and suffix protection of Garcia (2026) leaves it unchanged. Third, replacing SnapKV scoring with uniform-random scoring destroys accuracy at every budget, ruling out a generic regularization explanation (Appendix A.4.9, Table 20).

External validation The partition is not an artifact of synthetic RULER data. On three LongBench subtasks with Qwen2.5-14B (Table 12, Appendix A), the gate calls all three classes correctly *a priori* from D alone: lcc (capacity-bound, $\Delta = +0.146$), repobench-p (Liu et al., 2024) (mixed), and hotpotqa (Yang et al., 2018) (evict-robust, $\Delta = 0.000$). The partition is sufficiently reproducible to motivate asking whether it can be detected before decoding.

4.1 Analysis of PAGE

Predicting Unsafe Eviction Figure 2 shows that D separates the two task classes on the fitting cell (Qwen2.5-1.5B, RULER 4K): NIAH-MK3 sits below $\tau = 0.07$ and the dilution-prone tasks above it. Per-input AUC reaches 1.000 on that cell, dropping to 0.978, 0.899, and 0.741 on three held-out cells (Appendix A.3.2). D is a useful one-sided prefill signal: reliable at closing the gate on capacity-bound inputs, only weakly discriminative on dilution-prone ones. On Yi-1.5-9B and Llama-3.1-8B the drop ordering transfers but $\tau = 0.07$ does not; the z-scored variant with an unlabeled pilot maintains AUC ≈ 0.80 (Table 26).

Catastrophic Failure Prevention The cleanest single cell demonstrates the safety property without aggregation. On Mistral 4K NIAH-MK3 (Figure 3), plain SnapKV collapses from 0.99 to 0.00 at $b = 0.0625$, while gated SnapKV holds 0.89 at every budget. The gate closes on $90/100$ inputs, with $10/100$ false positives where it opens and eviction destroys the needle. Pooled over the 7 SnapKV cells at 4K and 16K, gating cuts the fixed-budget harm rate from 0.75 to 0.026 on the capacity-bound task (Table 29), yielding a $29\times$ reduction in harmful eviction that stands as the primary safety result.

Transfer Across Eviction Mechanisms Table 2 reports a 4×4 matrix of four base evictors and four models under the same gate, same τ, and no per-method tuning. Each Δ is gated minus plain accuracy at matched *nominal* budget, not a compression win. The gate composes with several tested eviction mechanisms, which means it works alongside them rather than advancing the frontier on its own. Dropping NIAH-MK3 takes the grand mean from $+22.9$pp to $+4.5$pp. In half the cells the non-MK3 Δ is exactly 0.000 because the gate opens on every non-MK3 input there, which is the intended behavior for a gate that should do nothing when eviction is safe. The defensible headline is the harm reduction (Table 29) and the matched-cache frontier (Section 4.1). Composition also extends outside the attention-score family, lifting a geometry-based evictor by $\Delta = +16.2$pp (Figure 8). Cross-architecture results appear in Tables 16 and 17. Full per-task decomposition, budget-axis breakdown, and cross-architecture tables are in Appendix A.4.1.

Table 2: Headline 4×4 method-agnostic matrix. Same gate ($\tau = 0.07$), no per-base-evictor tuning; Δ averages over budgets $b < 1.0$. All 16 cells are positive with grand mean $+22.9$pp, but the second Δ row shows what that is made of: excluding NIAH-MK3 the grand mean falls to $+4.5$pp and is exactly 0.000 in 8 of 16 cells, where the gate opens on every non-MK3 input and the two arms are the same system. The Qwen2.5-3B SnapKV cell ran at $\tau = 0.04$ and is reported at $\tau = 0.07$ by exact post-hoc re-evaluation (Appendix G). †Qwen2.5-14B H2O is not an independent cell (see text). Per-task decomposition: Table 11.

	SnapKV		H2O		StreamingLLM		PyramidKV	
Model	plain	gated	plain	gated	plain	gated	plain	gated
Qwen2.5-1.5B	0.438	0.559	0.196	0.357	0.033	0.195	0.417	0.570
Δ	+0.121		+0.162		+0.163		+0.152	
Δ no MK3	+0.000		+0.000		+0.000		+0.000	
Qwen2.5-3B	0.580	0.839	0.339	0.642	0.052	0.482	0.536	0.865
Δ	+0.259		+0.303		+0.430		+0.329	
Δ no MK3	+0.101		+0.098		+0.263		+0.144	
Qwen2.5-14B	0.709	0.853	0.668†	0.868†	0.080	0.330	0.632	0.860
Δ	+0.143		+0.200		+0.250		+0.228	
Δ no MK3	+0.000		+0.000		+0.000		+0.000	
Mistral-7B	0.623	0.810	0.400	0.636	0.102	0.365	0.593	0.823
Δ	+0.187		+0.236		+0.263		+0.231	
Δ no MK3	+0.015		+0.020		+0.053		+0.020	

Accuracy-Memory Frontier The gate falls back to the full cache when it closes, so the honest comparison is at matched *achieved* cache. At matched achieved kept-KV (Figure 4), gating beats plain SnapKV by $3 - 19$pp once the target retains more than about a third of the cache, and loses below that where the full-KV fallback dominates. PAGE recovers 69% of the headroom against a per-input oracle, almost entirely from protecting the capacity-bound anchor. This recovery is bounded by how often the gate opens, since $1 - p_{\text{open}}$ of the cache stays resident regardless of the nominal budget b, and so memory should be planned against that floor (Table 24).

The same floor sets the realized memory benefit. Against a $16\times$ nominal budget, realized compression is 1.8–$3.4\times$ with mean $2.9\times$, because the resident $1 - p_{\text{open}}$ fraction limits what eviction removes. Under static provisioning, where one gate-closed sequence forces the batch to the full cache, this benefit decays toward $1\times$ by batch 16 (Table 24). PAGE is therefore a per-input safeguard rather than a compressor, with a memory benefit that is real at moderate batch and small kept-KV targets and negligible once either the batch is large or the target is aggressive.

Scorer Independent of Capacity Floor The gate's value depends on whether the capacity-bound regime is a property of the input rather than the scorer. On Qwen2.5-1.5B RULER 4K, a faithful ManifoldKV+Ada-KV implementation (Datta et al., 2026; Feng et al., 2025b) and the attention-free KeyDiff (Park et al., 2025) both fall to chance on NIAH-MK3 below 25% budget, like the four attention-score evictors. A geometry scorer that reaches 0.72 on two-key NIAH against SnapKV's 0.16 still collapses on three-key retrieval. The capacity limit is therefore substrate-dependent rather than universal; full tables and the Llama-3.1-8B rescue result are in Appendix A.4.4 (Table 15).

Per-Input Gating Beyond Task-Level Routing A natural objection is that a simple task-level policy could achieve the same result. The task-label oracle closes on every NIAH-MK3 input and opens on every other, the best any task-level policy can achieve without test-time labels. In 8 of 16 cells PAGE reproduces the oracle exactly; in 5 cells it *beats* the oracle by up to $+0.198$ by closing on dilution-prone inputs the oracle wrongly opens, and in 3 Mistral cells it loses by at most 0.009 through false positives on MK3. Overall the oracle reaches $+20.1$pp against PAGE's $+22.9$pp. Task-level routing captures a substantial portion of the effect, but can fail on heterogeneous tasks where within-task variation makes per-input resolution useful. Full per-cell results appear in Appendix A.4.1 (Table 14).

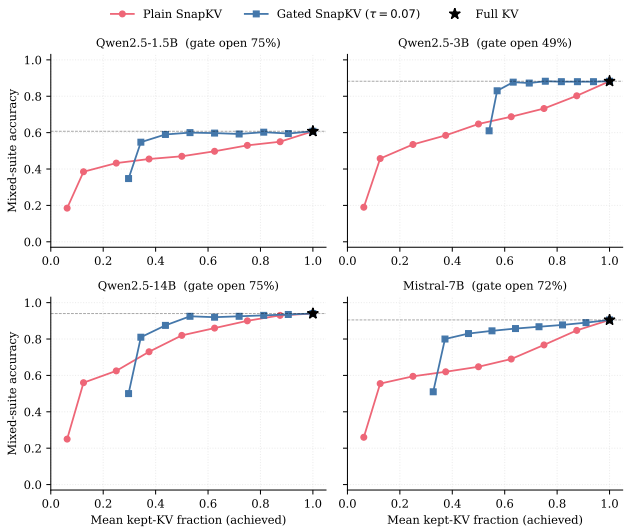


Figure 4: Accuracy vs. *achieved* kept-KV fraction on the mixed 4K suite, SnapKV. The gate falls back to the full cache when it closes, so gated points sit at larger kept-KV than their nominal budget and the honest comparison is at matched achieved cache. Above a target of $\approx 1/3$ kept-KV (Qwen3B: 0.57) the gated curve is the Pareto frontier by up to +19pp; at the leftmost gated point plain eviction dominates it. Gated results use the post-hoc $\tau = 0.07$ reconstruction (Appendix G).

Comparison with Published Methods and a Pre-Registered 32K Test Wrapping CapKV (Yang et al., 2026a) on LongBench yields +1.5/+8.3pp on two subtasks ($N = 68$ and 24), though sample sizes are small and no formal claim is made. Wrapping pretrained DBTrimKV (Bui et al., 2026) improves matched-budget accuracy by +23.3pp, but the gate fires on only 1.7% of records, so at matched *memory* plain DBTrimKV wins. PAGE therefore does not advance the frontier over a strong trained method at matched memory, and its value lies in the label-free per-input decision to retain the full cache when eviction is unsafe. Details appear in Appendix A.4. A pre-registered out-of-range test at 32K (Qwen2.5-1.5B, $N = 100$) fits $\rho/(1 - A_{\text{full}})$ on 4K/16K and hashes predicted intervals before sweeping. One of three predictions lands strictly inside its interval while both misses over-recover in different ways. What survives is ordinal rather than quantitative, as MK3's ratio stays far below VT's at every context length even though constant-ratio proportionality fails. We accordingly claim a directional relation rather than a quantitative law (Table 23, Appendix A.5).

5 Mechanistic Interpretation and Scaling Law

Dilution Mechanism We adopt the dilution mechanism of Bui et al. (2026) directly and claim no theoretical novelty. For a head reading a relevant set $\mathcal{R}$ against noise $\mathcal{N}$ with pre-eviction signal mass $\alpha_{\mathcal{R}}$, near-tie distractors force the dilution δ_t large and preferential retention reduces it. Their Proposition 3.1 establishes that near-tie distractors force attention dilution in a single head, and Corollary 3.2 shows preferential retention of high-attention positions reduces it. The partition follows from the mechanism: capacity-bound tasks concentrate the signal ($\alpha_{\mathcal{R}} \to 1$), so imperfect scoring can only *lose* signal and eviction strictly degrades the head; dilution-prone tasks spread it ($\alpha_{\mathcal{R}} \ll 1$), so eviction of noise positions monotonically improves the head Signal-Noise Ratio (SNR).

Scaling Relation Combining the dilution mechanism with the headroom $1 - A_{\text{full}}(M, T)$ gives the scaling formula:

$$\rho_{\text{KV}}(\mathsf{T}, M, T) \;\approx\; \big(1 - A_{\text{full}}(M, T)\big) \cdot p_{\text{recov}}\big(\rho_{\mathcal{R}}^2/\gamma\big), \qquad p_{\text{recov}} : [0, \infty) \to [0, 1], \tag{4}$$

where $\rho_{\mathcal{R}}$ is the retained fraction of relevant mass and γ the noise-retention ratio. The alignment between D and the partition is empirical: D ranks NIAH-MK3 smallest in every Qwen and Mistral cell (Table 26), and $\rho/(1 - A_{\text{full}})$ fits a through-origin slope on dilution-prone tasks while remaining

near zero on NIAH-MK3 regardless of headroom (Table 27, Figure 9). The pre-registered 32K test preserves the ordinal structure of these results, yet it disproves the constant-ratio proportionality implied by a direct interpretation of Eq. (4).

The theoretical bridge from D to δ_t requires a margin-to-noise ratio roughly three times the empirically observed values. The sufficient-condition analysis therefore motivates the mechanism, but does not validate the deployed gate. Consequently, the paper's empirical claims do not depend on this theoretical certification, instead rest on the observed partition and the direct gating results. We provide full derivations in Appendices B, C, and D.

6 LIMITATIONS

Architectural scope D presumes softmax attention over a materialized key-value cache and does not apply to linear-attention or state-space architectures. Attempting to run it on layers that expose no key-value cache in the standard form fails outright rather than degrading, which bounds the family of models these results speak to.

The gate is inactive on dilution-prone-only workloads On the dilution-prone LongBench subset (qasper, triviaqa, trec, multifieldqa_en) the pooled gated Δ is within $\pm$1pp of plain on all four models (Qwen2.5-1.5B 0.000, Qwen2.5-3B $+0.002$, Mistral-7B $+0.003$, Qwen2.5-14B -0.008). The partition predicts neutrality where a workload contains no capacity-bound inputs, and we observe it, including the mildly negative tail.

One-sided prediction and remaining scope gaps On LongBench passage_count at Qwen2.5-14B the gate closes on all inputs (mean $D = -0.013$) yet eviction is empirically safe, producing a harmless false positive that costs compression rather than accuracy. The symmetric failure arises at the single-distractor boundary: NIAH-MK2 is empirically capacity-bound yet its mean $D = 0.122$ co-locates with genuinely dilution-prone tasks, so the gate opens and eviction proceeds as a false negative that the mean-D threshold cannot resolve without per-input recovery information (Appendix A.2.5). D is therefore a one-sided predictor that is reliable at detecting dilution-prone inputs but not a perfect classifier. A related consequence for deployment is that the gate-open rate p_{open} is workload-dependent and not known before running, so static provisioning cannot size the cache from p_{open} in advance and must either profile it per workload or plan against the $1 - p_{\text{open}}$ resident floor. Multilingual and code tasks remain out of scope.

Threshold portability The portability of the empirical partition depends on the fixed threshold that operationalizes it. On Yi-1.5-9B and Llama-3.1-8B, the drop ordering transfers but $\tau = 0.07$ does not. It mis-classifies niah_multivalue on both models and inverts the drop-vs-ρ prediction on Yi. A z-scored variant repairs this with an unlabeled pilot, but it remains a calibration step, and per-input separation stays at AUC ≈ 0.80. The residual gap traces to interior profile structure that the early-minus-late difference cannot distinguish. The statistic D summarizes the per-layer agreement sequence a_ℓ by its endpoints, discarding the shape between them. A statistic that reads the full depth-profile rather than just its endpoints would be a stronger per-input predictor and is an open problem (Appendix A.3.3).

7 CONCLUSION

We show that KV-cache eviction is heterogeneous: inputs fall into a capacity-bound regime where compression causes severe degradation and a dilution-prone regime where selective retention can preserve or improve accuracy. The early-to-late head-agreement drop D separates these regimes from the prefill alone, supporting this claim strongly on Qwen and Mistral and partially across architectures with per-model calibration. PAGE wraps any base evictor in a gate on D and prevents catastrophic failures: on Mistral-7B NIAH-MK3 a $99\% \rightarrow 0\%$ collapse under plain eviction becomes a flat 89% under the gate, with the same protection on realistic lcc code completion and a lifted accuracy-memory frontier where moderate compression is the target. These gains come with real costs: the safeguard requires one attention-exposing prefill pass, its memory benefit is bounded by $1 - p_{\text{open}}$ and decays toward $1\times$ at batch 16 under static provisioning. The open problem it leaves is a signal that transfers across model families without a per-model pilot.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs. We formulated every mathematical claim and supplied the ingredients of its argument, and AI assistance was limited to writing up the resulting steps. Among the remaining tasks carrying required disclosure, we did not use generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean or reformat datasets, or interpret results. Translation and qualitative or thematic data analysis are not applicable to this work. Among tasks carrying recommended disclosure, we used generative AI tools to draft parts of this paper and to edit the manuscript for readability. We have reviewed all AI-assisted work: every proof in Appendices C, D, and E was checked line by line by the authors against the assumptions and scope statements of Appendix B, and all AI-drafted prose was revised by the authors and verified against the logs and tables it describes. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code and per-input logs are available at `https://anonymous.4open.science/r/PAGE-018239`. All runs use NVIDIA A100 80GB GPUs with PyTorch 2.11, Transformers 5.9, and Python 3.13, in bfloat16 with greedy (argmax) decoding, `max_new_tokens` $= 128$, and a fixed seed. Models are the public Hugging Face checkpoints of Qwen2.5-1.5B/3B/14B/32B-Instruct, Mistral-7B-Instruct-v0.3, Yi-1.5-9B-Chat, Llama-3.1-8B-Instruct, and Qwen3-4B-Instruct-2507, and tasks come from `simonjegou/ruler` (Jegou, 2024; Devoto et al., 2025) and LongBench (Bai et al., 2024).

The remaining material is in the appendix. Appendix G collects the implementation details needed for exact reproduction, including the cache-slicing and long-context conventions and the layer-binning convention behind the reported values of D. Appendix A.4 describes how inputs are drawn and how each experiment is constructed, and Appendix A.3.6 documents the diagnostic wrappers together with the code-path and hardware effects measured on them, including one wrapper that failed validation and whose numbers are therefore not reported. Variation is bounded in two places. Appendix A.6.3 reports three input-draw replicates on five SnapKV cells, and Appendix A.6.2 gives confidence intervals for the headline numbers. Each table caption states its per-cell sample sizes. The two published head-to-heads run in the baselines' own pinned environments, which Appendix A records with the released logs.

For the theoretical claims, Appendix B states the mechanism, the assumptions, and the notation, with full proofs in Appendices C, D, and E, and sharpened treatments of three deferred items in Appendix F. Each part ends with a scope statement saying what it does and does not establish. The gate is one thresholded scalar (Eq. (3)) and decoding is greedy, so every gated number in the paper can be recomputed from the released per-input logs without rerunning a model.

ETHICS STATEMENT

This work studies inference-time efficiency of publicly released language models on public benchmarks. It involves no human subjects, no personal data, and no new model training. We see no ethical risk specific to this method beyond those of the underlying models.

REFERENCES

Jinwoo Ahn, Ingyu Seong, Akhil Kedia, Junhan Kim, Hyemi Jang, Kangwook Lee, and Yongkweon Jeon. LookaheadKV: Fast and accurate KV cache eviction by glimpsing into the future without generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=RVLMGPXt2i`.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu

Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2025. URL https://arxiv.org/abs/2403.04652.

Samhruth Ananthanarayanan, Ayan Sengupta, and Tanmoy Chakraborty. Understanding the physics of key-value cache compression for llms through attention dynamics, 2026. URL https://arxiv.org/abs/2603.01426.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Ngoc Bui, Hieu Trung Nguyen, Arman Cohan, and Rex Ying. Make each token count: Towards improving long-context performance with kv cache eviction, 2026. URL https://arxiv.org/abs/2605.09649.

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. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=ayi7qezU87.

Ting-Yun Chang, Harvey Yiyun Fu, Deqing Fu, Chenghao Yang, Jesse Thomason, and Robin Jia. Value-aware stochastic KV cache eviction for reasoning models. In *COLM 2026 Workshop on Efficient Reasoning*, 2026. URL https://openreview.net/forum?id=HysrOiWFIA.

Alex Chen, Renato Geh, Aditya Grover, Guy Van Den Broeck, and Daniel Mingyi Israel. The pitfalls of KV cache compression. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41530–41553, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1926. URL https://aclanthology.org/2026.acl-long.1926/.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: fast and memory-efficient exact attention with io-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Debajyoti Datta, Trishala Neeraj, Bibek Paudel, Vyom Sharma, and Subhabrata Mukherjee. ManifoldKV: Training-free KV cache compression via euclidean outlier detection. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=fo77TzkulW.

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

Zican Dong, Peiyu Liu, Junyi Li, Zhipeng Chen, Han Peng, Shuo Wang, and Xin Zhao. ForesightKV: Optimizing KV cache eviction for reasoning models by learning long-term contribution. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=znV8JHv8b8.

Shaoting Feng, Yuhan Liu, Hanchen Li, Xiaokun Chen, Samuel Shen, Kuntai Du, Zhuohan Gu, Rui Zhang, Yuyang Huang, Yihua Cheng, Jiayi Yao, Qizheng Zhang, Ganesh Ananthanarayanan, and Junchen Jiang. Evicpress: Joint kv-cache compression and eviction for efficient llm serving, 2025a. URL https://arxiv.org/abs/2512.14946.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=tcisuhGsQZ.

Yuan Feng, Haoyu Guo, Junlin Lv, S Kevin Zhou, and Xike Xie. DefensiveKV: Taming the fragility of KV cache eviction in LLM inference. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=nJgS06sX3O.

Yuan Feng, Junlin Lv, Haoyu Guo, Yukun Cao, S Kevin Zhou, and Xike Xie. CriticalKV: Optimizing KV cache eviction from an output perturbation perspective. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=vMwlcLyI4b.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LeatkxrBCi.

Gabriel Garcia. Protection is (nearly) all you need: Structural protection dominates scoring in globally capped kv eviction, 2026. URL https://arxiv.org/abs/2605.18053.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein,

Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

T. E. Harris. A lower bound for the critical probability in a certain percolation process. *Mathematical Proceedings of the Cambridge Philosophical Society*, 56(1):13–20, 1960. doi: 10.1017/S0305004100034241.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In

First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Long Hu, Nan Jia, Rui Wang, Jiahui Li, Qingyi Yang, Yixue Hao, Xianzhi Li, and Xiaofei Liao. TAKE: Task-aware chunked KV cache eviction for efficient long-context LLM prefill, 2026. URL https://openreview.net/forum?id=kMLfUshPwo.

Simon Jegou. Ruler. Hugging Face Datasets, 2024. URL https://huggingface.co/datasets/simonjegou/ruler. Dataset.

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, 2023. URL https://arxiv.org/abs/2310.06825.

Malik Khalaf, Yara Shamshoum, Nitzan Hodos, Yuval Sieradzki, and Assaf Schuster. Anchorkv: Anchor-residual kv cache compression, 2026. URL https://arxiv.org/abs/2608.02901.

Jang-Hyun Kim, Dongyoon Han, and Sangdoo Yun. Fast kvzip: Efficient and accurate llm inference with gated kv eviction, 2026. URL https://arxiv.org/abs/2601.17668.

Steven Kolawole and Virginia Smith. Epiphany-aware kv cache eviction without the attention matrix, 2026. URL https://arxiv.org/abs/2606.26472.

Xuan Li, Yining Wang, Yuchen Liu, Guanjun Liu, Delai Qiu, Shengping Liu, Jiaen Liang, Wei Huang, Jun Yu, and Junnan Zhu. Later: Efficient test-time reasoning via latent exploration and explicit verification, 2026a. URL https://arxiv.org/abs/2605.07315.

Yu Li, Binxu Li, and Tian Lan. Momentkv: Closing the directional gap in kv cache eviction for long-context inference, 2026b. URL https://arxiv.org/abs/2606.01563.

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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=poE54GOq2l.

Xiaolin Lin, Jingcun Wang, Olga Kondrateva, Yiyu Shi, Bing Li, and Grace Li Zhang. Compresskv: Semantic-retrieval-guided kv-cache compression for resource-efficient long-context llm inference, 2026a. URL https://arxiv.org/abs/2606.24467.

Yuping Lin, Jiayuan Ding, Yue Xing, Pengfei He, Jiliang Tang, and Subhabrata Mukherjee. A simple plug-in for improving eviction-based kv cache compression, 2026b. URL https://arxiv.org/abs/2605.23258.

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

Luca Moschella, Laura Manduchi, and Ozan Sener. Learning to evict from key-value cache. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=0OevIlRMYN.

Junyoung Park, Dalton Jones, Matthew J Morse, Raghavv Goel, Mingu Lee, and Christopher Lott. Keydiff: Key similarity-based KV cache eviction for long-context LLM inference in resource-constrained environments. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uBaFH7aQnC.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,

Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Ziyao Tang, Pengkun Jiao, Xinhang Chen, Wei Liu, Shiyong Li, and Jingjing Chen. Predicting future utility: Global combinatorial optimization for task-agnostic KV cache eviction. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FQLxcBsKIb.

Guangtao Wang, Shubhangi Upasani, Chen Wu, Darshan Gandhi, Jonathan Lingjie Li, Changran Hu, Bo Li, and Urmish Thakker. LLMs know what to drop: Self-attention guided KV cache eviction for efficient long-context inference. In *Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference*, 2025. URL https://openreview.net/forum?id=qg9dlCcNzr.

Hao Wen, Xinrui Wu, Yi Sun, Feifei Zhang, Liye Chen, Jie Wang, Yunxin Liu, Yunhao Liu, Ya-Qin Zhang, and Yuanchun Li. Budgetthinker: Empowering budget-aware llm reasoning with control tokens, 2025. URL https://arxiv.org/abs/2508.17196.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EytBpUGB1Z.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Zhanchao Xu, Haoyang Li, Qingfa Xiao, Fei Teng, Chen Jason Zhang, Lei Chen, and Qing Li. From rigid to dynamic: Entropy-guided adaptive inference for long-context llms, 2026. URL https://arxiv.org/abs/2606.09508.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jiaming Yang, Chenwei Tang, Liangli Zhen, and Jiancheng Lv. Rethinking KV cache eviction via a unified information-theoretic objective. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=2fPbFUJ87k.

Ning Yang, Chengzhi Wang, Yibo Liu, Baoliang Tian, and Haijun Zhang. Compilerkv: Risk-adaptive kv compression via offline experience compilation, 2026b. URL https://arxiv.org/abs/2602.08686.

Xintong Yang, Hao Gu, Binxing Xu, Lujun Li, Bei Liu, Jiacheng Liu, Qiyuan Zhu, Yike Guo, and Sirui Han. Indexmem: Learned KV-cache eviction with latent memory for long-context LLM inference. In *Forty-third International Conference on Machine Learning*, 2026c. URL https://openreview.net/forum?id=SKXkGLbXdB.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question

answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Zihao Zeng, Bokai Lin, Tianqi Hou, Hao Zhang, and Zhijie Deng. In-context kv-cache eviction for llms via attention-gate, 2025. URL https://arxiv.org/abs/2410.12876.

Zhiyuan Zhai, Bingcong Li, Bingnan Xiao, Ming Li, and Xin Wang. Adaptive test-time compute allocation for reasoning llms via constrained policy optimization, 2026. URL https://arxiv.org/abs/2604.14853.

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, 2026. URL https://arxiv.org/abs/2605.08234.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: heavy-hitter oracle for efficient generative inference of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Lang Zhou, Shuxuan Li, Zhuohao Li, Shi Liu, Wei-Shi Zheng, and Zhilin Zhao. Ut-aca: Uncertainty-triggered adaptive context allocation for long-context inference, 2026. URL https://arxiv.org/abs/2603.18446.

Xiabin Zhou, Wenbin Wang, Minyan Zeng, Jiaxian Guo, Xuebo Liu, Li Shen, Min Zhang, and Liang Ding. Dynamickv: Task-aware adaptive kv cache compression for long context llms. In *EMNLP (Findings)*, pp. 8042–8057, 2025. URL https://aclanthology.org/2025.findings-emnlp.426/.

A Deferred Tables, Derivations, and Secondary Experiments

A.1 Extended Related Work

Section 2 positions PAGE against the decision axes prior work optimizes. This subsection gives the full treatment, starting with the five neighbors closest to our setup before turning to concurrent, older, and adjacent work.

Concurrent 2026 diagnostics and wrappers Zhang et al. (2026) ask a similar headline question, namely when value-aware KV eviction helps. Their work is complementary to ours along a clear axis. Their score is *value-aware* (a per-input support-coupling score $\phi(x)$ over value states), whereas our signal is *attention topology*. Beyond this signal difference, their unit of analysis is the selector rather than the task, their $\phi(x)$ drives diagnostic grid splits rather than an eviction gate, and they report that diagnostic signals transfer across selectors only conditionally, which makes the single transferred threshold we report non-obvious. Chen et al. (2026) document instruction-type-dependent degradation under five eviction methods, with some instruction types becoming entirely ignored, the closest published task-level failure taxonomy. We cede the hurts-degradation observation to them and claim instead the joint helps-versus-hurts partition plus the a priori predictor. Ananthanarayanan et al. (2026) compute head-consensus levels and report how they evolve with depth, describing distinct profiles by architecture (early consensus with late diversification on one family, late convergence on another) and relating those profiles descriptively to compression tolerance. Their statistic is not purely *within-layer*: they too report a depth-wise comparison, substantially the diagnosis Appendix A.3.3 reaches independently. We differentiate instead on the three axes that do hold: per input rather than per architecture, computed online at prefill rather than offline, and thresholded to make a deployment decision rather than used as a descriptive metric. To our knowledge the statistic has not been proposed as an eviction gate before, and the ablation of Table 3 shows it separates the partition cleanly (AUC 1.000) on the selecting cell and transfers across families. On the failure side, Chang et al. (2026) (VaSE) trace catastrophic eviction failures to high-magnitude

Table 3: Gate-signal ablation on Qwen2.5-1.5B-Instruct (RULER 4K, $N = 50$ per task, one shared prefill per input). Each cheap prefill statistic is tested for separating the capacity-bound anchor (NIAH-MK3) from the dilution-prone tasks (vt, fwe, qa_1, multivalue). "Sep." = a single threshold puts the capacity-bound task mean(s) strictly on one side of all four dilution task means. AUCs are per-input and oriented per signal (0.5 = chance). "MK mono." = task means are monotone along the MK1→MK2→MK3 distractor gradient; r = Pearson correlation with D.

Signal	Sep. MK3	AUC MK3	AUC MK3+MK2	MK mono.	r w/ D
head-agreement drop D (paper)	✓	1.000	**0.859**	✓	$+1.00$
mean attention entropy (norm.)	–	0.508	0.670	–	$+0.08$
top-32 attention mass share	–	0.784	0.574	✓	$+0.22$
max single-position share	–	0.723	0.708	✓	$+0.32$
key-norm dispersion (std/mean)	–	0.570	0.560	–	$+0.01$
early−late entropy drop	–	0.805	0.575	–	-0.64

value states, an axis orthogonal to our attention-topology analysis, and ManifoldKV (Datta et al., 2026) shows that a Euclidean-outlier key scorer improves NIAH-MultiKey retention under compression, which is why we define "capacity-bound" relative to the scorer being gated (Section 4.1). SAGE-KV (Wang et al., 2025) shares our *premise*: after prefilling, the model's own attention already indicates which tokens can be dropped, read once rather than per decode step, with a head-level analysis. The difference is what is done with the premise. SAGE-KV always compresses and uses the signal to pick *which* tokens and heads to keep, whereas we use a signal from the same source to decide *whether* to evict at all, do no head-level selection, and, unlike SAGE-KV, characterize a task class on which eviction cannot be made safe by any choice of tokens. KeyDiff (Park et al., 2025) is an attention-free, FlashAttention-compatible evictor whose stated motivation is precisely that materializing attention is expensive. We address it empirically in the scorer-independence analysis (Section 4.1). KeyDiff collapses on NIAH-MK3 below 25% budget, as does every scorer we tested. Its effectiveness therefore does not confer safety on the capacity-bound class.

On the wrapper side, VECTOR (Lin et al., 2026b) and CriticalKV (Feng et al., 2026b) are plug-and-play augmentations over multiple base evictors, and both always compress under a fixed budget and adapt *which* tokens are kept, whereas our gate decides *whether* to evict at all, a priori from a label-free attention signal. The serving system EvicPress (Feng et al., 2025a) also chooses whether to evict per context, but as part of a profiled quality-latency optimization over storage tiers rather than from an attention statistic, and without a task-type partition. Cross-family transfer of eviction control is itself known through CompilerKV's (Yang et al., 2026b) portable per-head retention tables. We instead transfer one fixed threshold across the Qwen and Mistral families, extended to Llama-architecture models and Qwen3 via an unlabeled per-model standardization.

Further concurrent evictors and allocators, EntropyInfer (Xu et al., 2026), CompressKV (Lin et al., 2026a), MomentKV (Li et al., 2026b), EpiKV (Kolawole & Smith, 2026), DefensiveKV (Feng et al., 2026a), LookaheadKV (Ahn et al., 2026), and TAKE (Hu et al., 2026), adapt budgets or scores per head, layer, token, or task, but none gates the eviction decision per input or identifies a task class where eviction hurts monotonically. AnchorKV (Khalaf et al., 2026) is also "safety-aware" but in a different sense from ours: it biases retention scores away from harmful-prompt directions for jailbreak defense while always compressing, whereas our "collapse-avoiding" sense is about not destroying task accuracy. Learned gates (Attention-Gate (Zeng et al., 2025), Fast KVzip (Kim et al., 2026)) share the "gate" name but train their gating parameters, unlike our inference-time scalar. A separate line uses learnable per-layer gates for cross-model KV-cache *fusion* (C2C (Fu et al., 2026)), which addresses communication between models, orthogonal to eviction within one.

Beyond the concurrent methods above, an older set of baselines and training-based predictors anchors the comparison. SnapKV (Li et al., 2024), H2O (Zhang et al., 2023), StreamingLLM (Xiao et al., 2024), PyramidKV (Cai et al., 2025), and Ada-KV (Feng et al., 2025b) are the standard training-free eviction baselines. We use SnapKV-style scoring as the base for our gate and demonstrate orthogonality on H2O, StreamingLLM, and PyramidKV. ForesightKV (Dong et al., 2026) and Learning-to-Evict (Moschella et al., 2026) train predictors of future utility and require labels or rollouts to do so. UT-ACA (Zhou et al., 2026) and LU-KV (Tang et al., 2026) are calibration-based

context allocators on a single substrate. None of these proposes a binary task-type partition or an a priori predictor.

Finally, three methods adapt compute on axes adjacent to but distinct from ours: AdaCompute-GBM (Zhai et al., 2026) on the self-consistency sample-count axis, BudgetThinker (Wen et al., 2025) on the CoT chain-length axis, and LaTER (Li et al., 2026a) on latent-then-explicit reasoning. These work on substrates orthogonal to the KV-cache axis we study.

A.2 Additional Partition Analyses

This subsection presents the robustness checks, confound analyses, and boundary diagnostics that support the empirical partition of Section 4.

A.2.1 Saturation, budget-insensitivity, and context-length scaling

A cell below the dilution-prone threshold for two different reasons, and only saturation case is harmless. Qwen2.5-3B has $\rho \approx 0$ on VT at 4K because $A_{\text{full}} \approx 1$, and the same model reaches $\rho = 0.32$ on FWE at 16K once headroom opens. Mistral-7B repeats the pattern: at 4K it is largely saturated ($A_{\text{full}} = 1.00$ on VT, 0.82 on FWE), and at 16K VT and FWE recover strongly ($\rho = 0.32$ and 0.22, niah_multivalue $0.06 \to 0.21$). Budget-insensitivity is the informative case instead, because headroom exists and recovery does not occur. On Mistral, QA_1 has $\rho = 0.04/0.00$ at 4K/16K despite $A_{\text{full}} = 0.81/0.70$, and QA_2 has $\rho = 0.01/0.00$ despite $A_{\text{full}} = 0.50/0.56$, and QA_2 on Qwen-14B behaves the same way. These are all QA, where the residual errors resemble parametric-knowledge failures that no cache budget can repair, a phenomenon distinct from the capacity limit exhibited by NIAH-MK3.

The headroom argument predicts the behaviour as context grows, and Qwen2.5-14B tests it directly by re-entering the dilution regime at long context. At 4K, three dilution-prone tasks (VT, FWE, QA_2) sit at $\rho \approx 0$: saturation, since $A_{\text{full}} \in [0.92, 1.00]$ leaves no headroom for recovery, as Eq. (4) predicts. At 16K, two of five dilution-prone tasks cross the threshold (FWE $0.00 \to 0.20$, QA_1 $0.06 \to 0.10$), niah_multivalue stays above it ($0.08 \to 0.07$), and the capacity-bound NIAH-MK3 stays at exactly 0.00 at both scales. VT ($0.00 \to 0.03$) and QA_2 ($0.02 \to 0.03$) remain below threshold because the inputs that the model gets wrong at 4K stay wrong at every budget at 16K. Headroom opens, dilution returns, and NIAH-MK3 stays pinned: the partition behaves as the scaling formula dictates, not as task identity dictates.

A.2.2 The partition is not a task-name heuristic

Task identity and dilution-proneness could in principle be confounded, but they are not. The NIAH family contains structurally different subtasks. niah_multivalue (one key, many values) is dilution-prone in seven of the eight cells measured (Table 1), while NIAH-MK3 (three near-tie keys, one of which is the answer) is capacity-bound in all of them, and the distractor sweep (Table 5) places niah_multikey_1 (no near-tie distractors) on the dilution-prone side as well. The *near-tie distractor* structure, not the surface task category, determines the class, and the head-agreement drop in Section 3 catches the distinction. A non-RULER test-case confirms this: LongBench passage_retrieval_en (pick 1 of 30 paragraphs, $N = 100$, Qwen2.5-1.5B, prompts 10 – 16K tokens) looks NIAH-like but has no near-tie distractors, and it is evict-robust in practice. Accuracy stays flat from $b = 1.0$ to $b = 0.0625$ ($0.29 \to 0.26$, $\rho = 0.01$), while its mean drop $D = 0.105$ sits squarely in the dilution-prone cluster. The gate opens on all $100/100$ inputs. The predictor correctly classifies this surface-similar but structurally different task a priori.

A.2.3 Protection-matched ablation

The partition could be an artifact of our relatively light protection setting. Garcia (2026) argue that structural protection dominates scorer choice, so we re-ran the Qwen2.5-1.5B 4K partition tasks with their 10% prefix + 10% suffix protection ($n_{\text{sink}} = w = 410$, $N = 100$ per task). Protection overrides the budget, so all nominal $b \leq 0.1875$ share an effective floor of 824 kept tokens (≈ 0.21 of context). Under this heavy protection the partition remians unchanged. NIAH-MK3 still collapses (0.65 full-KV $\to 0.09$ at the floor, an 86% relative loss despite the floor cache being $3.3\times$ our light-protection cache at nominal $b = 0.0625$), so the capacity-bound failure is not protection-fixable. The

Table 4: Mean head-agreement drop D and gate-open fraction on four additional RULER tasks, $N = 100$ per cell, $\tau = 0.07$. niah_multiquery (four required keys, no real distractors) closes the gate on two of three models and sits nearest the anchor on the third, confirming a second capacity-bound task. qa_2 (29 real hard-negative documents), niah_single_1 (no distractors), and cwe (ten exact output tokens) all open the gate on most or all inputs, so neither distractor count nor exact-token output length predicts the class.

Task	Qwen2.5-1.5B	Qwen2.5-3B	Mistral-7B
niah_multiquery	D=0.076, open 32%	D=0.021, open 100%	D=0.042, open 100%
qa_2	D=0.213, open 0%	D=0.097, open 10%	D=0.086, open 18%
niah_single_1	D=0.153, open 0%	D=0.061, open 87%	D=0.121, open 0%
cwe	D=0.200, open 0%	D=0.103, open 0%	D=0.135, open 0%

eviction-helps signal survives on the dilution-prone side (ρ: FWE $0.06 \to 0.13$, QA_1 $0.08 \to 0.07$, VT $0.06 \to 0.03$, aggregate $0.077 \geq 0.05$), and FWE's floor accuracy (0.31) exceeds its full-KV accuracy (0.22). Heavy protection is also not free: protected VT drops to 0.00 at the floor while our light-protection run holds 0.75 at nominal $b = 0.125$, because protected prefix/suffix tokens crowd out the scorer-selected middle tokens the task requires. In our harness w is both the recency protection and the SnapKV scoring window, so this run is a protection-heavy SnapKV variant rather than an exact replication of their protocol.

Table 5: Distractor-count sweep on the NIAH-MultiKey family (Qwen2.5-1.5B, RULER 4K, $N = 50$ per task). Mean head-agreement drop D shrinks monotonically as the near-tie distractor count grows.

Task	mean D	gate-open frac	ρ_{plain}	ρ_{gated}	plain at b=0.0625
niah_multikey_1	+0.155	1.00	0.04	0.04	0.36
niah_multikey_2	+0.122	1.00	0.00	0.00	0.00
niah_multikey_3	+0.045	0.00	0.00	0.00	0.00

A.2.4 Boundary and confound analysis: full table

A single capacity-bound task leaves open whether NIAH-MK3 names a class or an isolated construction. We tested the boundary directly on four further RULER tasks chosen by mechanism rather than by outcome: a second near-tie retrieval task predicted capacity-bound, a real-hard-negative task predicted dilution-prone despite carrying more distractors than any other task we tested, a single-needle control with no distractors at all predicted dilution-prone, and a code-completion task requiring ten exact low-redundancy output tokens predicted capacity-bound. Table 4 reports the outcome at $\tau = 0.07$, $N = 100$ per cell.

The results confirm that the partition is not defined by distractor count or output length. niah_multiquery is capacity-bound on two of three models, giving the class a confirmed second member. qa_2 carries the most distractors of any task tested yet sits comfortably on the dilution-prone side. What separates the two classes is whether distractors are near-tie surface-form variants of the answer: the distractor sweep isolates this directly (Table 5). The class has two entry conditions: a near-tie surface-form distractor, or a conjunctive multi-key answer. Both conditions concentrate the answer into a small, non-redundant token set, which is the property the gate responds to.

A.2.5 Near-tie distractor count drives D

The capacity-bound classification rests on a mechanism claim: that near-tie distractors, not task identity, distractor count, or output length in general, drive D downward. Section A.2 already shows that real hard-negative distractors do not have this effect, since qa_1 and qa_2 carry the most distractors among the tasks we tested and exhibits the largest D (Table 26). Exact-token output length also does not, since cwe requires ten exact low-redundancy tokens yet keeps the gate open on every input tested (Table 4). This subsubsection isolates the near-tie case directly. The RULER NIAH-MultiKey family lets us sweep the near-tie distractor count with the number of documents

Table 6: Gate-signal ablation repeated on held-out cells. "sep." counts the signals whose mean on NIAH-MK3 lies strictly below their mean on every dilution-prone task, which is the property the gate thresholds, and "margin" is that gap for D. "margin$^-$" excludes niah_multivalue, the exclusion Appendix A.3.3 places on the Llama-architecture ordering claim. AUC is per-input separability of NIAH-MK3 from the pooled dilution-prone tasks. The wrapper column records which memory workaround produced the cell (Appendix A.3.6).

Cell	Status	Wrap.	D MK3	min dil.	margin	margin$^-$	AUC(D)	sep.
Qwen2.5-1.5B 4K	fitting	stock	0.0448	0.1114	$+0.0667$	$+0.0667$	1.000	1/6
Qwen2.5-14B 4K	held out	v1	0.0455	0.0487	$+0.0032$	$+0.0436$	0.899	1/6
Llama-3.1-8B 4K	held out	v1	0.0710	0.0607	$\mathbf{-0.0103}$	$+0.0073$	0.741	**0/6**
Mistral-7B 16K	held out	v3	0.0544	0.0756	$+0.0212$	$+0.0393$	0.978	1/6

held fixed: niah_multikey_1 (no near-tie distractors), niah_multikey_2 (one), niah_multikey_3 (two). The dilution mechanism of Bui et al. (2026) predicts that more near-tie distractors should produce larger expected dilution δ_t and, via App. D, smaller head-agreement drop D. On Qwen2.5-1.5B at RULER 4K, $N = 50$ per task, we observe exactly this: a clean monotone gradient (Table 5).

D shrinks by $\sim 3.5\times$ from $0 \to 2$ near-tie distractors, and per-input ranges between MK2 and MK3 do not overlap ($\max_{\text{MK3}} D = 0.062 < \min_{\text{MK2}} D = 0.088$). The partition predictor separates 2-distractor inputs from ≤ 1-distractor inputs 100% of the time at $\tau = 0.07$. The monotone gradient replicates on Mistral-7B-v0.3 ($D = 0.081, 0.072, 0.060$ for MK1/MK2/MK3, $N = 50$ per task), confirming the mechanism is not Qwen-specific, though the finer per-input MK2/MK3 non-overlap is not reproduced (their ranges overlap on Mistral).

This clean separation between 0 and 2 distractors leaves open what happens at the single-distractor boundary, and MK2 is exactly that case. It is empirically capacity-bound at low budgets (plain SnapKV collapses from 80% at b=1 to 0% at b=0.0625) yet is classified as dilution-prone by the predictor at $\tau = 0.07$: MK2's mean $D = 0.122$ sits between the MK3 cluster ($D \leq 0.062$) and the MK1 cluster ($D \geq 0.099$), co-located with genuinely dilution-prone FWE ($D = 0.098$) and niah_multivalue ($D = 0.107$). No single mean-D threshold separates MK2 from the dilution-prone cluster on this model, so the predictor falls back to its weakest form here. It correctly classifies MK3 and the standard dilution-prone tasks but mis-classifies single-near-tie MK2. Per-input recovery information beyond mean D would be needed to catch it at deployment.

A.3 Signal and Method Details

A.3.1 Gate-signal ablation: extended analysis

Table 3 gives the headline comparison; Section 3 summarises the key findings. This subsection provides the extended analysis of the signal's two free parameters.

The signal's two free parameters D depends on a top-k parameter and a bin fraction. The justification for $k = 32$ was that it is commensurate with $|\mathcal{R}|$, which is the quantity the gate exists to avoid knowing, so an empirical check tests that choice. Over $k \in \{8, 16, 32, 64, 128\}$ on the selecting cell and on a held-out Mistral-7B cell ($N = 400$ inputs each), $k = 32$ maximises both the separation margin and the per-input AUC on *both* cells (Qwen2.5-1.5B: margin $+0.068$, AUC 1.000; Mistral-7B: margin $+0.020$, AUC 0.977). This empirical validation replaces the original heuristic. The choice is not robust upward. The MK3-lowest ordering on which the gate depends breaks at $k = 128$ on Qwen2.5-1.5B and at $k \geq 64$ on Mistral-7B.

The bin fraction is the less sensitive parameter, but only where D works at all. Recomputing D with early and late bins at $1/4$, $1/3$ and $1/2$ of the layer stack, from the stored per-layer agreements at no additional inference cost, leaves the task ordering intact on every Qwen cell we retained. The margin narrows as the bins widen but never changing sign (Qwen2.5-1.5B 4K: $+0.084$, $+0.068$, $+0.041$; Qwen2.5-3B 4K: $+0.047$, $+0.042$, $+0.025$; Qwen2.5-3B 16K: $+0.087$, $+0.067$, $+0.045$). On the two Llama-architecture cells the ordering is inverted at *every* bin fraction (Llama-3.1-8B -0.002, -0.011, -0.024; Yi-1.5-9B -0.033, -0.031, -0.032). This is the same transfer failure Section 4.1 reports for the fixed τ on those families rather than a separate defect of the binning. No choice of

bin fraction repairs it. The thirds convention is therefore not load-bearing where the signal transfers, and is not the reason the signal fails where it does not. The Mistral per-layer logs are corrupt and excluded, so this sweep covers five of the seven cells.

One methodological point falls out of the k sweep. Sweeping k at the deployed fixed $\tau = 0.07$ inverts the conclusion. $k = 8$ then posts $\Delta = +0.239$ on Qwen2.5-1.5B and $+0.391$ on Mistral-7B, apparently beating $k = 32$, while its gate-open fraction is 0.00. The gate has silently become full-KV fallback, and the apparent gain is the full cache. The location and scale of D move with k (pooled mean D is $+0.138$ at $k = 32$ but negative at $k = 8$ and $k = 128$), so any sweep of a parameter that rescales D must recalibrate τ alongside it, or it ablates the threshold rather than the parameter. This is the same failure mode as the Qwen3 fixed-τ cell of Section 3, and we report gate-open fractions beside every Δ in the sweep for that reason.

A.3.2 Held-out signal ablation

Table 3 is measured on the cell D was selected. An AUC of 1.000 there is the number a reader should trust least. We therefore repeat the ablation on three cells that played no part in selecting the statistic (Table 6), and the result splits into what survives held-out testing and what does not.

What survives is the qualitative claim: D remains the only one of the six statistics that achieves task-level separation anywhere, and it does so on two of the three held-out cells. Every alternative fails on every cell, held out or not, so the claim Table 3 makes now rests on more than the measurement that selected the signal.

What does not survive is the fitting cell's exact number. The per-input AUC of 1.000 is a property of that cell alone. Held out, it is 0.899, 0.978, and 0.741, so we report 1.000 as a fitting-cell figure rather than as the separability of D in general.

The two weak held-out cells share a common cause: both are decided by a single task, niah_multivalue. On Llama-3.1-8B, NIAH-MK3 sits at $D = 0.0710$ while niah_multivalue sits at 0.0607, giving an unrestricted margin of -0.0103. Consequently, D separates on two of the three held-out cells, not three, once niah_multivalue is counted. The other three dilution-prone tasks lie above MK3, and against the nearest of them (VT at 0.0783) the margin is $+0.0073$ if niah_multivalue is excluded. That figure is the one Appendix A.3.3 derives from the depth profile, reached here from an independent measurement and it already places this exclusion on the task-level ordering claim for a stated reason: niah_multivalue is eviction-fragile on Llama-architecture models, so a low D there is arguably the correct call rather than a mis-ordering. We did not register that exclusion before running this ablation, so we report it here as a post-hoc diagnostic rather than a confirmed prediction, and Table 6 gives both margins so either case is checkable. On Qwen2.5-14B the margin of $+0.0032$ is likewise set by niah_multivalue at 0.0487, and excluding it the margin is $+0.0436$ against FWE.

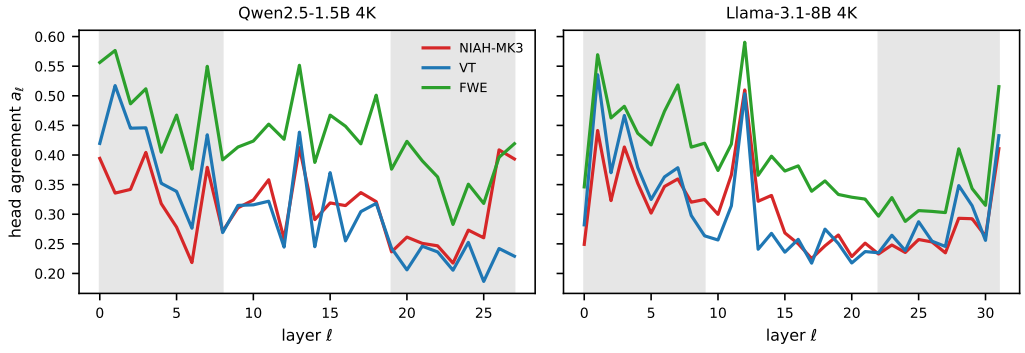


Figure 5: Per-layer head agreement a_ℓ, averaged over inputs ($N = 100$ per task, RULER 4K). Shaded bands are the early and late thirds that define D. Left: on Qwen2.5-1.5B the profile decays with depth and the capacity-bound task holds the highest late plateau, as (A3) predicts. Right: on Llama-3.1-8B the profile is non-monotone and NIAH-MK3 is nearly indistinguishable from VT under the same difference, which is why the fixed τ does not transfer to that family.

A.3.3 THE PER-LAYER PROFILE BEHIND D

The small Llama-3.1-8B margin above is a symptom. This subsubsection diagnoses the cause by examining the per-layer profile that D summarizes. Assumption (A3) in App. D asserts a layer structure: early layers respond to local context, late layers to task structure, which is why an early-minus-late difference should separate the two classes. The paper elsewhere reports only the binned scalar D. We plot the profile a_ℓ itself (Figure 5). It is recomputed from the released per-layer logs at zero GPU cost.

On Qwen2.5-1.5B the profile behaves as (A3) describes. Agreement decays with depth, and NIAH-MK3 holds a markedly higher late plateau (0.283) than the dilution-prone tasks (0.227 on VT). The difference separates them by $+0.068$. On Llama-3.1-8B it does not. The profile is non-monotone, with a pronounced mid-stack spike, and NIAH-MK3 and VT become nearly indistinguishable under the difference ($D = +0.071$ against $+0.078$, a margin of $+0.007$, an order of magnitude smaller than on Qwen).

We report this because it is the mechanistic content behind a failure the paper otherwise leaves unexplained. The fixed threshold does not transfer to the Llama architecture. Per-input AUC there is ≈ 0.80 rather than 1.000, not because the signal is absent but because the depth profile it summarises has a different shape. An early-minus-late difference is the right summary of a monotone profile and the wrong summary of a profile with interior structure, which is also why the depth-correlation statistic comes out anti-correlated on Llama rather than merely weak. A statistic that reads the profile shape rather than its endpoints is the natural next step, and this is the concrete form of the "per-input statistic beyond mean D" that Section 6 names as future work. We attempted it, from the stored per-layer agreements at no additional inference cost, and report the outcome because it is informative in both directions.

Three endpoint-free variants of the profile were tested against D on five cells: the late-plateau level alone, the regression slope of a_ℓ on depth, and the fraction of adjacent layer pairs that decrease. The late-plateau level, which the difference between MK3 at 0.283 and VT at 0.227 above suggests as the natural candidate, *fails*: its per-input AUC is at or below chance on every cell (0.412, 0.201, 0.203, 0.006, 0.298), and on Llama-3.1-8B it is almost perfectly anti-correlated. The observation that motivated it is real on Qwen and does not generalise. The regression slope matches D but never exceeds it (1.000, 1.000, 1.000, 0.714, 0.800 against D's 1.000, 1.000, 1.000, 0.786, 0.803), which indicates the endpoint difference is already extracting most of what the profile carries. A monotonicity statistic reads higher than D on one cell (Yi-1.5-9B, 0.912 against 0.803) and at chance on another (Llama-3.1-8B, 0.552). At $N = 50$ capacity-bound inputs per cell, with three statistics tested across five cells and no correction for that multiplicity, we treat the single high value as unexplained rather than as a result, and do not build on it.

The conclusion we draw is narrow and negative: none of the three shape statistics transfers uniformly, and none improves on D in a way that replicates, so the open problem of Section 6 stands unresolved.

A.3.4 CALIBRATION RECIPE DETAILS

Section 3 summarizes the calibration recipe. This subsubsection gives the full recipe, its validation, its sensitivity to τ, and the architecture-normalized variant.

Two senses of "label-free" PAGE never observes correctness, making it *accuracy-label-free*, which is weaker than *task-label-free*. The fixed-τ recipe below needs 20 inputs from a known capacity-bound task and 20 from a known dilution-prone one. The z-scored variant assumes neither, and we use the stronger term only for it.

Recipe For a new (model, context):

1. Pick 20 inputs from a known capacity-bound task (NIAH-MK3) and compute the head-agreement drop D_i on each. Average to get $\bar{D}_{\mathcal{C}}$.
2. Pick 20 inputs from a known dilution-prone task (VT) and compute $\bar{D}_{\mathcal{D}}$.
3. Set

$$\tau = \tfrac{1}{2}\big(\bar{D}_{\mathcal{C}} + \bar{D}_{\mathcal{D}}\big). \tag{5}$$

40 prefills, no labels, no decode, < 5 minutes on a single A100.

Validation On four cells (Table 7). The recipe lifts Qwen2.5-3B 16K from $+3.4$pp to $+5.4$pp (the only failure cell turned into a win) and lifts Mistral 4K from $+18.7$pp to $+22.3$pp. It is neutral on average across the four cells but with smaller variance.

Table 7: Calibration-recipe validation. $\bar{D}_{\mathcal{C}}$ uses NIAH-MK3 ($N = 20$); $\bar{D}_{\mathcal{D}}$ uses VT ($N = 20$). Bolded τ-choice column is the winner per cell. The pilot means here are computed on 20-input samples and therefore differ slightly from the $N \geq 50$ per-task means of Table 26.

Cell	$\bar{D}_{\mathcal{C}}$	$\bar{D}_{\mathcal{D}}$	τ_{cal}	gated@τ_{cal}	gated@$\tau = 0.07$
Qwen2.5-1.5B 4K	$+0.045$	$+0.161$	$+0.103$	$+0.121$	$+0.121$
Qwen2.5-3B 4K	-0.002	$+0.076$	$+0.037$	$+0.213$	$\mathbf{+0.259}$
Qwen2.5-3B 16K	-0.021	$+0.068$	$+0.024$	$\mathbf{+0.054}$	$+0.034$
Mistral 4K	$+0.060$	$+0.096$	$+0.078$	$\mathbf{+0.223}$	$+0.187$

Table 8: The three τ variants used in the paper, their label requirements, and where each is applied, referenced from Section 3. Accuracy labels are never used by any variant. The fixed value produces all headline numbers; the sole exception is the Qwen2.5-3B 4K SnapKV matrix cell, run at $\tau = 0.04$ and reported at $\tau = 0.07$ by exact post-hoc re-evaluation (Appendix G).

Variant	Accuracy-label-free	Task-label-free	Used for
Fixed $\tau = 0.07$	✓	✓	Qwen2.5 and Mistral headline results
z-scored per-model pilot	✓	✓	Cross-family transfer (Yi, Llama, Qwen3)
40-input midpoint (Eq. (5))	✓	–	Calibration validation (Table 7)

Threshold sensitivity Sweeping $\tau \in \{0.01, 0.025, 0.04, 0.055, 0.07, 0.085, 0.10, 0.13, 0.16, 0.20\}$ on Qwen2.5-1.5B at RULER 4K (mixed suite, $N = 100$, budgets $\{0.5, 0.25, 0.125, 0.0625\}$, SnapKV) shows an operational plateau $\tau \in [0.055, 0.10]$ inside which Δ varies by under 1pp. Both false-positive (gate fires on capacity-bound MK3) and false-negative (gate stays closed on dilution-prone VT/FWE/QA_1) rates stay at or below 9%. The deployed $\tau = 0.07$ sits in the middle of this plateau (Figure 6). Two scope conditions attach to that statement. The sweep is on the cell that selected the signal, and we do not claim the plateau transfers. It is a fitting-cell measurement, reported here because the deployed constant should not be a critical threshold on the cell it came from. Δ alone is also the wrong quantity to look for a plateau in, since raising τ closes the gate more often and substitutes full-KV fallback for eviction, which raises Δ monotonically while destroying compression. We therefore read the plateau jointly with the false-positive and false-negative rates above, and report achieved kept-KV alongside Δ wherever the threshold varies. Outside it the behaviour changes character. Below 0.04, FP $\geq 64\%$ and $\Delta \leq +0.06$. Above 0.13, FN $\geq 33\%$ and the method degenerates toward "always full KV," which on this full-KV-strong task set still yields a large Δ, since the sweep's raw maximum is in fact at $\tau = 0.20$, but only at the price of never compressing, so it is not a useful operating point. The deployed threshold is not critically sensitive to the exact value.

Architecture-normalized variant: the z-scored drop The recipe above calibrates a per-cell threshold from two labeled pilot tasks. A stronger variant needs no task labels at all: standardize the drop per model, $z(x) = (D(x) - \mu_M)/\sigma_M$, with μ_M, σ_M pooled over an unlabeled ~ 100-input pilot, and use one global threshold θ_z. Fitting $\theta_z = -0.69$ on the five Qwen/Mistral cells *only* and transferring it zero-shot to the two held-out Llama-architecture models: task-level separation of NIAH-MK3 from {QA_1, QA_2, VT, FWE} succeeds in 7 of 7 cells (pooled cross-cell AUC 0.965), where raw D with one threshold manages 4 of 7. The reason is that raw D is an architecture-invariant *ordering* whose location and scale are model-specific: $\tau = 0.07$ sits at $z = +0.69$ on Yi but $z = -0.97$ on Llama. These are two opposite failure modes that standardization removes. Downstream, the z-gate lifts Llama-3.1-8B from $\Delta = +0.109$ to $\mathbf{+0.174}$ (NIAH-MK3 gate-open $0.56 \to 0.30$, MK3 accuracy $0.46 \to 0.72$) and gives Yi $\Delta = +0.145$ with *real* compression (kept-KV 0.22–0.32 versus $0.76 - 0.89$ under the fixed-τ fallback). Two honest scopes: niah_multivalue

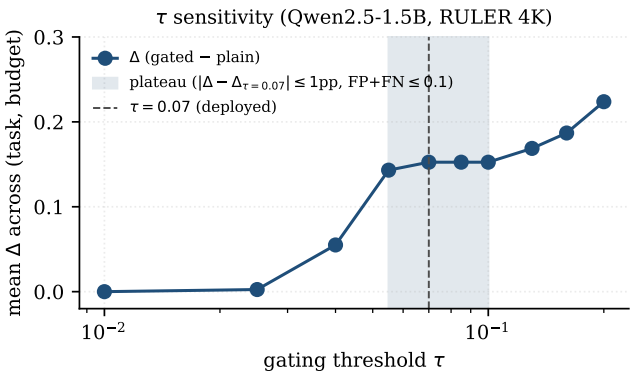


Figure 6: τ sensitivity on Qwen2.5-1.5B, RULER 4K. Mean Δ (gated $-$ plain) over (task, budget) cells. Shaded band: operational plateau where $|\Delta - \Delta_{\tau=0.07}| \leq 1$pp and FP+FN ≤ 0.10. The deployed $\tau = 0.07$ sits in the middle of the plateau.

Table 9: Head-pair subsampling (Qwen2.5-1.5B, RULER 4K, $H = 12$, 66 pairs, $L = 28$, $N = 30$ inputs per task over four tasks, 120 input-task rows in total). "Margin" is the gap between mean D on NIAH-MK3 and the smallest mean D among the dilution-prone tasks; "ordering" counts the draws in which that margin stayed positive.

Pairs kept	count	mean D MK3	min mean D dil.	margin	ordering
100%	66	0.0413	0.1131	+0.0717	1/1
50%	33	0.0404	0.1109	+0.0705	20/20
25%	16	0.0456	0.1179	+0.0724	20/20
10%	7	0.0280	0.1035	+0.0754	20/20
5%	3	0.0541	0.1206	+0.0665	20/20
2%	1	0.0529	0.1264	+0.0735	20/20

must be excluded from the task-level ordering claim (its mean D sits below MK3 on both Llama-architecture models, which is empirically the right call there, since it is eviction-fragile on both), and per-input MK3-vs-rest AUC on the Llama-architecture models is intrinsically ≈ 0.80, since no monotone rescaling changes within-model ranking. A depth-correlation statistic over the per-layer profile scores AUC 0.22 on Llama. This does not mean the depth profile is uninformative there. An AUC of 0.22 is *anti*-correlated, and inverting its sign gives 0.78, comparable to D's own ≈ 0.80 on that family. The profile carries signal on Llama with the opposite sign to the one assumption (A3) predicts, which Appendix A.3.3 shows directly.

A fourth family where fixed τ fails outright: Qwen The starkest transfer test is Qwen3-4B-Instruct, a different model generation from the Qwen2.5 calibration set, whose drop distribution is compressed to a different scale entirely ($\mu = 0.034$, $\sigma = 0.017$ pooled over 400 inputs). There the fixed $\tau = 0.07$ sits at $z = +2.0$ and the gate essentially never fires. It keeps 99.4% of the cache and its apparent $\Delta = +0.43$ is a pure full-KV-fallback artifact with 0.6% compression. The z-scored threshold ($\tau_{\text{Qwen3}} = \mu + \theta_z \sigma = 0.022$, same $\theta_z = -0.69$) restores a working gate. NIAH-MK3 stays closed (open fraction 0.02, accuracy $0.05 \rightarrow 0.98$) while the three dilution-prone tasks open ($0.94 - 1.00$), yielding $2.3\times$ compression (43% mean kept-KV) at $\Delta = +0.235$ over plain and recovering 54.5% of the per-input oracle headroom. The NIAH-MK3-smallest ordering transfers cleanly. Qwen3 is thus the cleanest evidence for the paper's central claim about the predictor: the drop is an architecture-invariant *ordering* whose scale must be standardized per model. Here fixed-τ does not merely underperform. It does nothing.

The recipe is the deployment-friendly by default as it always fixes the failure cell. The z-scored variant extends it label-free across architecture families. We report fixed-τ numbers throughout as the conservative default to show both the robustness of one constant and what standardization buys.

A.3.5 HEAD-PAIR AND LAYER SUBSAMPLING

How many head pairs does D need? Layer count is only half of the $O(LH^2k)$ cost. The other half is the pair count, and it is the half that scales badly: 66 pairs at $H = 12$ but 780 at $H = 40$. The released logs cannot answer this, because the test averages over pairs before writing, so we re-ran the test-case retaining the per-pair Jaccards and subsampled offline (Table 9). Random subsets are drawn 20 times per fraction, since a single unrepresentative draw would understate what subsampling can do.

The ordering on which the gate depends survives *every* draw down to a single pair out of 66. At 2% of the pairs the margin is $+0.0735$, against $+0.0717$ using all of them. We read this as a statement about the cost model rather than a recommendation to gate on a single pair: the H^2 term is not load-bearing, so the gate can be made substantially cheaper on exactly the wide-head models where it is most expensive, which is the deployment objection of Section 6. The caveats are that this is one model at one context length, that it is an ordering result on task means rather than a per-input one. The margin is not monotone in the fraction (the 5% row dips to $+0.0665$), which is sampling noise at 120 inputs.

How many layers does D need? The gate's cost grows with the number of layers scanned, so we ask whether a subset suffices. Recomputing D from layer subsets of the stored per-layer agreements needs no GPU. We first check validity: recomputing over all L layers reproduces the `drop` the deployed gate actually thresholded, bit-exactly on the two Qwen cells and to 4×10^{-4} on Mistral, which is two orders of magnitude below τ. Table 10 then reports whether each subset preserves the ordering the gate relies on, NIAH-MK3 ranking below every other task. Halving the layers is safe on all three cells, though the margin shrinks by roughly a quarter on Qwen2.5-1.5B and by an order of magnitude on Qwen2.5-3B. Beyond that the signal degrades and is cell-dependent: every-4th holds only on Qwen2.5-3B, the early bin alone inverts the ordering everywhere, and the single first-and-last-layer pair happens to work on Qwen2.5-1.5B while failing badly on Mistral. We therefore do not recommend subsampling below $L/2$, and we read the incidental successes as a caution against tuning the layer set per cell.

Table 10: Effect of computing D from a subset of layers rather than all of them. "Margin" is the gap between mean D on NIAH-MK3 and the smallest mean D among the other tasks; a positive margin means a separating threshold still exists on that subset, and "kept" marks whether the task ordering is preserved.

	Qwen2.5-1.5B (L=28)		Qwen2.5-3B (L=36)		Mistral-7B (L=32)
Layer subset	margin	kept	margin	kept	margin
all layers	$+0.0676$	✓	$+0.0415$	✓	$+0.0174$
every 2nd	$+0.0520$	✓	$+0.0125$	✓	$+0.0152$
every 4th	-0.0017	–	$+0.0035$	✓	-0.0194
early bin only	-0.0410	–	-0.0280	–	-0.0225
first+last layer	$+0.0660$	✓	-0.0532	–	-0.1541

A.3.6 GATE-SIGNAL DIAGNOSTIC

The diagnostic requests every layer's full $[H, T, T]$ attention, which reaches roughly 120 GiB for Qwen2.5-14B at 4K and cannot be run directly. Two wrappers make it fit. The retention wrapper (v1) calls the stock attention and slices the returned tensor, and the query-chunk wrapper (v3) chunks the query axis while calling the stock function unchanged. We validated each against the stock script on the fitting cell before using it: v1 agrees exactly ($D = 0.0412$ against 0.0412 on the same host) and v3 agrees to $+0.000029$, which is floating-point reassociation. We report this validation because a third wrapper failed the same check. Rewriting attention through SDPA moved mean D from 0.0448 to 0.0300, a 33% shift, reproducibly on two different GPUs, so the cause is the output path perturbing hidden states and compounding across layers rather than a hardware difference. No reported number uses this third wrapper. Separating that 0.0148 code artifact from the 0.0035 hardware effect required varying code path and hardware together, since either comparison alone against the released numbers confounds the two.

Table 11: Per-task Δ (gated $-$ plain, budgets $b < 1.0$, $\tau = 0.07$) for all 16 matrix cells, with the gate-open fraction on non-MK3 inputs. Rows marked † have $\Delta_{\text{non-MK3}}$ exactly 0.000 with the gate open on every non-MK3 input, so gated $\equiv$ plain there by construction. Recomputed from the released per-input logs.

Base evictor	Model	MK3	VT	FWE	QA_1	all	no MK3	open (non-MK3)
SnapKV	Qwen2.5-1.5B	+0.484	0.000	0.000	0.000	+0.121	+0.000†	1.000
	Qwen2.5-3B	+0.734	+0.004	+0.300	0.000	+0.259	+0.101	0.653
	Qwen2.5-14B	+0.573	0.000	0.000	0.000	+0.143	+0.000†	1.000
	Mistral-7B	+0.701	0.000	+0.046	0.000	+0.187	+0.015	0.923
H2O	Qwen2.5-1.5B	+0.647	0.000	0.000	0.000	+0.162	+0.000†	1.000
	Qwen2.5-3B	+0.920	+0.017	+0.277	0.000	+0.303	+0.098	0.653
	Qwen2.5-14B	+0.800	0.000	0.000	0.000	+0.200	+0.000†	1.000
	Mistral-7B	+0.883	0.000	+0.043	+0.017	+0.236	+0.020	0.923
StreamingLLM	Qwen2.5-1.5B	+0.650	0.000	0.000	0.000	+0.163	+0.000†	1.000
	Qwen2.5-3B	+0.930	+0.030	+0.760	0.000	+0.430	+0.263	0.653
	Qwen2.5-14B	+1.000	0.000	0.000	0.000	+0.250	+0.000†	1.000
	Mistral-7B	+0.890	0.000	+0.140	+0.020	+0.263	+0.053	0.923
PyramidKV	Qwen2.5-1.5B	+0.610	0.000	0.000	0.000	+0.152	+0.000†	1.000
	Qwen2.5-3B	+0.883	0.000	+0.433	0.000	+0.329	+0.144	0.653
	Qwen2.5-14B	+0.913	0.000	0.000	0.000	+0.228	+0.000†	1.000
	Mistral-7B	+0.863	0.000	+0.060	0.000	+0.231	+0.020	0.923
mean		+0.780	+0.003	+0.129	+0.002	**+0.229**	**+0.045**	0.894

A.4 Experimental Details and Secondary Results

This subsection collects the full tables, decompositions, and secondary experiments that support the experimental validation of Section 4, including the random-eviction control (Appendix A.4.9, Table 20) and the per-layer eviction ablation (Appendix A.4.10, Table 21).

A.4.1 Per-task decomposition of the matrix

Table 2 reports one number per cell; decomposing that number by task explains where it comes from. Table 11 breaks down every matrix cell by task and reports the gate-open fraction that explains the zeros, and two structural facts follow from it. First, $\Delta_{\text{non-MK3}}$ is exactly 0.000 in 8 of 16 cells, and in each of those the gate opens on 100% of non-MK3 inputs, so the gated and plain arms execute identically: the zero is an identity, not a null result. Second, the mean Δ_{MK3} is +0.780 while the grand mean is $+0.229 \approx \frac{1}{4}\Delta_{\text{MK3}}$, the MK3 share of the mixed suite. Since per-task accuracy is additive under reweighting, the headline for a workload whose capacity-bound share is f is $\Delta(f) = f \cdot \Delta_{\text{MK3}} + (1-f) \cdot \Delta_{\text{rest}}$, which is near-linear through the origin (Figure 7): $\Delta(0) = +0.045$, $\Delta(0.25) = +0.229$ (the reported suite), $\Delta(0.5) = +0.412$. A deployment therefore sees the gate's benefit in proportion to how much capacity-bound traffic it actually serves, and the realistic-workload estimate of that share is in Table 12.

Table 12: The partition on a non-synthetic workload: three LongBench subtasks, Qwen2.5-14B, SnapKV, $\tau = 0.07$. D is the mean head-agreement drop and "open" the fraction of inputs on which the gate fires. "Sensitivity" is the fraction of full-cache-correct answers plain eviction destroys at $b = 0.0625$ (the capacity signal) and ρ the fraction of inputs on which eviction recovers a full-cache failure (the dilution signal). The gate's class call is made from D alone, before decoding, and matches the empirical class on all three.

Subtask	N	A_{full}	sens.	ρ	D	open	empirical class	Δ
lcc	96	0.292	0.64	0.031	0.0543	0.250	capacity-bound	+0.146
repobench-p	82	0.317	0.42	0.061	0.0541	0.256	mixed	+0.049
hotpotqa	100	0.560	0.12	0.020	0.0973	0.860	evict-robust	+0.000

Budget convention Δ averages over the eviction budgets $b < 1.0$. Including $b = 1.0$ leaves the value unchanged because the gated and plain arms coincide there. Restricting instead to

$\{0.125, 0.25, 0.5\}$ gives $+0.240$, and to the four budgets shared by all cells $+0.243$. We report $b < 1.0$ throughout.

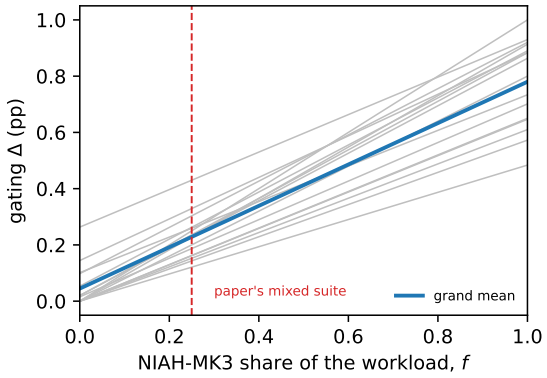


Figure 7: Gating Δ as a function of the capacity-bound share f of the workload, $\Delta(f) = f\Delta_{\text{MK3}} + (1-f)\Delta_{\text{rest}}$, mean-line slope 0.7354. Grey lines are the 16 matrix cells, the blue line their mean, and the dashed marker the $f = 0.25$ composition of our mixed suite. Because Δ_{rest} is exactly zero in half the cells, the mean is close to linear through the origin, so the headline $+22.9$pp should be read as a statement about a workload that is one quarter capacity-bound rather than as a property of the method.

A.4.2 Budget-axis decomposition

Table 13 decomposes the Qwen2.5-3B 4K SnapKV cell along the budget axis, complementing the budget-averaged Δ values reported in Table 2 and Figure 8.

Table 13: Qwen2.5-3B 4K, SnapKV, mixed suite. Plain accuracy collapses from 0.88 to 0.19 as the budget shrinks; gated holds $0.61 - 0.88$ across the entire range.

budget b	plain	gated	Δ
1.000	0.882	0.882	+0.000
0.875	0.802	0.880	+0.078
0.750	0.733	0.880	+0.147
0.625	0.688	0.880	+0.193
0.500	0.647	0.882	+0.235
0.375	0.585	0.873	+0.287
0.250	0.535	0.877	+0.343
0.125	0.458	0.830	+0.372
0.0625	0.190	0.610	+0.420

A.4.3 The task-label oracle

Section 4.1 summarizes the task-label oracle comparison. Table 14 gives the full per-cell results.

It has three parts: (i) in 8 of 16 cells PAGE reproduces the oracle *exactly*, so there the gate is a perfect task classifier and D buys label-freeness rather than per-input resolution, (ii) in 5 cells PAGE *beats* the oracle, by up to $+0.198$, which no task-level policy can do, since the surplus comes from closing on individual dilution-prone inputs that the oracle wrongly opens, (iii) in 3 Mistral cells PAGE loses by at most 0.009 through false positives on MK3. Overall the oracle reaches $+20.1$pp against PAGE's $+22.9$pp.

So most of the matrix Δ is a task-level effect that a cheating baseline could match, and we say so. The per-input claim rests on the cells where PAGE exceeds the oracle. Llama-3.1-8B NIAH-MK3, where the mean $D = 0.071$ sits just *above* the fixed τ so that a task-level policy keyed on that

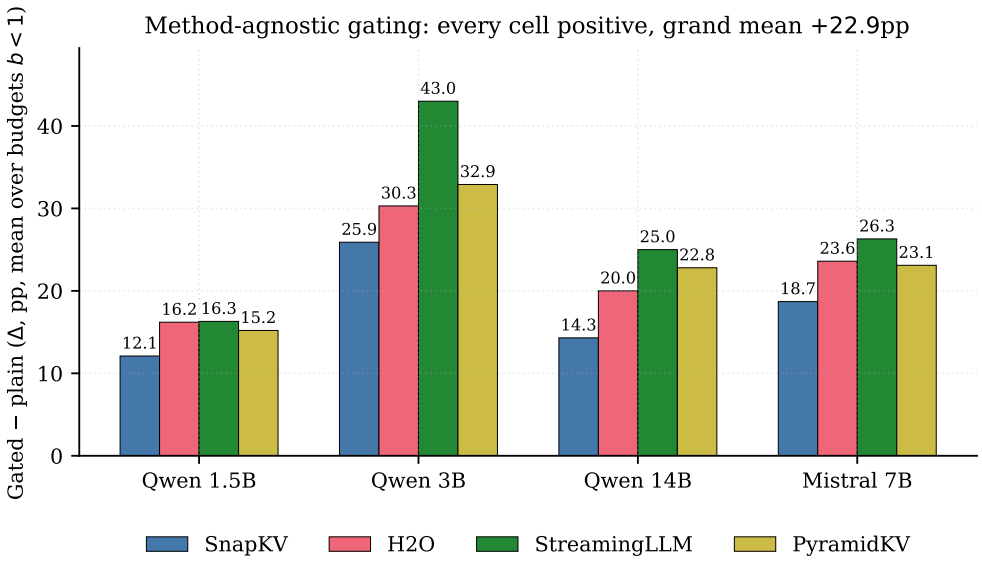


Figure 8: The headline 4×4 method-agnostic matrix. Bars are gated − plain accuracy averaged over budgets. The wrapper is positive in every one of 16 cells with a single $\tau = 0.07$ and no per-base-evictor tuning. Grand mean $+22.9$pp; excluding NIAH-MK3, $+4.5$pp (Table 11).

mean would open everywhere and gain nothing, does not support this argument, for two reasons. First, the oracle gap on that cell is *negative* under both thresholds we report. PAGE reaches $+0.089$ against the oracle's $+0.162$ at $\tau = 0.07$, and $+0.125$ against $+0.162$ under the z-scored threshold, so the cheating baseline wins there rather than loses. Second, the premise itself is not robust. The margin is 0.0012 in D, and whether the task mean falls above or below the z-scored threshold flips under three of four plausible definitions of the unlabeled pilot, which Section 3 does not pin down. Both quantities are recomputable from the released per-input log. The reconstruction reproduces the deployed gate exactly on all 1200 evicting rows.

A.4.4 Scorer-independence: full tables

Section 4.1 summarizes the scorer-independence result. The full tables follow.

The single-mask ManifoldKV variant used in the main text is not a fair test of a geometry scorer's full capability, so we also built a faithful ManifoldKV+Ada-KV implementation (Datta et al., 2026; Feng et al., 2025b): score $s_i = \|k_i - \mu\|_2$ against the per-(layer, kv-head) key centroid μ, with Ada-KV head-wise budget reallocation (safeguard floor `floor_alpha` $= 0.5$ plus a global top-k over the remaining per-head scores). Ragged per-head keep counts are realized by masking evicted keys to $-\infty$ per kv-head rather than physically pruning, which reproduces exactly the logits a ragged compressed cache would produce with no over-allocation, and the measured kept fraction matches the nominal budget to four decimals. We add KeyDiff (Park et al., 2025) (negative cosine to the same centroid) as the published attention-free scorer.

Before trusting a negative result from this scorer, we first confirm it is competent. Table 15(a) is the control that makes the negative result interpretable: on two-key NIAH the geometry scorer reproduces exactly the multi-key advantage its authors report, reaching 0.72 at $b = 0.5$ where SnapKV is at 0.16. A scorer that can do this is not failing MK3 for want of a better scoring rule, and indeed every scorer we test still collapses on MK3. Table 15(b) shows that at $b \leq 0.25$, SnapKV, KeyDiff, and ManifoldKV (post- and pre-RoPE, uniform and Ada-KV allocation) are all at or near 0.00 against a full-cache 0.66. The geometry scorer does not beat a properly per-head SnapKV at any budget here and is weaker at $b = 0.5$ (0.06-0.10 versus 0.32). The conclusion we draw is the one the gate is built on, stated with the scope Section 6 establishes. On this cell, on capacity-bound inputs, the recoverable information is not present in any 25% subset of the cache, so none of the scoring rules we test (attention-based, geometric, or attention-free) recovers it, and the only safe action is not to

Table 14: PAGE against a task-label oracle that closes on NIAH-MK3 and opens elsewhere. Δ is gated minus plain over budgets $b < 1.0$. A gap of exactly 0.000 means the gate reproduced the oracle's decisions on every input; a positive gap requires within-task per-input variation.

Base evictor	Model	PAGE Δ	oracle Δ	gap
SnapKV	Qwen2.5-1.5B	$+0.121$	$+0.121$	0.000
	Qwen2.5-3B	$+0.259$	$+0.183$	$+0.076$
	Qwen2.5-14B	$+0.143$	$+0.143$	0.000
	Mistral-7B	$+0.187$	$+0.195$	-0.008
H2O	Qwen2.5-1.5B	$+0.162$	$+0.162$	0.000
	Qwen2.5-3B	$+0.303$	$+0.230$	$+0.073$
	Qwen2.5-14B	$+0.200$	$+0.200$	0.000
	Mistral-7B	$+0.236$	$+0.243$	-0.007
StreamingLLM	Qwen2.5-1.5B	$+0.163$	$+0.163$	0.000
	Qwen2.5-3B	$+0.430$	$+0.232$	$\mathbf{+0.198}$
	Qwen2.5-14B	$+0.250$	$+0.250$	0.000
	Mistral-7B	$+0.263$	$+0.247$	$+0.015$
PyramidKV	Qwen2.5-1.5B	$+0.152$	$+0.152$	0.000
	Qwen2.5-3B	$+0.329$	$+0.221$	$+0.108$
	Qwen2.5-14B	$+0.228$	$+0.228$	0.000
	Mistral-7B	$+0.231$	$+0.240$	-0.009
grand mean		$\mathbf{+0.229}$	$\mathbf{+0.201}$	$+0.028$

evict. That floor is substrate-dependent rather than universal. The same geometry scorer reaches 0.990 at $b = 0.5$ on Llama-3.1-8B at 8K, and the budget below which every scorer fails ranges from $b \leq 0.125$ to $b \leq 0.25$ across the models we ran.

Table 15: Scorer-independence of the capacity limit (Qwen2.5-1.5B, RULER 4K, $N = 50$). (a) The geometry scorer reproduces the published two-key advantage, so it is competent. (b) On three-key NIAH-MK3 every scorer collapses below 25% budget. Accuracy at nominal budget b.

(a) two-key NIAH, the positive control

Scorer	full	$b = 0.5$	$b = 0.25$
SnapKV	0.76	0.16	0.04
ManifoldKV (post-RoPE)	0.76	**0.72**	**0.56**
ManifoldKV (pre-RoPE)	0.76	**0.76**	**0.52**

(b) three-key NIAH-MK3

Scorer (allocation)	full	$b = 0.5$	$b = 0.25$	$b = 0.125$	$b = 0.0625$
SnapKV (uniform per-head)	0.66	0.20	0.02	0.00	0.00
SnapKV (Ada-KV)	0.66	0.32	0.00	0.00	0.00
KeyDiff (uniform)	0.66	0.02	0.00	0.00	0.00
KeyDiff (Ada-KV)	0.66	0.02	0.00	0.00	0.00
ManifoldKV L2 post-RoPE (uniform)	0.66	0.02	0.00	0.00	0.00
ManifoldKV L2 post-RoPE (Ada-KV)	0.66	0.06	0.00	0.00	0.00
ManifoldKV L2 pre-RoPE (uniform)	0.66	0.02	0.00	0.00	0.00
ManifoldKV L2 pre-RoPE (Ada-KV)	0.66	0.10	0.00	0.00	0.00

A.4.5 Detect-and-retry as an alternative design

The gate is not the only way to handle capacity-bound inputs. A natural objection is that one could evict unconditionally and simply re-run the inputs that fail, avoiding the gate and its attention-exposing pass entirely. The comparison is analytic, so we state it rather than measure it. Writing $p_{\mathcal{C}}$ for the proportion of capacity-bound inputs and c_{pre}, c_{dec} for prefill and decode cost, PAGE pays c_{gate} on every input plus a memory cost of $(1 - p_{\mathrm{open}})$ full caches, where c_{gate} is 17ms on Qwen2.5-1.5B and 89ms on Mistral-7B at 4K (Table 28). Detect-and-retry instead pays nothing up

front but incurs a second full prefill and decode on every input it retries, so its expected overhead is $(p_{\mathcal{C}} \cdot \mathrm{TPR} + (1 - p_{\mathcal{C}}) \cdot \mathrm{FPR}) \cdot (c_{\mathrm{pre}} + c_{\mathrm{dec}})$, and every capacity-bound input its detector misses is returned wrong rather than merely late. At the rate $p_{\mathcal{C}} \approx 0.25$ measured on the realistic workload (Table 12) and with c_{gate} two orders of magnitude below a full prefill-plus-decode, retry breaks even only if its detector is both near-perfect in recall and has a false-positive rate below roughly 0.03, since each false positive costs a doubled request. The two designs also differ in kind rather than degree: the gate spends bounded memory to make a failure impossible, while retry spends unbounded tail latency to make it recoverable, and only the gate is available to a system that cannot re-run a request. We note this as a design point rather than a claim, since we did not build a decode-side distress detector, and a good one would be a contribution in its own right.

Table 16: Cross-architecture SnapKV-only runs on two Llama-architecture models. These use the four most aggressive budgets $\{0.0625, 0.125, 0.25, 0.5\}$ rather than the eight-budget grid of Table 2, so the Δ values are not directly comparable to the matrix rows (aggressive-budget-only averaging inflates Δ).

Model	plain SnapKV	gated SnapKV	Δ
Yi-1.5-9B	0.464	0.854	$+0.390$
Llama-3.1-8B	0.509	0.618	$+0.109$

Table 17: Best plain evictor vs. best gated evictor per model, over the four base evictors of Table 2 at the same budget grid.

Model	Best plain	Best gated	Δ
Qwen2.5-1.5B	0.438 (SnapKV)	0.570 (PyramidKV)	$+0.132$
Qwen2.5-3B	0.580 (SnapKV)	0.865 (PyramidKV)	$+0.285$
Qwen2.5-14B	0.709 (SnapKV)	0.868 (H2O)	$+0.159$
Mistral-7B	0.623 (SnapKV)	0.823 (PyramidKV)	$+0.201$

Static per-route policy A serving team with a known capacity-bound endpoint has a third option that costs nothing at all: route by endpoint and disable eviction on it, with no attention pass and no detector. Where that is available it is the right baseline, and by Table 14 a task-level policy already captures roughly 88% of what PAGE achieves on our matrix, so the honest comparison is narrow. PAGE earns its cost only where routing is unavailable or insufficient: mixed endpoints that serve both classes, traffic whose class is not known at admission, and the within-task variation of Section 4.1, where inputs of the same nominal task fall on opposite sides. We did not evaluate a routed deployment, and a reader whose traffic is cleanly separable by endpoint should treat static routing as the default and the gate as the fallback for what routing cannot reach.

A.4.6 Composition beyond attention-score evictors

The comparison above assumes the gate's signal generalizes beyond attention-score scoring, an assumption worth testing directly. To check composition beyond attention-score-based bases, we add a ManifoldKV-style scorer (Datta et al., 2026): per-(layer, head) Euclidean distance of each key to the head's key centroid, averaged into a single shared keep mask (a single-mask variant of their per-layer design, forced by our uniform-cache harness). On the Qwen2.5-1.5B 4K mixed suite ($N = 100$/task, four budgets), the same gate at the same $\tau = 0.07$ lifts it by $\Delta = +16.2$pp with no negative (task, budget) cell, and on NIAH-MK3 the gate recovers full-KV accuracy ($0.00 \to 0.65$) at every budget. Two observations follow. The gate signal is attention-based and scorer-independent, so composition extends outside the attention-score family, but our single-mask variant does *not* reproduce ManifoldKV's reported multi-key rescue (plain accuracy 0.01 at $b = 0.5$). Since their per-layer, per-head design may behave differently, we treat their published numbers as the authoritative claim and scope ours accordingly (Section 6).

A.4.7 CROSS-ARCHITECTURE, SCALE, AND BEST-PLAIN RESULTS

Table 16 extends the SnapKV gate to two Llama-architecture models, and Table 17 compares the best plain vs. best gated evictor per model.

On architecture, Yi-1.5-9B ($\Delta = +39.0$pp) and Llama-3.1-8B ($\Delta = +10.9$pp, Table 16) are Llama-architecture descendants distinct from Qwen and Mistral. Only the SnapKV column was run for compute, and on a smaller, more aggressive budget grid (see caption). The sign is positive on both, but on these two models part of the gain comes from the gate closing conservatively on tasks where plain SnapKV collapses, rather than from a correctly recovered partition. The per-task breakdown and the shared Llama-arch family caveat (niah_multivalue mis-classifies on both Yi and Llama) are in Section A.5.3 and Appendix A.3.3.

On scale, a SnapKV-only run on Qwen2.5-32B-Instruct (RULER 4K mixed suite, $N = 100$) confirms the predictor scales but also illustrates the saturation arm of the scaling formula. The drop ordering transfers cleanly (NIAH-MK3 smallest at $D = 0.025$, then FWE 0.052, VT 0.079, QA_1 0.103). At the same time the 32B model is near-saturated at 4K ($A_{\text{full}} = 1.00$ on VT and NIAH-MK3, $0.93 - 0.94$ on FWE/QA_1), so by Eq. (4) there is essentially no dilution headroom, and the gate's value is capacity-bound protection rather than recovery: plain SnapKV collapses at $b = 0.0625$ (NIAH-MK3 0.00, VT 0.00, FWE 0.09) while the gate holds the full cache on the low-drop tasks, giving mean $\Delta = +0.335$ at $\tau = 0.07$ (driven by NIAH-MK3 $+0.88$ and FWE $+0.44$) and $+0.237$ at the z-scored threshold with more compression (0.45 vs. 0.63 kept-KV). We report this as a scale point for the partition and predictor, not as additional dilution evidence.

The best-plain-versus-best-gated comparison still leaves open why the matrix's gains differ across base evictors rather than landing at one number, which the per-evictor behaviour explains. StreamingLLM keeps only 36 tokens regardless of budget, so on the dilution-prone fraction it is near zero plain and near zero gated, and gating contributes only via the gate-closed NIAH-MK3 inputs that go from 0 to full-KV accuracy. SnapKV is budget-scaled and degrades more gracefully, so it has less room for gating to recover. PyramidKV's depth-weighted scoring biases toward bottom layers but produces similar absolute gains to SnapKV. The pattern is that gating Δ tracks how much the base evictor collapses on the capacity-bound fraction. Two implementation caveats apply: our PyramidKV is a depth-weighted variant of the published per-layer budget schedule, and under two-pass prefill scoring (used on Qwen2.5-14B for memory) the H2O score only sees the last w queries, so the 14B H2O column degenerates toward SnapKV-style scoring. Its Δ there should be read as a near-duplicate of the SnapKV cell rather than an independent base evictor.

A.4.8 HEAD-TO-HEAD TABLES

Head-to-head with CapKV on LongBench CapKV (Yang et al., 2026a) is the closest published method that reports on LongBench. No public CapKV codebase exists yet, since their preprint is dated 2026-04-28, so we re-implement CapKV's per-(layer, kv-head) statistical-leverage score $s_i = w_i \cdot v_i^\top A^{-1} v_i$ with $A = I + \sum_i w_i v_i v_i^\top$ and $w_i = \exp(\langle k_i, \mu_q \rangle \beta / \sqrt{d})$ from their equations, then wrap it in our gate. Two LongBench subtasks, Qwen2.5-3B-Instruct, budget $b = 0.5$, CapKV temperature $\beta = 5$ (their default; we write β for their τ to avoid clashing with our gate threshold). Results in Table 18.

Table 18: Gated vs. plain CapKV (our re-implementation) on two LongBench subtasks, Qwen2.5-3B-Instruct, $b = 0.5$.

Subtask	N	plain CapKV	gated CapKV	Δ
LongBench/qasper	68	0.132	0.147	$+0.015$
LongBench/triviaqa	24	0.875	0.958	$+0.083$

Both columns are positive at matched budget. Conditional on the gate firing (drop ≥ 0.07, the same $\tau = 0.07$ we use everywhere), plain and gated agree by construction. On the gate-closed subset where the wrapper falls back to full KV, the difference is $+4.3$pp on qasper ($N_{\text{closed}} = 23$) and $+18.2$pp on triviaqa ($N_{\text{closed}} = 11$). The headline numbers are not a reproduction of CapKV's Table 1 (different model, different metric, 16K eager cap on prompt length) and we report them as a directional head-to-head: at these sample sizes ($N = 68, 24$) wrapping CapKV in the gate is weakly

positive overall and concentrates its gain on the gate-closed subset, consistent with the orthogonality claim of Section 4.1. We do not promote "never hurts" to a formal claim at these sample sizes, and larger LongBench runs are clean follow-up work.

Head-to-head with DBTrimKV on RULER Bui et al. (2026) (DBTrimKV) is the published trained-gate evictor the paper's mechanism section adopts. The authors ship pretrained gates on Hugging Face for several Qwen3 variants (no Qwen2.5 checkpoint exists, so the head-to-head is necessarily on a different model family). We run DBTrimKV-Qwen3-4B-Instruct-2507 on RULER 4K, $N = 30$ inputs per task on (NIAH-MK3, FWE, VT, QA_1), at three memory budgets $M \in \{128, 256, 512\}$. The wrapper uses the same fixed $\tau = 0.07$ calibrated on Qwen2.5-1.5B, with no per-model tuning. Results in Table 19.

Table 19: Gated vs. plain DBTrimKV (pretrained Qwen3-4B-Instruct-2507 gates) on RULER 4K at three memory budgets. The gate fires on only 1.7% of records, so the win is dominated by the full-KV fallback.

	N	plain DBTrimKV	gated DBTrimKV	Δ
all (3 budgets $\times$ 4 tasks)	360	0.617	0.850	$\mathbf{+0.233}$
per task: NIAH-MK3	90	0.467	1.000	$+0.533$
per task: FWE	90	0.222	0.600	$+0.378$
per task: VT	90	0.989	1.000	$+0.011$
per task: QA_1	90	0.789	0.800	$+0.011$
per budget: $M = 128$	120	0.475	0.850	$+0.375$
per budget: $M = 256$	120	0.617	0.850	$+0.233$
per budget: $M = 512$	120	0.758	0.850	$+0.092$

Threshold behaviour: the gate fires rarely on Qwen3 at $\tau = 0.07$ At our default threshold the gate is open on $6/360 = 1.7\%$ of records. The remaining $\sim 98\%$ trigger the full-KV fallback, which explains the large Δ on NIAH-MK3 ($+0.533$): the gate closing turns off DBTrimKV's compress step entirely. As a consequence, cache occupancy under our wrapper is $\sim 14.5\times$ plain DBTrimKV at the same nominal M. The $+0.233$ is therefore a *quality* win at matched budget, not a *memory-budget* win at matched cache size. The rare gate firing here is the same Qwen3 scale mismatch analyzed in Section 3: Qwen3-4B's drop distribution ($\mu = 0.034$) puts the fixed $\tau = 0.07$ at $z = +2.0$, so a per-model z-scored threshold ($\tau_{\text{Qwen3}} = 0.022$) is the right deployment default and, applied to the same suite, restores a genuinely compressing gate (43% kept-KV, $\Delta = +0.235$). The headline is that wrapping the strongest published trained-gate evictor with our partition-aware gate strictly improves matched-budget accuracy, with the gain concentrated on the capacity-bound task as the mechanism predicts.

A.4.9 Random-eviction control

We replace SnapKV scoring with uniform-random per-position scoring at every budget on a 4K, 4-key synthetic NIAH test-case ($N = 150$, Qwen2.5-1.5B). To be clear about the source, this 4-key test-case is our own construction and is *not* a RULER task: the NIAH-MKk names used elsewhere in the paper refer to RULER's `niah_multikey_`k (Hsieh et al., 2024), which tops out at three keys. The score-driven sweep retains accuracy at moderate budgets, while the uniform-random sweep collapses immediately (Table 20).

The per-input recovery rate ρ under random scoring is 0.027 (vs. 0.013 under SnapKV). The four "recovered" inputs under random are isolated successes among an otherwise destroyed distribution, and the control rules out a generic regularisation explanation: any-eviction-helps is falsified.

A.4.10 Per-layer eviction ablation

The default gating uses a single key-mask shared across layers. A per-layer variant derives each layer's keep mask from that layer's own last-window attention with the same total budget per layer (uniform cache shape). Table 21 reports ρ on the four dilution-prone RULER tasks where per-layer might plausibly help, on Qwen2.5-1.5B at 4K and 16K.

Table 20: SnapKV vs. uniform-random scoring at matched budget on the 4K 4-key synthetic NIAH test-case. Random eviction destroys accuracy long before the budget gets aggressive; the dilution effect requires informed scoring.

budget b	SnapKV acc	Random acc	Δ
1.000	0.827	0.827	$+0.000$
0.875	0.820	0.400	$+0.420$
0.750	0.820	0.173	$+0.647$
0.625	0.827	0.040	$+0.787$
0.500	0.833	0.033	$+0.800$
0.375	0.807	0.007	$+0.800$
0.250	0.760	0.000	$+0.760$
0.125	0.207	0.000	$+0.207$
0.0625	0.033	0.000	$+0.033$

Table 21: Per-layer vs. single-mask SnapKV on Qwen2.5-1.5B. Per-layer wins clearly on aggregation tasks at 4K (FWE $1.8\times$, niah_multivalue $2.1\times$); ties at 16K and on sequential-integration tasks.

Task	Context	Single-mask ρ	Per-layer ρ	Ratio
VT	4K	0.060	0.040	0.67
FWE	4K	0.060	**0.110**	1.83
QA_1	4K	0.080	0.070	0.88
niah_multivalue	4K	0.070	**0.150**	2.14
VT	16K	0.190	0.190	1.00
FWE	16K	0.040	0.050	1.25
QA_1	16K	0.140	0.140	1.00
niah_multivalue	16K	0.100	0.100	1.00

A.5 Deployment Details

This subsection collects the tables and secondary analyses behind the deployment constraints of Section 6.

A.5.1 Deferred experiment tables

The 8-cell scaling table Table 22 reports the gating Δ on the mixed suite across four models and two context lengths. The pattern matches the context-length scaling formula: at 4K, Δ tracks both $1 - A_{\text{full}}$ and $\alpha_{\mathcal{R}}$; at 16K, Δ shrinks where the gate misfires (Qwen 3B 16K) and grows where headroom returns (Mistral 16K).

Table 22: Cross-(model, context) gating Δ. Same $\tau = 0.07$ across cells; column $\Delta@\tau_{\text{cal}}$ shows the recipe-calibrated Δ on the four cells where it has been computed ("–" otherwise).

Model	Context	plain mean	gated@$\tau = 0.07$	Δ	$\Delta@\tau_{\text{cal}}$
Qwen2.5-1.5B	4K	0.438	0.559	$+0.121$	$+0.121$
Qwen2.5-3B	4K	0.580	0.839	$+0.259$	$+0.213$
Qwen2.5-14B	4K	0.709	0.853	$+0.143$	–
Mistral-7B	4K	0.623	0.810	$+0.187$	$+0.223$
Qwen2.5-1.5B	16K	0.372	0.404	$+0.032$	–
Qwen2.5-3B	16K	0.530	0.564	$+0.034$	$+0.054$
Qwen2.5-14B	16K	0.796	0.867	$+0.071$	–
Mistral-7B	16K	0.535	0.651	$+0.116$	–

Pre-registered 32K predictions Table 23 gives the full per-task predictions and measurements behind the out-of-range test of Section 4.1.

Table 23: Pre-registered 32K predictions vs. measurements (Qwen2.5-1.5B, $N = 100$ per task). Predicted intervals = 4K/16K ratio band $\times$ measured 32K headroom, frozen before the sweep.

Task	A_{full}(32K)	predicted ρ	measured ρ (Wilson 95%)	verdict
VT	0.81	$[0.048, 0.106]$	$0.14\ [0.085, 0.221]$	outside (above)
FWE	0.49	$[0.021, 0.263]$	$0.05\ [0.022, 0.112]$	inside
NIAH-MK3	0.14	≤ 0.02	$0.04\ [0.016, 0.098]$	outside (above; CI overlaps)

Table 24: Achieved compression of the gated deployment at $\tau = 0.07$ (SnapKV, 4K mixed suite). Kept-KV is the mean over inputs of the retained cache fraction; "grid mean" averages the eight nominal budgets of Table 2. Plain SnapKV keeps exactly the nominal fraction (0.0625 at $b = 0.0625$; grid mean 0.453). The gate-open fraction is a per-input property, independent of b, and sets the compression floor $1 - p_{\text{open}}$.

		gated kept-KV fraction			compression
Model	p_{open}	b=0.0625	b=0.25	grid mean	@ b=0.0625
Qwen2.5-1.5B	0.75	0.297	0.437	0.584	3.4×
Qwen2.5-3B	0.49	0.541	0.632	0.728	1.8×
Qwen2.5-14B	0.75	0.297	0.437	0.584	3.4×
Mistral-7B	0.72	0.327	0.462	0.602	3.1×

A.5.2 Achieved compression and batching decay

Achieved compression of the gated deployment Table 24 reports the per-model gate-open fraction and realized kept-KV behind the compression figures quoted in the abstract and Section 1. Realized compression at $b = 0.0625$ ranges from $1.8\times$ on Qwen2.5-3B, whose gate-open fraction is lowest at 0.49 and whose fallback therefore costs most, to $3.4\times$ on Qwen2.5-1.5B and Qwen2.5-14B, with a mean of $2.9\times$. The $3.4\times$ endpoint alone understates this spread. We report the range and the mean instead, since the spread is a property of the workload the gate sees rather than measurement noise.

Batching decay under static provisioning Table 25 evaluates the compression decay per model under static batch provisioning, where a batch is provisioned for its largest resident cache.

A.5.3 Per-task drop ordering

Table 26 reports the mean drop per task ($N = 100$ inputs per task on the Qwen and Mistral columns, $N = 50$ on Yi and Llama; Mistral 4K qa_1 has $N = 78$ and Mistral 16K niah_multivalue $N = 98$ due to run truncation). NIAH-MK3 is the smallest drop in every Qwen and Mistral cell, while on the two Llama-architecture models it ranks second-smallest, so the *ordering* of the partition is fully preserved across Qwen and Mistral and partially preserved on the Llama-architecture models, even when the absolute threshold needs slight tuning. This is what gives the calibration recipe in Eq. (5) its robustness.

Llama-class cross-architecture transfer is partial On Yi-1.5-9B and Llama-3.1-8B (both Llama-architecture descendants), niah_multivalue ranks smallest by D (Yi: $D = -0.013$; Llama: $D = +0.064$), and NIAH-MK3 ranks 2nd-smallest. On the five Qwen and Mistral cells, NIAH-MK3 is the smallest. We attribute the swap to a Llama-arch artifact: the multi-value retrieval task happens to have unusually concentrated early-layer attention on these models, which inverts the early-vs-late drop sign. The capacity-bound side of the partition (small D, gate closes, full-KV fallback) still operates on both models, but the fixed $\tau = 0.07$ does *not* transfer as a threshold. On Llama the mis-classification is confined to niah_multivalue, where the wrapper holds full KV instead of compressing, which forfeits the eviction speed-up but does not damage accuracy, because plain SnapKV is already robust there. The Llama-3.1-8B gated SnapKV Δ is $+10.9$pp on the matrix-comparable mixed suite, driven almost entirely by NIAH-MK3 ($+43.5$pp) where plain SnapKV collapses to 0% and gated holds $\sim 44\%$ (still well below the full-KV 1.00). The other tasks have

Table 25: Compression under static batch provisioning at $\tau = 0.07$ (SnapKV, $b = 0.0625$). p_{open} and k_{open} are from Table 24. Compression decays toward $1\times$ as batch size grows because a single gate-closed sequence forces the whole batch to the full cache.

Model	B=1	B=4	B=8	B=16	B=32
Qwen2.5-1.5B	3.37×	1.42×	1.10×	1.01×	1.00×
Qwen2.5-3B	1.85×	1.06×	1.00×	1.00×	1.00×
Qwen2.5-14B	3.37×	1.42×	1.10×	1.01×	1.00×
Mistral-7B	3.06×	1.33×	1.07×	1.00×	1.00×

Table 26: Per-task mean head-agreement drop D across three architecture families (Qwen, Mistral, and Llama-architecture: Yi-1.5 and Llama-3.1). This table uses the strict-thirds bins, early $= [0, \lfloor L/3 \rfloor)$ and late $= [\lfloor 2L/3 \rfloor, L)$; the deployed gate uses equal-size bins and differs by at most 0.004 where $3 \nmid L$ (Appendix G). Bold marks the smallest drop per column. NIAH-MK3 is the smallest drop on Qwen and Mistral cells; on Yi-1.5 and Llama-3.1 it ranks 2nd-smallest behind niah_multivalue (Llama-arch family artifact; see prose). Yi-1.5-9B serves as a parallel Llama-class architecture with a distinct training corpus (different tokenizer, deeper layer geometry).

Task	Qwen 1.5B 4K	Qwen 3B 4K	Qwen 3B 16K	Mistral 4K	Mistral 16K	Yi-1.5-9B 4K	Llama-3.1-8B 4K
QA_1	0.218	0.105	0.099	0.096	0.132	0.080	0.097
QA_2	0.191	0.097	0.092	0.087	–	0.079	0.101
VT	0.152	0.076	0.069	0.095	0.096	0.065	0.083
niah_multivalue	0.107	0.045	0.070	0.079	0.075	−0.013	**0.064**
FWE	0.098	0.038	0.046	0.079	0.091	0.043	0.114
NIAH-MK3	**0.040**	−0.004	−0.021	**0.060**	**0.053**	0.019	0.073

$\Delta \approx 0$ because Llama's plain SnapKV is robust on this budget range (≥ 0.86 at $b = 0.0625$ on QA_1). The NIAH-MK3 gate-open fraction on Llama is 0.56 despite mean $D = 0.071$ sitting above τ: per-input D dispersion closes the gate on the other $\sim 44\%$ of NIAH-MK3 inputs, which is what rescues the cell. This cell does *not* support a per-input claim over a task-level policy, since the task-label oracle outperforms PAGE here under both thresholds (Section 4.1). The dispersion explains the cell's Δ, not an advantage over knowing the label. On Yi the mis-classification is broader: at $\tau = 0.07$ the gate also stays closed on FWE and VT (gate-open fractions 0.00 and 0.10), so most of Yi's $+39.0$pp comes from the full-KV fallback firing on tasks where plain SnapKV collapses. That outcome is accuracy-preserving but it is *not* evidence that the raw threshold transfers to Yi, since the drop-vs-ρ prediction is inverted there. The z-scored variant of Section 3 resolves this: with a per-model standardization from an unlabeled pilot, Yi compresses for real (kept-KV $0.22 - 0.32$) while still gaining $\Delta = +0.145$, and Llama improves to $\Delta = +0.174$. We state the fixed-τ failure plainly in Section 4.1.

A.5.4 Scaling formula verification

Table 27 reports the headroom $1 - A_{\text{full}}$, the observed ρ, and the ratio. A through-origin fit on the dilution-prone rows gives a slope of 0.34. The Qwen2.5-3B 4K FWE row (ratio 0.04) is excluded from that fit: at $N = 100$, $\rho = 0.01$ is a single recovery, Wilson $[0.002, 0.054]$, and is better read as counting noise than as part of the trend. On NIAH-MK3 the ratio is ≤ 0.02 regardless of headroom, so the partition is visible in the ratio alone.

A.5.5 Cost of the gate and end-to-end efficiency

Table 28 reports the end-to-end efficiency of gated eviction.

Deployment cost at scale We measured the two-pass gate on Qwen2.5-32B (64 layers, 40 heads): the gate costs 350ms at 4K (145ms re-forward +204ms head-agreement drop, $\approx 18\%$ of the 1.9s prefill), and the 16K two-pass run *OOMs* on a contended 80GB card (weights + KV + the $w \times T$ attention transient). The drop cost is quadratic in head count: a head-count sweep gives an exponent of $2.0 - 2.4$ (per-model floors 2.1ms on Qwen-1.5B, 27ms Mistral-7B, 87ms Qwen-32B, extrapolating to ≈ 291ms on a 70B-class model), confirming the $O(LH^2k)$ complexity. The $w \times T$ attention transient is $LHwT \cdot 2$ bytes (5GiB at 32K on Qwen-32B, $\approx$ 10GiB on 70B-class). A

Table 27: Verification of the scaling formula. The ratio $\rho/(1 - A_{\text{full}})$ is approximately constant on dilution-prone tasks across (model, context) cells and near zero on NIAH-MK3 regardless of headroom.

Sweep	$1 - A_{\text{full}}$	ρ	ratio
Qwen2.5-1.5B 4K VT	0.18	0.06	0.33
Qwen2.5-1.5B 16K VT	0.75	0.19	0.25
Qwen2.5-3B 4K VT	0.00	0.00	–
Qwen2.5-3B 16K VT	0.09	0.05	0.55
Qwen2.5-3B 4K FWE	0.24	0.01	0.04
Qwen2.5-3B 16K FWE	0.62	0.32	0.52
Qwen2.5-1.5B 4K NIAH-MK3	0.35	0.01	0.03
Qwen2.5-1.5B 16K NIAH-MK3	0.74	0.00	0.00

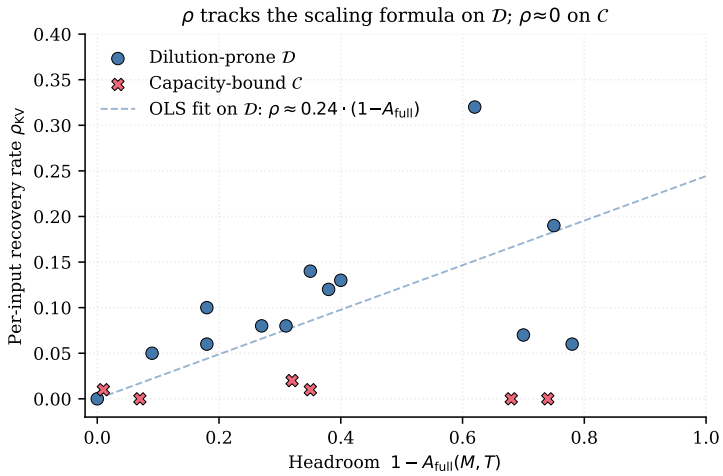


Figure 9: Per-input recovery rate ρ vs. accuracy headroom $1 - A_{\text{full}}$ across the $(model, context, task)$ cells of Table 27. Dilution-prone tasks (filled circles) lie on a band with a through-origin fitted slope of 0.34. NIAH-MK3 (open triangles) lies on the $\rho = 0$ axis regardless of headroom. The two populations are visibly separated.

70B model at 32K is therefore not viable in a paged-FlashAttention stack: the two-pass re-forward needs contiguous (non-paged) KV and a non-fused eager kernel, and materializes a per-sequence transient that paged attention exists to avoid. PAGE is best read as an offline/prefill-time gate, not a fused-serving drop-in.

A.5.6 Additional limitations

Two further limitations concern which baselines appear where. DBTrimKV's pretrained gates exist only for Qwen3 variants, so Section 4.1 runs on Qwen3-4B-Instruct-2507 rather than the Qwen2.5-keyed matrix, which is why DBTrimKV stays a standalone head-to-head rather than a matrix row. The matrix itself covers only training-free policies for the same reason: trained or learned evictors (DBTrimKV, CapKV, IndexMem, DynamicKV) are cited but not included in it.

The last limitation concerns how confidently NIAH-MK3 can stand in for the capacity-bound class generally. It is the only task where empirical capacity-boundedness (accuracy collapse under eviction) and the gate's capacity-bound classification coincide. On LongBench passage_count at Qwen2.5-14B the gate closes on all inputs (mean $D = -0.013$) yet eviction is empirically *safe* (accuracy flat under $16\times$ eviction, $\rho = 0$): a benign *false positive* (over-conservative, costing compression not accuracy), the dual of the MK2 false negative (Appendix A.2.5). The head-agreement drop is therefore a one-sided predictor, reliable at firing on dilution-prone inputs, not a perfect clas-

Table 28: End-to-end efficiency of gated eviction (RULER 4K, batch 1, bf16, greedy, 128 decode steps, medians; A100-80GB). "Gated expected" is the gate-open-weighted mix $0.75 \cdot (b{=}0.125) + 0.25 \cdot \text{full}$ at the measured mixed-suite gate-open fraction ($\tau{=}0.07$); "gated measured" actually gates each input ($15/20$ open, matching the $N{=}400$ calibration fraction of 0.75). On the 1.5B model (2 KV heads), 4K decode is weight-bound and eviction buys *memory* ($8\times$ plain, $2.9\times$ gated expected), not speed; on Mistral-7B (8 KV heads, 513 MiB KV) eviction already yields a $1.29\times$ decode speedup at 4K. Measured KV matches the analytic $L \cdot H_{\text{kv}} \cdot d \cdot T \cdot 4\,\text{B}$ formula to 0.1 MiB, validating the 16K/32K scaling rows.

Model	Config	Decode tok/s	KV cache (MiB)	Peak decode (MiB)
Qwen2.5-1.5B ($N{=}20$)	Full KV	37.9	106.1	3090
	SnapKV $b{=}0.125$	37.8	13.2 ($8.0\times$)	2976
	SnapKV $b{=}0.0625$	37.5	6.6 ($16.1\times$)	2968
	Gated (measured, $\tau{=}0.07$)	37.7	13.2 / 110.5 (open/closed)	2976
	Gated (expected mix)	37.9	36.5 ($2.9\times$)	3004
Mistral-7B-v0.3 ($N{=}12$)	Full KV	18.9	512.8	14429
	SnapKV $b{=}0.125$	24.5 ($1.29\times$)	64.1 ($8.0\times$)	13929

Prefill-time overhead (Qwen2.5-1.5B, 4K, median ms)		Analytic KV at $b{=}0.125$ gated (0.75 open)	
Prefill (eager)	344	Qwen2.5-1.5B, 16K / 32K	448 → 154 / 896 → 308 MiB
SnapKV scoring	1.2	Mistral-7B, 16K / 32K	2.0 → 0.69 / 4.0 → 1.38 GiB
Head-agreement gate	17.2	(gate on Mistral-7B: 88.5 ms, $O(LH^2k)$ head-pair growth)	
Eviction + cache rebuild	3.2		

Table 29: Fixed-budget harm rate $\Pr_x[A(b,x) < A_{\text{plain}}(1,x)]$, gated / plain at matched nominal budget, pooled over the seven SnapKV cells (Qwen2.5-1.5B/3B/14B and Mistral-7B at 4K; Qwen2.5-1.5B/3B and Mistral-7B at 16K), $N = 500$ per (task, budget), $\tau = 0.07$. Unlike $r(x)$ above this is not empty: the gate can and does harm inputs. On the capacity-bound task it cuts harm by $29\times$ at $16\times$ compression; on QA_1 it is exactly inactive (identical rates, the gate opens on every input); on VT at $16\times$ both arms are harmed almost equally, so the gate does not help there.

Task	$b = 0.0625$	$b = 0.125$	$b = 0.25$	$b = 0.5$
NIAH-MK3	**0.026** / 0.750	**0.026** / 0.744	**0.024** / 0.724	**0.018** / 0.546
FWE	**0.236** / 0.414	**0.162** / 0.298	**0.118** / 0.224	**0.064** / 0.116
QA_1	0.110 / 0.110	0.064 / 0.064	0.030 / 0.030	0.010 / 0.010
VT	0.684 / 0.694	0.108 / 0.110	0.034 / 0.036	0.026 / 0.026

sifier. Multilingual and code tasks remain out of scope. We expect the mechanism to transfer, since it is architecture-level, but we have not measured it.

A.6 Statistical Validation

A.6.1 Per-input recovery distributions

The mean ρ hides the shape of the per-input recovery, so we also report its distribution. For each input x we form the recovery indicator $r(x) = \text{sign}(\max_b A_{\text{gated}}(b,x) - A_{\text{plain}}(1,x)) \in \{-1, 0, +1\}$, where the maximum is over the budget sweep and $A_{\text{plain}}(1,x)$ is the full-cache plain accuracy. Figure 10 pools all (M,T) cells per task, and Figure 11 reports the fraction of strictly recovered inputs ($r(x) > 0$) within each (M,T) cell.

What $r(x)$ can and cannot show The budget sweep includes $b = 1.0$, and at $b = 1.0$ the gated arm keeps the entire cache and reproduces the plain full-cache outcome. We audit this on all 16 matrix cells plus the four 16K cells: at $b = 1.0$, $n_{\text{kept}} = T$ and $A_{\text{gated}} = A_{\text{plain}}$ on every one of the 5,120 rows. Hence $\max_b A_{\text{gated}}(b,x) \geq A_{\text{gated}}(1,x) = A_{\text{plain}}(1,x)$ and $r(x) \geq 0$ is an *identity*, not a measurement. The empty $r = -1$ bin is therefore true by construction and carries no information about whether gating ever harms an input. $r(x)$ should instead be read only as an *oracle-budget recovery* measure, that is, how often some budget in the sweep beats the full cache.

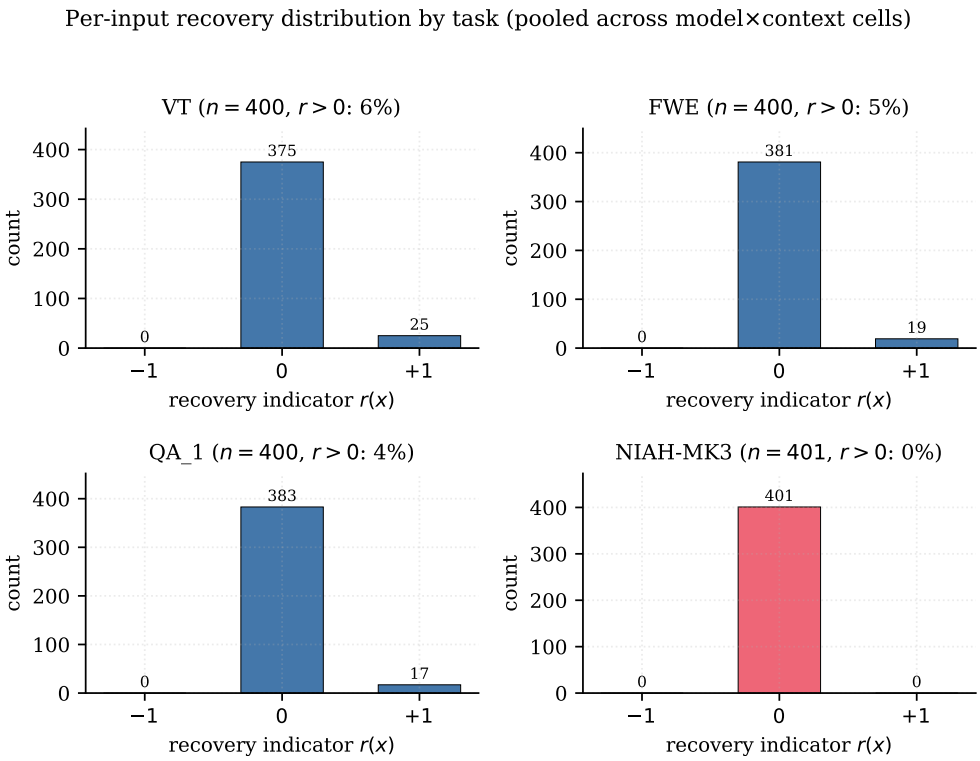


Figure 10: Per-input recovery indicator $r(x)$ pooled across model and context cells, faceted by RULER task. The three dilution-prone tasks show a non-trivial right tail at $r{=}{+}1$; NIAH-MK3 is concentrated entirely at $r{=}0$, consistent with the capacity-bound regime. The $r{=}{-}1$ bin is empty by construction, not by measurement, because the budget sweep includes $b = 1.0$; see the identity argument above and Table 29 for the harm rate this figure cannot report.

The question it cannot answer, whether a *deployed* gate at a *fixed* budget ever loses an answer the full cache would have got, is answered by the harm rate below.

Fixed-budget harm rate Holding b fixed makes "worse than the full cache" a real case. Table 29 reports $\Pr_x[A(b,x) < A_{\text{plain}}(1,x)]$ for the gated and plain arms at matched b, pooled over the seven SnapKV (model, context) cells ($N = 500$ per task and budget).

The 16K Qwen2.5-1.5B cell is `gated_16k_qwen15b_sdpa.jsonl`, the SDPA-prefill run reported throughout the paper. An earlier aborted eager run left a nine-row NIAH-MK3-only stub under a similar name; pooling that stub instead drops 50 inputs per task from FWE, QA_1 and VT and shifts the column above to $0.029/0.796$, $0.291/0.444$, $0.118/0.118$ and $0.751/0.756$. We state this because the two files are easy to confuse and the released analysis script pins the correct one.

The histograms show that ρ is not driven by a few outliers: across the three dilution-prone tasks we recover a few percent to over ten percent of inputs in many cells. NIAH-MK3 sits at $r{=}0$ on essentially every input, confirming that the capacity-bound regime is a genuine null rather than a noisy average. This input-level view matches the partition diagnosed by the scaling formula and rules out the case that gating helps only a small number of inputs by chance. It does not license any claim that gating is harmless: Table 29 shows that at a fixed budget the gate harms 2.6% of NIAH-MK3 inputs and 23.6% of FWE inputs at $16\times$ compression.

A.6.2 Confidence intervals for the headline numbers

Tables 30 and 31 report 95% intervals for every headline delta, computed as described in the uncertainty paragraph of Section 4. Full per-cell derivations ship with the released logs.

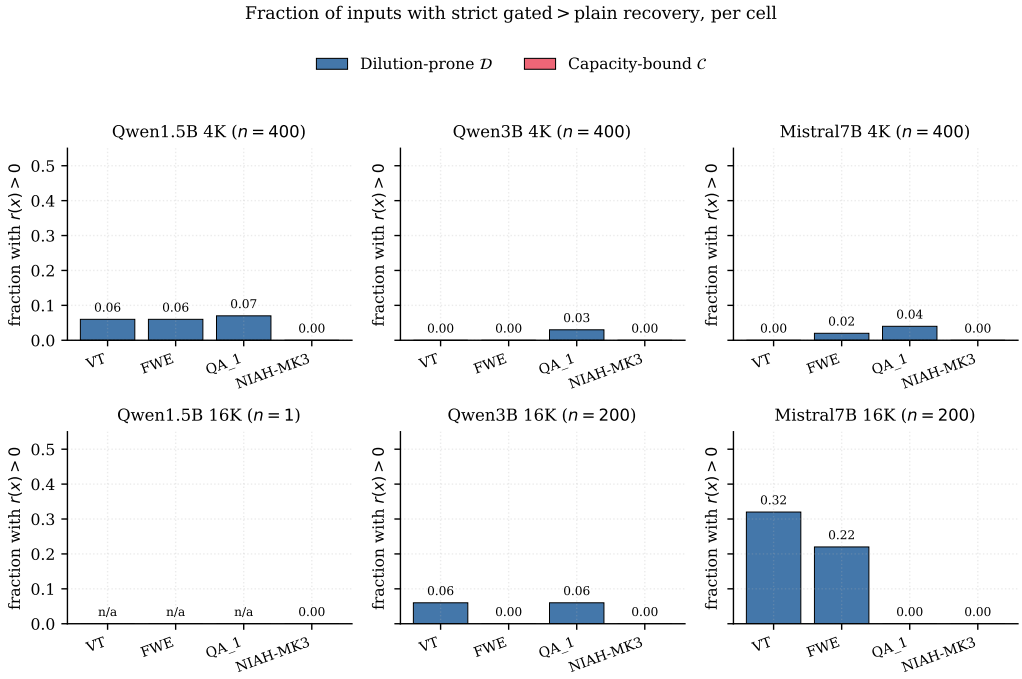


Figure 11: Fraction of inputs with strict recovery $r(x) > 0$ per task, faceted by $(\text{model}, \text{context})$ cell. Dilution-prone bars are non-trivial in every cell with data, while NIAH-MK3 stays at or near zero, visualising the partition input by input.

A.6.3 Seed replicates

The intervals above are cluster bootstraps over a fixed set of inputs, so they quantify sampling within one draw and say nothing about the draw itself. We therefore re-ran five SnapKV cells under three draws of the RULER split, varying only the shuffle seed (Table 32). Decoding is greedy, so there is no decoding stochasticity to average over.

Seed SD ranges from 0.004 to 0.011 against a grand-mean Δ of $+0.229$, so draw-to-draw spread runs roughly an order of magnitude below the effect it measures. The published SnapKV entries of Figure 8 fall inside the replicate range for Qwen2.5-1.5B, Qwen2.5-3B and Mistral-7B. The Qwen2.5-14B entry, $+0.143$, sits 0.001 above the largest replicate, which we attribute to N: that cell is $N = 200$ in Table 2 and $N = 400$ here. Restricted to NIAH-MK3, where nearly all of Δ lives, the spread widens to $0.009 - 0.045$ on a much larger effect. Qwen2.5-1.5B is the loosest cell at 0.045 on a mean near $+0.45$, still an order of magnitude below its own effect.

All fifteen runs used one host, and the replicates of a cell are never split across machines. That matters because the reproduction gate (Appendix A.3.6) finds a small but real host effect on mean D, which would otherwise appear in this table as seed variance.

B Theory: Mechanism, Partition, and Scaling Derivations

Section 5 states the mechanism, the partition it implies, and the scaling formula in summary form. This section provides the full derivations, followed by the sufficient-condition bridge connecting the mechanism to the observable predictor. At the end of each part, an explicit scope statement defines what is and is not established, so the claims can be read exactly. Notation follows Bui et al. (2026) $\alpha_{\mathcal{R}}$ the pre-eviction signal mass, γ the noise-retention ratio, $\rho_{\mathcal{R}}, \rho_{\mathcal{N}}$ the retained-mass fractions, δ_t the dilution, D the head-agreement drop, σ_z the pre-softmax logit-noise scale, m the logit margin, $\kappa := m/\sigma_z$, and H, L, T the head, layer, and length counts. Throughout, the T key positions split as $T = R + N$ into a relevant set $\mathcal{R}$ ($|\mathcal{R}| = R$) and a noise set $\mathcal{N}$ ($|\mathcal{N}| = N$), with $R \ll N$.

Table 30: 95% confidence intervals for the 16 gating deltas of Table 2. Cluster bootstrap over inputs: we resample the (task, id) pairs with replacement (10,000 replicates); each input contributes its per-budget paired gated−plain differences, preserving the pairing and the within-input correlation across budgets; intervals are $2.5 - 97.5$ percentiles. All gated outcomes are the exact $\tau = 0.07$ post-hoc reconstruction of Appendix G. $N = 400$ inputs per cell (4 tasks $\times$ 100), $N = 200$ for Qwen2.5-14B. *Every one of the* 16 *intervals excludes zero*; the smallest lower bound is $+0.093$ (Qwen2.5-1.5B, SnapKV). These are 16 simultaneous comparisons, so we also adjust for multiplicity. Under Benjamini-Hochberg at $q = 0.05$, with per-cell p-values from the same cluster bootstrap, *all* 16 *cells survive*, the least significant at $p \leq 5 \times 10^{-4}$ against a threshold of 0.050. The adjustment does not change any conclusion drawn from this table, which is why we report it.

Model	Base evictor	Δ	95% CI	N
Qwen2.5-1.5B	SnapKV	$+0.121$	$[+0.093, +0.151]$	400
Qwen2.5-1.5B	H2O	$+0.162$	$[+0.126, +0.199]$	400
Qwen2.5-1.5B	StreamingLLM	$+0.163$	$[+0.128, +0.200]$	400
Qwen2.5-1.5B	PyramidKV	$+0.152$	$[+0.118, +0.188]$	400
Qwen2.5-3B	SnapKV	$+0.259$	$[+0.225, +0.295]$	400
Qwen2.5-3B	H2O	$+0.303$	$[+0.261, +0.347]$	400
Qwen2.5-3B	StreamingLLM	$+0.430$	$[+0.383, +0.477]$	400
Qwen2.5-3B	PyramidKV	$+0.329$	$[+0.286, +0.372]$	400
Qwen2.5-14B	SnapKV	$+0.143$	$[+0.108, +0.181]$	200
Qwen2.5-14B	H2O	$+0.200$	$[+0.152, +0.250]$	200
Qwen2.5-14B	StreamingLLM	$+0.250$	$[+0.190, +0.310]$	200
Qwen2.5-14B	PyramidKV	$+0.228$	$[+0.175, +0.285]$	200
Mistral-7B	SnapKV	$+0.187$	$[+0.154, +0.220]$	400
Mistral-7B	H2O	$+0.236$	$[+0.196, +0.278]$	400
Mistral-7B	StreamingLLM	$+0.263$	$[+0.220, +0.305]$	400
Mistral-7B	PyramidKV	$+0.231$	$[+0.191, +0.272]$	400

Table 31: 95% CIs for the remaining headline deltas. Cross-architecture rows (Table 16) use the four-budget grid and the same cluster bootstrap; 16K rows are the second half of Table 22; the DBTrimKV row is Table 19 ($N = 120$ inputs, budgets $\{128, 256, 512\}$). The Mistral NIAH-MK3 showcase entries are Wilson score intervals on $N = 100$. The only headline delta whose interval includes zero is Qwen2.5-3B at 16K, the gate-misfire cell that Section 4.1 already reports as a shrunken effect.

Quantity	Point	95% CI	N
Yi-1.5-9B Δ (SnapKV, 4-budget)	$+0.390$	$[+0.338, +0.444]$	200
Llama-3.1-8B Δ (SnapKV, 4-budget)	$+0.109$	$[+0.069, +0.154]$	200
Qwen2.5-1.5B 16K Δ	$+0.032$	$[+0.013, +0.053]$	200
Qwen2.5-3B 16K Δ	$+0.034$	$[-0.002, +0.071]$	200
Qwen2.5-14B 16K Δ	$+0.071$	$[+0.034, +0.107]$	120
Mistral-7B 16K Δ	$+0.116$	$[+0.079, +0.156]$	200
DBTrimKV wrapper Δ (RULER 4K)	$+0.233$	$[+0.172, +0.297]$	120
Mistral 4K NIAH-MK3, gated acc. @ b=0.0625	0.89	$[0.814, 0.937]$	100
Mistral 4K NIAH-MK3, plain acc. @ b=0.0625	0.00	$[0.000, 0.037]$	100

Two distinct noise scales appear and should not be conflated: σ_z is the pre-softmax logit-noise scale (Appendix D), and σ_v is the per-distractor value-vector variance under isotropy (Appendix C).

Mechanism and partition We adopt as stated the dilution mechanism from Bui et al. (2026), which suffices to motivate our gating predictor. Fix a frozen decoder language model M. A single attention head reads from positions $t \in \{1, \ldots, T\}$, each carrying a value v_t and a pre-eviction probability α_t with $\sum_t \alpha_t = 1$. Partition positions into a relevant set $\mathcal{R} = \{t : r_t = 1\}$ ($|\mathcal{R}| = R$)

Table 32: Seed replicates of the SnapKV cells. Three draws of the RULER 4K split, identical in every other respect: greedy decoding, $\tau = 0.07$, Δ averaged over budgets $b < 1.0$, $N = 400$ (4 tasks $\times$ 100). The lower block restricts to NIAH-MK3.

Model	seed 1	seed 2	seed 3	mean	SD	range
All four tasks						
Qwen2.5-1.5B	+0.113	+0.099	+0.122	+0.111	0.011	0.022
Qwen2.5-3B	+0.255	+0.267	+0.275	+0.266	0.010	0.019
Qwen2.5-14B	+0.133	+0.140	+0.142	+0.138	0.005	0.010
Mistral-7B	+0.187	+0.194	+0.183	+0.188	0.006	0.011
Llama-3.1-8B	+0.078	+0.071	+0.078	+0.076	0.004	0.007
NIAH-MK3 only						
Qwen2.5-1.5B	+0.454	+0.398	+0.486	+0.446	0.045	0.089
Qwen2.5-3B	+0.738	+0.747	+0.729	+0.738	0.009	0.019
Qwen2.5-14B	+0.530	+0.559	+0.568	+0.552	0.020	0.037
Mistral-7B	+0.709	+0.710	+0.669	+0.696	0.023	0.041
Llama-3.1-8B	+0.310	+0.282	+0.311	+0.301	0.016	0.029

and a noise set $\mathcal{N}$ ($|\mathcal{N}| = N$), with $R \ll N$. Let $U_t \subseteq \mathcal{R}$ denote the *useful* set at step t (the positions whose retention preserves the next-token margin), and define dilution as the attention mass *outside* U_t:

$$\delta_t := 1 - \sum_{i \in U_t} \alpha_{t,i}. \tag{6}$$

Their Proposition 3.1 establishes that near-tie distractors (α_t comparable across $\mathcal{N}$) force δ_t to be large, and their Corollary 3.2 shows that preferential retention with $\rho_U \geq \rho_D$ (where ρ_U, ρ_D are the retained mass on useful and distractor positions) reduces δ_t.

The partition follows from this mechanism. Define $\alpha_{\mathcal{R}} := \sum_{t \in \mathcal{R}} \alpha_t$ as the pre-eviction signal mass. Two regimes emerge:

- *Dilution-prone* ($\mathcal{D}$): $\alpha_{\mathcal{R}} \ll 1$. Most attention is on noise; eviction of noise positions monotonically improves the head SNR.
- *Capacity-bound* ($\mathcal{C}$): $\alpha_{\mathcal{R}} \to 1$. Attention is concentrated on the few relevant positions. Imperfect scoring can only *lose* signal; eviction strictly degrades the head.

The partition we observe on RULER is the empirical version of $\alpha_{\mathcal{R}}$ onto the task structure: NIAH-MK3 forces a single concentrated signal ($\alpha_{\mathcal{R}} \to 1$), while VT, FWE, QA, and niah_multivalue all spread signal across many positions ($\alpha_{\mathcal{R}} \ll 1$).

Context-length scaling Fix the task and its relevant-set size R. Let $T = R+N$. Assume distractor attention is approximately uniform on $\mathcal{N}$: $\alpha_t \approx (1 - \alpha_{\mathcal{R}})/N$ for $t \in \mathcal{N}$. Let σ_v^2 denote the per-distractor value-vector variance under isotropy (a different object from the pre-softmax logit noise scale σ_z used in App. D). Then the head SNR before eviction is

$$\mathrm{SNR} = \frac{\alpha_{\mathcal{R}}^2}{\sigma_v^2 \sum_{t \in \mathcal{N}} \alpha_t^2} \approx \frac{\alpha_{\mathcal{R}}^2 \cdot N}{\sigma_v^2 (1 - \alpha_{\mathcal{R}})^2}.$$

If $\alpha_{\mathcal{R}}$ decreases with T (more distractors compete for softmax mass), the pre-eviction SNR decreases and the noise level grows.

Let the kept set be $\mathcal{K}$ and define $\rho_{\mathcal{R}} := \sum_{t \in \mathcal{K} \cap \mathcal{R}} \alpha_t / \alpha_{\mathcal{R}}$ (the retained fraction of relevant mass) and analogously $\rho_{\mathcal{N}}$. Under noise-retention ratio $\gamma \in (0, 1)$ (i.e. $|\mathcal{K} \cap \mathcal{N}| = \gamma N$), the post-eviction SNR ratio $\mathrm{SNR}_{\mathrm{post}}/\mathrm{SNR}_{\mathrm{pre}}$ is a monotone-increasing function of the *gain argument* (Eq. (10)).

$$g := \rho_{\mathcal{R}}^2 / \gamma \tag{7}$$

The squaring survives the renormalisation step, and consequently, g is not a function of $\rho_{\mathcal{R}}/\gamma$. The accuracy headroom $1 - A_{\mathrm{full}}(M, T)$ scales with T on dilution-prone tasks.

The single-budget improvement event factorises into a headroom indicator $A := \mathbf{1}[\hat{A}(\mathrm{SNR}_{\max}) < 1]$ and a recovery indicator $B := \mathbf{1}[\exists b : \mathrm{SNR}_{\mathrm{post}} > \mathrm{SNR}_{\mathrm{pre}}]$. The recovery probability is

$p_{\text{recov}}(g) := \Pr_x[B]$, with $p_{\text{recov}}(g) = 0$ for $g \leq 1$ and non-decreasing above. Since the operational $b^*(x)$ is the argmax over the budget grid, and a calibrated (nested top-k) scorer ensures recovery budgets are not near-disjoint, the union-bound factor reduces to unity. The per-input expected accuracy improvement at a single budget b thus obeys the exact scaling relation:

$$\rho_{\text{KV}} = (1-A_{\text{full}})\, p_{\text{recov}} + \text{Cov}_x(A, B), \qquad (1-A_{\text{full}})\, p_{\text{recov}} \;\leq\; \rho_{\text{KV}} \;\leq\; \min\{1-A_{\text{full}},\, p_{\text{recov}}\}. \tag{8}$$

The lower endpoint uses $\text{Cov}_x(A, B) \geq 0$, which holds because both indicators are monotone in the single driver $\alpha_{\mathcal{R}}$ (Chebyshev's association inequality (Harris, 1960)); the upper endpoint is Fréchet–Hoeffding. Eq. (4) (main text) predicts three observed effects in one formula: context-length amplification on dilution-prone tasks (Qwen2.5-1.5B VT goes $\rho = 0.06 \to 0.19$ as $T = 4\text{K} \to 16\text{K}$, in lockstep with $1 - A_{\text{full}} = 0.18 \to 0.75$); larger-model saturation at short context (Qwen2.5-3B 4K VT has $1 - A_{\text{full}} \approx 0.00$, so $\rho \approx 0$); and re-entry of larger models into the dilution regime at long context (Qwen2.5-3B 16K FWE has $1 - A_{\text{full}} = 0.62$ and $\rho = 0.32$). Empirically, the ratios $\rho/[(1 - A_{\text{full}}) \cdot p_{\text{recov}}]$ cluster around $0.25 - 0.55$ on dilution-prone tasks.

Scope. This derivation establishes the identity and two-sided bound with constant 1 *inside* the deterministic-SNR surrogate. It does not establish the surrogate itself (accuracy as a fixed function of head SNR is assumed); the covariance is pinned only to a sign, not a value. Where headroom is driven by $\alpha_{\mathcal{R}}$-independent factors (the QA parametric-knowledge cells) both bounds pinch $\rho \to 0$ despite headroom, matching the observed QA exceptions as a sign prediction, not a proof about those tasks.

Connecting the predictor to the mechanism: a sufficient condition for non-trivial dilution The mechanism uses the unobservable $\alpha_{\mathcal{R}}$. Our gating predictor uses the head-agreement drop D (Eq. (3)), which is observable. We sketch a proof connecting the two in Appendix D, and leave the formal version with explicit, tight constants to future work. Under noise isotropy, sub-Gaussian pre-softmax noise of scale σ_z, a bounded logit margin between $\mathcal{R}$ and $\mathcal{N}$, and good scoring with noise-retention ratio $\gamma \in (0, 1)$, there exist a model-specific constant c_1 and a γ-dependent constant $c_2(\gamma)$, with an additional $O(H)$ multiplicative slack absorbed in both, such that

$$D \;\geq\; c_1 \cdot \log(1/\gamma) \;\Longrightarrow\; \mathbb{E}[\delta_t] \;\geq\; c_2(\gamma) \cdot (1 - A_{\text{full}}), \tag{9}$$

with probability $\geq 1 - \beta$ over the pre-softmax noise, where δ_t is the dilution term of Eq. (6). The converse (capacity-bound $\Rightarrow$ small D) holds under the same surrogate given a κ-separation assumption $m \geq 2\sigma_z\sqrt{\log(2H^2LT^2/\beta)}$: then, by the top-$k$ concentration result stated and proved as Lemma 1 in Section D, late-layer heads share a common top-k set, $a_{\text{late}} \to 1$, and $\Pr[D \geq \tau] \leq \beta = 2H^2LT^2\, e^{-\kappa^2/4}$, so observing $D \geq \tau$ certifies the input is not capacity-bound except on a noise tail.

Measured margin (why the bridge does not certify the gate). The converse is only informative when $\beta < 1$, which needs $\kappa \geq 10.11$ at $(H, L, T) = (12, 28, \approx 4\text{K})$. We measured κ directly on Qwen2.5-1.5B RULER 4K NIAH-MK3: median $\kappa = 3.33$, about $3\times$ too small. The gap is robust: sweeping the top-k split over $k \in \{8, 16, 32, 64\}$ gives $\kappa = 4.01, 3.68, 3.33, 2.99$ (monotone, none rescues the bound); a second architecture gives 3.71 (Mistral-7B); and, decisively, a dilution-prone task gives $\kappa = 3.40$ (VT), close to the capacity-bound 3.33 and so the margin itself is not what distinguishes the classes (the gate separates them through D, not κ). A factor of three in κ is a factor $> 10^9$ in β. The margin condition is a union bound over $T^2 \cdot HL$ failure events and is correspondingly loose, so its failure leaves the top-k concentration *uncertified rather than false*: the bound is therefore *uninformative at the paper's own parameters*, and the class separation the gate relies on is empirical (Table 26), not certified by the surrogate.

Sign vs. ordinal. As stated the converse predicts $D \leq 0$ on capacity-bound inputs, which Table 26 contradicts in five of seven cells (measured D is mildly positive, $+0.02$ to $+0.07$). Relaxing the common-set idealisation to an $O(1)$ needle neighbourhood weakens the conclusion to $\Pr[D \leq o(1)] \geq 1 - \beta - o(1)$, which the data satisfy but which no longer distinguishes the classes by sign. We therefore rely on the ordering based separation only.

Scope. Establishes the sufficient direction and, under the named κ-separation and common-set assumptions, the converse and its $e^{-\kappa^2/4}$ rate; also (App. H content) an $O(H^2\varepsilon)$ approximate identification of U_t with $\mathcal{I}^{(\ell^*)}$ under ε-close heads plus a consensus–usefulness assumption, which degrades

exactly in the single-decisive-head (retrieval-head, Wu et al., 2025) regime where the paper already declines to apply the bridge. Does not establish that real attention satisfies the margin assumption (measured false, $\kappa \approx 3$ vs. ≥ 10.1 needed), that capacity-bound heads literally share one common set, that the $O(H)$ pairwise-to-H-wise slack or the heuristic $\log(1/\gamma)$ factor are tight, or a standalone necessity proof. The dilution mechanism motivates the bridge, not the reverse. The bridge is the sufficient-condition span from the mechanism to the observable D, and it does not reach that observable at the measured margin. Neither the bridge nor its converse carries the empirical claims, which rest on the measured partition (Table 26) and the direct gating results. We retain this analysis as motivation for why an early-minus-late agreement difference is a reasonable signal, while noting that the signal's validity is established empirically, not by this bridge.

C Proof sketch of the scaling proposition

We expand the sketch from Section 5.

Setup Let $T = R + N$ with R relevant and N noise positions. Assume noise isotropy: $\alpha_t \approx (1-\alpha_{\mathcal{R}})/N$ for $t \in \mathcal{N}$, with $\alpha_{\mathcal{R}} := \sum_{t\in\mathcal{R}} \alpha_t$, and let σ_v^2 be the per-distractor value-vector variance under isotropy (distinct from the pre-softmax logit noise scale σ_z of App. D). The pre-eviction head SNR is

$$\mathrm{SNR}_{\mathrm{pre}} = \frac{\alpha_{\mathcal{R}}^2}{\sigma_v^2 \sum_{t\in\mathcal{N}} \alpha_t^2} \approx \frac{\alpha_{\mathcal{R}}^2 N}{\sigma_v^2 (1-\alpha_{\mathcal{R}})^2}.$$

Eviction with noise-retention ratio $\gamma \in (0,1)$ keeps a fraction $\rho_{\mathcal{R}}$ of the relevant mass and a fraction $\rho_{\mathcal{N}}$ of the noise mass. After softmax renormalisation over the kept set, the post-eviction relevant mass is $\widetilde{\alpha}_{\mathcal{R}} = \rho_{\mathcal{R}}\alpha_{\mathcal{R}}/Z$ with $Z = \rho_{\mathcal{R}}\alpha_{\mathcal{R}} + \rho_{\mathcal{N}}(1-\alpha_{\mathcal{R}})$, and the retained noise positions number γN with per-position mass $\rho_{\mathcal{N}}(1-\alpha_{\mathcal{R}})/(Z\gamma N)$. Substituting,

$$\mathrm{SNR}_{\mathrm{post}} = \frac{\widetilde{\alpha}_{\mathcal{R}}^2}{\sigma_v^2 \sum_{t\in\mathcal{K}\cap\mathcal{N}} \widetilde{\alpha}_t^2} \approx \frac{\rho_{\mathcal{R}}^2 \alpha_{\mathcal{R}}^2 \, \gamma N}{\sigma_v^2 \, \rho_{\mathcal{N}}^2 (1-\alpha_{\mathcal{R}})^2} = \frac{\rho_{\mathcal{R}}^2}{\rho_{\mathcal{N}}^2 \gamma^{-1}} \cdot \mathrm{SNR}_{\mathrm{pre}}. \tag{10}$$

Under good scoring ($\rho_{\mathcal{R}} \geq \rho_{\mathcal{N}}$ with $\rho_{\mathcal{N}} \approx \gamma$ on a calibrated scorer), the gain factor reduces to $\rho_{\mathcal{R}}^2/\gamma$, so the post-eviction SNR strictly exceeds the pre-eviction SNR. The renormalisation Z is the step that turns the raw fractions $(\rho_{\mathcal{R}}, \rho_{\mathcal{N}})$ into a usable SNR ratio. Without it the argument would collapse to $\rho_{\mathcal{R}}^2/\gamma$ absent the squaring.

Factorisation We work inside the surrogate, where the true accuracy indicator $\mathbf{1}[A(b,x) > A(b_{\max},x)]$ is replaced by the surrogate indicator $\mathbf{1}[\hat{A}(\mathrm{SNR}(b,x)) > \hat{A}(\mathrm{SNR}(b_{\max},x))]$ with $\hat{A} : [0,\infty) \to [0,1]$ strictly increasing and saturating at 1. This replacement is the surrogate's content: it asserts that accuracy is a deterministic function of head SNR. Within the surrogate the improvement event factorises as

$$\mathbf{1}[\hat{A}(\mathrm{SNR}_{\mathrm{post}}) > \hat{A}(\mathrm{SNR}(b_{\max},x))] = \mathbf{1}[\hat{A}(\mathrm{SNR}(b_{\max},x)) < 1] \cdot \mathbf{1}[\hat{A}(\mathrm{SNR}_{\mathrm{post}}) > \hat{A}(\mathrm{SNR}_{\mathrm{pre}})].$$

Both monotonicity and saturation are needed: if $\hat{A}$ were not strictly increasing the second indicator could fire without an SNR gain, and if $\hat{A}$ did not saturate at 1 the case $\hat{A}(\mathrm{SNR}(b_{\max},x)) = 1$ would not zero out the left-hand side. The first indicator is the headroom in the surrogate (a population-level proxy for the true full-cache failure probability $1 - A_{\mathrm{full}}$). The second depends only on $\rho_{\mathcal{R}}^2/\gamma$ via Eq. (10) once $\rho_{\mathcal{N}} \approx \gamma$ is folded in. The transfer from surrogate event to operational accuracy is mediated by the surrogate-fits-data assumption: we treat $\hat{A}$ as a good fit to the expected accuracy conditional on SNR, and deviations are absorbed into the empirical constant, fitted at 0.34. Marginalising over the prompt distribution and treating the two indicators as independent gives

$$\rho_{\mathrm{KV}}(\mathsf{T}, M, T) \approx \big(1 - A_{\mathrm{full}}(M,T)\big) \cdot p_{\mathrm{recov}}(\rho_{\mathcal{R}}^2/\gamma),$$

which is Eq. (4). The boundary conditions on p_{recov} follow from the surrogate, and both are statements about the gain argument $g = \rho_{\mathcal{R}}^2/\gamma$ of Eq. (7) rather than about $\rho_{\mathcal{R}}$ and γ separately. At $\rho_{\mathcal{R}} = 0$ no relevant mass survives, so $\mathrm{SNR}_{\mathrm{post}} = 0$, recovery is impossible and $g = 0$. At $\gamma = 1$ (no eviction) $\mathrm{SNR}_{\mathrm{post}} = \mathrm{SNR}_{\mathrm{pre}}$ identically, the second indicator never fires, and $g = \rho_{\mathcal{R}}^2 \leq 1$. More generally the second indicator requires $\mathrm{SNR}_{\mathrm{post}} > \mathrm{SNR}_{\mathrm{pre}}$, which by Eq. (10) is exactly $g > 1$. Hence $p_{\mathrm{recov}}(g) = 0$ on the whole dead zone $g \leq 1$, and is non-decreasing above it. Stating the two endpoints separately as $p_{\mathrm{recov}}(0) = p_{\mathrm{recov}}(1) = 0$ with monotonicity on $[0,1]$ would force $p_{\mathrm{recov}} \equiv 0$ there, which is the dead zone rather than a prediction of no recovery anywhere.

From surrogate to operational ρ_{KV} The factorisation argues about the improvement event at a single budget b, whereas the operational ρ_{KV} (Eq. (1)) is the per-input strict-Pareto frequency over the best budget $b^*(x)$. The two are related by a union bound over the discrete budget grid $\mathcal{B} = \{b_1, \ldots, b_K\}$:

$$\Pr_x\big[A(b^*(x), x) > A(b_{\max}, x)\big] = \Pr_x\big[\exists\, b \in \mathcal{B} : A(b, x) > A(b_{\max}, x)\big] \\ \leq \sum_{b \in \mathcal{B}} \Pr_x\big[A(b, x) > A(b_{\max}, x)\big].$$

Under monotone recovery (p_{recov} non-decreasing in $\rho_{\mathcal{R}}^2/\gamma$, with $\rho_{\mathcal{R}}$ determined by the scorer's top-k at each b), the right-hand side is bounded by $K \cdot \Pr_x[A(\widetilde{b}, x) > A(b_{\max}, x)]$ for the budget $\widetilde{b}$ that maximises the per-budget improvement probability. Under the independence approximation of the previous paragraph, this maximiser term equals Eq. (4), so

$$\rho_{\mathrm{KV}}(\mathsf{T}, M, T) \;\leq\; K \cdot \big(1 - A_{\mathrm{full}}(M, T)\big) \cdot p_{\mathrm{recov}}\big(\rho_{\mathcal{R}}^2/\gamma\big).$$

This is a one-sided upper bound at the cost of a multiplicative grid factor K. **The factor is an artifact of the union step and is not needed.** Appendix F shows that on any budget grid whose improvement events are nested, which is what a calibrated scorer produces, the correct leading constant is exactly 1, and characterises the nested-versus-disjoint boundary at which a real factor K could appear. We therefore state the bound with constant 1 and keep the K form only to show where the loose version comes from.

Independence bound via covariance Let A and B denote the headroom and recovery indicators viewed as Bernoulli random variables under the prompt-distribution measure. The exact decomposition is $\Pr[A \cdot B] = \Pr[A]\ \Pr[B] + \mathrm{Cov}(A, B)$, so the independence step incurs only the covariance residual. Both indicators are monotonically decreasing in $\alpha_{\mathcal{R}}$: larger $\alpha_{\mathcal{R}}$ raises $\mathrm{SNR}_{\mathrm{pre}}$, pushing A_{full} toward 1 (less headroom, so A trends to 0) and simultaneously saturating $\mathrm{SNR}_{\mathrm{post}}$ earlier under (A1) noise-isotropy (less recovery margin, so B trends to 0). Coupling A and B through the single scalar $\alpha_{\mathcal{R}}$ makes them comonotone in this driver, so Chebyshev's association (sum) inequality (Harris, 1960) gives $\mathrm{Cov}(A, B) \geq 0$, i.e. $\Pr[A \cdot B] \geq \Pr[A]\ \Pr[B]$. The lattice-theoretic FKG inequality is not needed here and does not apply directly: both indicators are monotone functions of one scalar driver, which is the elementary Chebyshev/Harris setting. Lemma 3 in App. F proves the same step, including the case where A carries an auxiliary driver independent of $\alpha_{\mathcal{R}}$. The Cauchy–Schwarz / Bernoulli-variance bound $|\mathrm{Cov}(A, B)| \leq \sqrt{\mathrm{Var}(A)\,\mathrm{Var}(B)} \leq 1/4$ then yields an absolute, distribution-free residual. Combining with the union-bound paragraph above,

$$\rho_{\mathrm{KV}}(\mathsf{T}, M, T) \;\leq\; K \cdot \big(1 - A_{\mathrm{full}}(M, T)\big) \cdot p_{\mathrm{recov}}\big(\rho_{\mathcal{R}}^2/\gamma\big) \;+\; K/4.$$

The $K/4$ term is rigorous but uninformative at our $K \leq 7$, since a bound of 1.75 on a probability says nothing. Both K factors disappear under the sharpened treatment of Appendix F, which replaces this paragraph's conclusion with the exact identity $\rho_{\mathrm{KV}} = (1 - A_{\mathrm{full}})\, p_{\mathrm{recov}} + \mathrm{Cov}(A, B)$ and a covariance term that is signed rather than merely bounded. The derivation above is included because it isolates the step that introduces the slack, though every empirical claim in the paper rests on the sharpened version.

D PROOF SKETCH: SUFFICIENT CONDITION FOR THE PARTITION

We sketch a proof of the bridge inequality Eq. (9): under the simplified attention model below, a large head-agreement drop D implies a non-trivial lower bound on Bui's dilution δ_t (Bui et al., 2026). The argument is structured as five numbered steps followed by a converse heuristic, an empirical-validation paragraph, and an explicit list of what is *not* proved. We give closed forms for the constants c_1 and $c_2(\gamma)$ rather than leave them implicit. The constants carry an $O(H)$ multiplicative slack from the pairwise-to-H-wise Jaccard step in Step 3, and tightening the constants is left to future work.

Per-head logit model With $\mathcal{R}, \mathcal{N}$ the relevant and noise sets as above, for layer $\ell \in \{1, \ldots, L\}$ and head $h \in \{1, \ldots, H\}$ the pre-softmax logits are $z_{h,i}^{(\ell)} = (q_h^{(\ell)})^\top k_i \eta_{h,i}^{(\ell)}$, where $\eta_{h,i}^{(\ell)}$ is zero-mean σ_z-sub-Gaussian noise independent across (ℓ, h, i). The subscript z distinguishes the pre-softmax

logit noise scale used here from the value-vector variance σ_v used in Appendix C. We write $\alpha_{h,i}^{(\ell)}$ for the post-softmax attention. Bui's dilution $\delta_t = 1 - \sum_{i \in U_t} \alpha_{t,i}$ is taken at the decoding query at step t. We drop the t subscript and write $\delta := \delta_t$.

Step 1 (the surrogate model) We use the multi-head lift of the single-head surrogate from Appendix C:

- **(A1)** *Signal coherence.* For each layer ℓ and head h, there is a head-private relevant subset $\mathcal{R}_h^{(\ell)} \subseteq \mathcal{R}$ such that for $i \in \mathcal{R}_h^{(\ell)}$ the deterministic part of the logit satisfies $(q_h^{(\ell)})^\top k_i \geq m_+$, and for $i \in \mathcal{N}$, $(q_h^{(\ell)})^\top k_i \leq m_-$, with $m_+ - m_- \geq m > 0$. This m is the *logit margin*; it is bounded above by $\log T$ in the well-trained regime.
- **(A2)** *Sub-Gaussian noise.* $\eta_{h,i}^{(\ell)}$ is zero-mean and σ_z^2-sub-Gaussian, independent across (ℓ, h, i). This is the standard noise model used in attention-as-feature-aggregation analyses.
- **(A3)** *Capacity-bound vs dilution-prone regimes.* In the capacity-bound regime $\mathcal{C}$, all heads at all layers share a common relevant set: $\mathcal{R}_h^{(\ell)} = \mathcal{R}$ for every (ℓ, h). In the dilution-prone regime $\mathcal{D}$, late-layer head subsets diverge: there exists a layer index $\ell^* > L/2$ such that for any head pair (h, h') at layer ℓ^*, $|\mathcal{R}_h^{(\ell^*)} \cap \mathcal{R}_{h'}^{(\ell^*)}| / |\mathcal{R}_h^{(\ell^*)} \cup \mathcal{R}_{h'}^{(\ell^*)}| \leq J_{\text{late}}$, while early-layer heads share their subsets, $|\mathcal{R}_h^{(1)} \cap \mathcal{R}_{h'}^{(1)}| / |\mathcal{R}_h^{(1)} \cup \mathcal{R}_{h'}^{(1)}| \geq J_{\text{early}}$, with $J_{\text{early}} > J_{\text{late}}$. Both J_{early} and J_{late} are model-specific population constants of the surrogate, not functions of L. This layer structure is architecture-dependent, not a universal claim: on Qwen2.5-1.5B the late-plateau ordering holds (capacity-bound NIAH-MK3 sits at 0.283, above dilution-prone VT at 0.227), whereas on Llama-3.1-8B the per-layer profile is non-monotone and the separation between regimes collapses (Appendix A.3.3, Figure 5). The results downstream of (A3) are stated within this surrogate and are not claimed to hold where the measured profile departs from it.
- **(A4)** *Identification of U_t.* We identify Bui's useful set U_t (a decoding-step quantity defined at the residual-stream level) with the head-common signal set $\mathcal{I}^{(\ell^*)} := \bigcap_h \mathcal{R}_h^{(\ell^*)}$ at the divergence layer ℓ^* (a single-layer attention quantity). The identification is a modelling choice (positions every head agrees are signal). It collapses time-step and layer-index axes, and within the surrogate the two are assumed to coincide rather than shown to. We flag it in the "What is not proved" list below.

The empirical content of (A3) is the layer-dependence we measure: early layers respond to local context (similar $\mathcal{R}_h^{(\ell)}$ across heads, regardless of task), late layers respond to task structure (head divergence on dilution-prone tasks, head convergence on capacity-bound tasks). See §3 for the corresponding operational definitions.

Step 2 (head-agreement drop as Jaccard divergence) The head-agreement drop D in Eq. (3) is the early-late difference of mean pairwise top-k Jaccard agreement, with k chosen to be commensurate with R (in the experiments $k = 32 \sim |\mathcal{R}|$). In the high-SNR limit of the surrogate (Lemma 1 below), each head's top-k set concentrates on its own private $\mathcal{R}_h^{(\ell)}$ with high probability. Hence at layer ℓ,

$$a_\ell \;\approx\; \frac{1}{\binom{H}{2}} \sum_{h<h'} \frac{|\mathcal{R}_h^{(\ell)} \cap \mathcal{R}_{h'}^{(\ell)}|}{|\mathcal{R}_h^{(\ell)} \cup \mathcal{R}_{h'}^{(\ell)}|}. \tag{11}$$

Binning into early third and late third and subtracting,

$$D \;=\; a_{\text{early}} - a_{\text{late}} \;\approx\; \overline{J}_{\text{early}} - \overline{J}_{\text{late}}, \tag{12}$$

the difference of mean pairwise Jaccard between layers in the early and late bins. D is therefore a direct measurement of the head-set divergence in (A3).

Lemma 1 (Top-k concentration on the private set)**.** *Under (A1), (A2) with $m \geq \sigma_z \sqrt{4 \log(2H^2LT^2/\beta)}$, the top-$k$ set of head (ℓ, h) equals $\mathcal{R}_h^{(\ell)}$ jointly across all HL heads with probability $\geq 1 - \beta$, when $k = |\mathcal{R}_h^{(\ell)}|$.*

Proof. A standard sub-Gaussian maximal inequality: for any pair (i, j) with $i \in \mathcal{R}_h^{(\ell)}$ and $j \in \mathcal{N}$, $\Pr[z_{h,i}^{(\ell)} < z_{h,j}^{(\ell)}] \le \exp(-m^2/(4\sigma_z^2))$. There are $R \cdot N \le T^2/4$ such relevant-vs-noise pairs at each layer-head. Substituting the threshold $m \ge \sigma_z\sqrt{4\log(2H^2LT^2/\beta)}$ gives a per-pair tail of $\beta/(2H^2LT^2)$. Union-bounding over up to T^2 pairs per layer-head and over all HL layer-heads gives a total tail of $HL \cdot T^2 \cdot \beta/(2H^2LT^2) = \beta/(2H) \le \beta$. □

The threshold $m \ge \sigma_z\sqrt{4\log(2H^2LT^2/\beta)}$ is the *margin condition*. If it holds, every step below holds with probability $\ge 1 - \beta$ over the noise.

Step 3 (head-set divergence forces union expansion) For each layer ℓ define the union of head-private sets $\mathcal{U}^{(\ell)} := \bigcup_{h=1}^{H} \mathcal{R}_h^{(\ell)}$ and the intersection $\mathcal{I}^{(\ell)} := \bigcap_{h=1}^{H} \mathcal{R}_h^{(\ell)}$. The mean pairwise Jaccard $\overline{J}_\ell$ at layer ℓ is a *pairwise* object, whereas $|\mathcal{U}^{(\ell)}|$ and $|\mathcal{I}^{(\ell)}|$ are H-wise objects, so the ratio $|\mathcal{U}|/|\mathcal{I}|$ is not equal to $1 - \overline{J}$ in general. Inclusion-exclusion only gives $|\mathcal{U}|/|\mathcal{I}| \le 1 + (H-1)(1-\overline{J})$ in the best case, and the union-vs-intersection gap can scale with H in the worst case. We therefore avoid an H-wise identity and use the weaker pairwise inequality. Under (A3), $\overline{J}_{\ell^*} \le J_{\text{late}}$, so there exists a head pair (h, h') at layer ℓ^* with $|\mathcal{R}_h^{(\ell^*)} \cap \mathcal{R}_{h'}^{(\ell^*)}| \le J_{\text{late}} \cdot |\mathcal{R}_h^{(\ell^*)} \cup \mathcal{R}_{h'}^{(\ell^*)}|$. Picking the head $h^\dagger$ with the largest private residual set $\mathcal{R}_{h^\dagger}^{(\ell^*)} \setminus \mathcal{I}^{(\ell^*)}$,

$$|\mathcal{U}^{(\ell^*)} \setminus \mathcal{I}^{(\ell^*)}| \;\ge\; |\mathcal{R}_{h^\dagger}^{(\ell^*)} \setminus \mathcal{I}^{(\ell^*)}| \;\ge\; (1 - J_{\text{late}}) \cdot \max_h |\mathcal{R}_h^{(\ell^*)}|. \tag{13}$$

This is strictly weaker than treating $(|\mathcal{U}| - |\mathcal{I}|)/|\mathcal{U}|$ as if it were equal to $1 - \overline{J}$, but it is correct, and the slack is absorbed in c_1 below.

Under (A4), Bui's useful set is $U_t = \mathcal{I}^{(\ell^*)}$, so positions in $\mathcal{U}^{(\ell^*)} \setminus \mathcal{I}^{(\ell^*)}$ are attended by some head but not by all, and from the perspective of the next layer's query they contribute attention mass outside U_t.

Step 4 (bridge to Bui's δ_t) Let $\bar{\alpha}$ denote the average per-head, per-position attention on a position in $\mathcal{R}_h^{(\ell^*)}$ at layer ℓ^*. Under (A1) with margin m, the softmax concentrates and $\bar{\alpha} \ge (1 - (N/R)e^{-m})/R$ on private positions, tight when $e^m \gg N/R$ (an earlier draft omitted the N/R factor, which is not negligible in the paper's own $R \ll N$ regime). Treating per-head value contributions as additive after the output projection (a modelling choice we flag below), the attention mass *outside* $\mathcal{I}^{(\ell^*)}$ contributed by the maximal-residual head is at least

$$\sum_{i \in \mathcal{R}_{h^\dagger}^{(\ell^*)} \setminus \mathcal{I}^{(\ell^*)}} \alpha_{h^\dagger, i}^{(\ell^*)} \;\ge\; \bar{\alpha} \cdot (1 - J_{\text{late}}) \cdot \max_h |\mathcal{R}_h^{(\ell^*)}|, \tag{14}$$

using Eq. (13). Under eviction with noise-retention ratio γ and a top-k scorer that does not a priori distinguish signal-side residuals from noise positions (an additional modelling assumption beyond (A1)-(A4) which we make explicit here), positions in $\mathcal{U}^{(\ell^*)} \setminus \mathcal{I}^{(\ell^*)}$ inherit the same effective per-position retention rate γ as the noise set. Those retained for only a subset of heads appear as distractors to the remaining heads after the cache is merged. Conditional on this retention,

$$\mathbb{E}[\delta_t] \;\ge\; \bar{\alpha} \cdot (1 - J_{\text{late}}) \cdot \max_h |\mathcal{R}_h^{(\ell^*)}| \cdot (1 - \gamma), \tag{15}$$

because the $(1 - \gamma)$ fraction of un-retained positions in the residual private sets no longer balance the heads that did keep them, leaving a residual distractor mass proportional to $(1 - \gamma)$. This is a new bound derived in our surrogate. It is structurally analogous to Bui's Corollary 3.2 (which relates pre- and post-eviction dilution via $\delta' = (\rho_D/\rho_U)\delta/[(1-\delta) + (\rho_D/\rho_U)\delta]$), but the input here is the head-disagreement-induced distractor set rather than a pre-existing dilution.

Step 5 (closing the bridge: solve for D) Eq. (15) uses $(1 - \overline{J}_{\text{late}})$, which we relate to the observable D via Eq. (12): $D = \overline{J}_{\text{early}} - \overline{J}_{\text{late}}$. Since $\overline{J}_{\text{early}} \le 1$ trivially (Jaccard is bounded by one), $D \le 1 - \overline{J}_{\text{late}}$, so substituting gives

$$\mathbb{E}[\delta_t] \;\ge\; \bar{\alpha} \cdot D \cdot \max_h |\mathcal{R}_h^{(\ell^*)}| \cdot (1 - \gamma). \tag{16}$$

The substitution $1 - \overline{J}_{\text{late}} \geq D$ holds always; we do not require the stronger $\overline{J}_{\text{early}} \approx 1$ assumption. The lesser the value of $\overline{J}_{\text{early}}$, the larger the gap between $1 - \overline{J}_{\text{late}}$ and D, and the looser this bound is relative to its tightest form.

The $\log(1/\gamma)$ factor on the right-hand side of Eq. (9) is a heuristic amplification term, not a rigorous consequence of the steps above. The intended argument is that after eviction the cache retains γN noise positions, and the maximum of γN sub-Gaussian logits concentrates near $\sigma_z\sqrt{2\log \gamma N}$. The multiplicative penalty on the head signal-to-noise budget should scale with the post-eviction noise level relative to the pre-eviction floor. A careful Taylor expansion of $\sqrt{2\log N} - \sqrt{2\log \gamma N}$ gives a difference of order $\log(1/\gamma)/\sqrt{2\log N}$, not $\log(1/\gamma)$ itself, so the log factor therefore captures only the worst-case noise-tail shift at small N and not the asymptotic regime. We retain the $\log(1/\gamma)$ factor in Eq. (9) as a convenient packaging of the eviction amplification, but we flag this ($\log(1/\gamma)$ heuristic) in the "What is not proved" gap list below (item 3). Readers willing to drop the factor entirely can read Eq. (9) as $D \geq c_1$ with $c_1 = 1/J_{\text{early}}$. This is the conservative no-amplification case, and it is the form on which we rely as the deployed threshold $\tau = 0.07$ (Section 3) is calibrated empirically rather than derived from c_1.

Two heuristic observations close the bridge to $(1 - A_{\text{full}})$. Both are leading-order estimates rather than rigorous bounds. Observation (i) enters the "What is not proved" gap list as the leading-order substitution of item 3, and observation (ii) rests on the accuracy surrogate flagged in item 2. (i) Under (A1) noise-isotropy and the high-SNR regime, the average per-position mass on a relevant position satisfies $\bar{\alpha} \approx \alpha_{\mathcal{R}}/R$. The set containment $\mathcal{R}_h^{(\ell^*)} \subseteq \mathcal{R}$ then gives the rigorous one-sided relation $\bar{\alpha} \cdot \max_h |\mathcal{R}_h^{(\ell^*)}| \leq \alpha_{\mathcal{R}}$. Replacing this one-sided relation by the equality $\bar{\alpha} \cdot \max_h |\mathcal{R}_h^{(\ell^*)}| \approx \alpha_{\mathcal{R}}$ requires $\max_h |\mathcal{R}_h^{(\ell^*)}| \approx R$, that is, that a single head's private set already covers most of $\mathcal{R}$. This is an order-of-magnitude assumption strictly stronger than union coverage $\bigcup_h \mathcal{R}_h^{(\ell^*)} \approx \mathcal{R}$, and it appears as a leading-order substitution in the "What is not proved" gap list below (item 3). Only the one-sided relation is used in the conservative case. (ii) Combining the headroom factorisation of Appendix C (Eq. (4)) with the strictly increasing accuracy surrogate gives $(1 - A_{\text{full}}) \leq C \cdot \alpha_{\mathcal{R}}$ on dilution-prone tasks for a constant C depending only on the slope of $\hat{A}(\cdot)$ near the operating SNR.

Combining (i) (as a leading-order estimate) and (ii) with Eq. (16) yields the operational bridge

$$\mathbb{E}[\delta_t] \;\gtrsim\; \frac{(1-\gamma)}{C} \cdot D \cdot (1 - A_{\text{full}}), \tag{17}$$

where $\gtrsim$ denotes leading-order in the surrogate. The substitution $\bar{\alpha} \cdot \max_h |\mathcal{R}_h^{(\ell^*)}| \approx \alpha_{\mathcal{R}}$ is the source of the $\gtrsim$ rather than $\geq$. The conservative rigorous case replaces $\alpha_{\mathcal{R}}$ on the right-hand side with $\bar{\alpha} \cdot \max_h |\mathcal{R}_h^{(\ell^*)}| \leq \alpha_{\mathcal{R}}$ and drops the bridge to $(1 - A_{\text{full}})$ entirely. We use $\gtrsim$ for the operational case.

Eq. (17) is the heuristic bridge, in the leading-order sense above, between the observable D and Bui's dilution under the empirical headroom proxy. Eq. (9) in Section 5 packages this relationship as a threshold inequality. Substituting $D \geq \tau$ into Eq. (17) gives $\mathbb{E}[\delta_t] \gtrsim ((1-\gamma)\tau/C) \cdot (1 - A_{\text{full}})$, so any chosen threshold τ on D produces a corresponding $c_2 = (1-\gamma)\tau/C$. Choosing $\tau = c_1 \log(1/\gamma)$ with $c_1 = 1/J_{\text{early}}$,

$$\boxed{c_1 \;=\; \frac{1}{J_{\text{early}}}, \qquad c_2(\gamma) \;=\; \frac{(1-\gamma)\cdot\log(1/\gamma)}{C \cdot J_{\text{early}}}.} \tag{18}$$

J_{early} enters the threshold as a model-specific scale that makes c_1 commensurate with the empirical Jaccard floor across heads. On our test models $J_{\text{early}} \in [0.4, 0.7]$, giving $c_1 \in [1.4, 2.5]$. The empirical $\tau = 0.07$ is well below this, consistent with the constants being loose (per the $O(H)$ slack in Step 3 and the heuristic substitution in observation (i) above). Both constants carry an $O(H)$ multiplicative slack from the pairwise-to-H-wise step of Step 3, and the $\log(1/\gamma)$ factor in $c_1 \log(1/\gamma)$ is heuristic (gap-list item 3). The conservative no-amplification case drops the $\log(1/\gamma)$ factor and gives the threshold $D \geq c_1 = 1/J_{\text{early}}$ with rate $c_2(\gamma) = (1-\gamma)/(C \cdot J_{\text{early}})$ directly.

The converse direction (heuristic, sufficient-side only) The partition predictor's operational claim is the contrapositive: *small D ⇒ low dilution ⇒ eviction is risky*. We sketch two arguments and flag that neither is a separate necessity proof.

The first is a comment on the sufficient-side lower bound: under (A3) for the capacity-bound regime $\mathcal{C}$ ($\mathcal{R}_h^{(\ell)} = \mathcal{R}$ for all (ℓ, h)), the union and intersection coincide, so the right-hand side of Eq. (15) is zero. Since Eq. (15) is a *lower* bound on $\mathbb{E}[\delta_t]$, a zero right-hand side does not bound δ_t from above. It only says the sufficient lower bound becomes uninformative.

The second is an upper bound on δ_t that does provide some necessary-side content, but only inside the surrogate. Under (A1) with margin m satisfying the threshold of Lemma 1, and (A3) regime $\mathcal{C}$ so that $U_t = \mathcal{R}$ for every head, the pairwise top-k Jaccard between heads is $\geq J_{\text{early}}$ at every layer (no late-layer divergence), so $|\mathcal{U}|/|\mathcal{I}| \to 1$ and there is no residual distractor mass. The attention mass on $\mathcal{N}$ for any single head is bounded by the sub-Gaussian tail $O(\exp(-m^2/(4\sigma_z^2)))$ by Lemma 1, and Bui's dilution is $\delta_t \leq O(\exp(-m^2/(4\sigma_z^2)))$ in the surrogate. This is the surrogate-level converse; lifting it to real attention requires both a lower bound on m for capacity-bound tasks (which we have not verified) and the surrogate identification (A4) of U_t with $\mathcal{I}^{(\ell^*)}$. The boxed Eq. (9)/(18) is the sufficient direction only.

Empirical validation Table 26 reports the mean head-agreement drop D across the five dilution-prone tasks (qa_1, qa_2, vt, niah_multivalue, fwe) and the capacity-bound task (NIAH-MK3) on Qwen 1.5B, Qwen 3B, and Mistral. In every (model, context) cell, NIAH-MK3 has the smallest drop (including the cell where it is negative, Qwen 3B 16K). The *ordering* of tasks by D is preserved across architectures, and this is the same partition that the observable signal D induces. Cells with $D \approx 0$ have $\mathbb{E}[\delta_t]$ bounded by the noise level in the surrogate, and the per-input recovery rate ρ in Table 1 is ≤ 0.02 on all such cells. The observed ordering is therefore consistent with the sufficient direction of the bridge on this set of tasks. We do not promote this to a formal claim. The bridge inequality carries no probability guarantee at the measured margin (Appendix D), so consistency with it is uninformative. The evidence here supports the empirical ordering of tasks by D rather than the bridge itself, and that ordering is a pattern across many cells for which we report no corrected significance test. The dilution term δ_t is not observed on these tasks, and the necessity direction is established only in the surrogate.

What is *not* proved We report the gaps explicitly to clarify the scope. The first two are substantive and carry the analysis; the remainder are inherited from the exact-recovery route of Steps 1–5 and are dissolved rather than repaired by the expected-agreement analysis of Appendix E, which does not traverse the $\alpha_{\mathcal{R}}$, $\mathcal{U}_t$, value-projection, or $\hat{A}(\text{SNR})$ machinery they concern.

1. **The early-overlap ordering (A3′).** That real capacity-bound inputs have strictly lower early-layer head-set overlap than dilution-prone ones. This is the one substantive empirical assumption of the replacement analysis. It is measured rather than assumed (Appendix E), but it is not derived. The stronger common-late-set clause of the original (A3) is not required here and thus not claimed. That clause is architecture-dependent, holding on Qwen and Mistral and failing on the Llama architecture (Appendix A.3.3). The replacement analysis rests on the early ordering alone.
2. **The accuracy surrogate $\hat{A}(\text{SNR})$.** The constant C depends on the slope of $\hat{A}$ near the operating SNR, and on tasks where $\hat{A}$ is steep (precise retrieval) versus shallow (aggregation) that slope differs by an order of magnitude. No task-agnostic constant exists, and we report the empirical band instead. This enters only the scaling relation of Eq. (4), never the gate signal D.
3. **Inherited from the exact-recovery route.** The constants $c_1, c_2(\gamma)$ of Eq. (18) carry an $O(H)$ slack from the pairwise-to-H-wise step. The pairwise Jaccard is not the H-wise union/intersection ratio. The substitution $\bar{\alpha} \cdot \max_h |\mathcal{R}_h^{(\ell^*)}| \approx \alpha_{\mathcal{R}}$ is leading-order. The identification of Bui's $\mathcal{U}_t$ with $\mathcal{I}^{(\ell^*)}$ (A4) is a modelling choice. The $\log(1/\gamma)$ amplification is heuristic, with true rate $\Theta(\log(1/\gamma)/\sqrt{\log N})$. The scorer-symmetry and value-additivity steps of Step 4 are assumptions. Each of these is an artifact of routing the observable through set unions and intersections under exact recovery. The first-moment analysis of Appendix E computes the observable directly and does not use any of them.

The sketch above is therefore a *structural* proof of the sufficient direction under a margin condition the measurements do not support. We retain it as the motivation for the mechanism and give the analysis that holds at the measured margin in Appendix E.

E EXPECTED-AGREEMENT ANALYSIS AT FINITE MARGIN

The bridge of Appendix D carries no guarantee at our own measured parameters. This section gives an explanation of the same observable that is correct at the measured margin-to-noise ratio, and that predicts the sign of D on capacity-bound inputs rather than being contradicted by it. The first-moment result is exact under its model; the per-input variance material is flagged as leading-order.

E′.1 Why the exact-recovery route is uninformative here Two modelling choices, not a real barrier, produce this failure. First, Lemma 1 demands that *every* head's top-k set equal its private relevant set exactly, jointly over all HL heads and all T^2 position pairs. That is a worst-case union bound over $2H^2LT^2$ tuples, and it forces $\kappa \geq 2\sqrt{\log(2H^2LT^2/\beta)}$, which at our measured cell $(H, L, T) = (12, 28, \approx 3956)$ and $\beta = 1$ is $\kappa \geq 10.11$. The context length is not what drives this: $\sqrt{\log T} \approx 2.88$ at $T = 4096$, so the threshold comes from the H^2L head-and-layer multiplicity inside the logarithm. Second, the observable is then reconstructed from head-wise unions and intersections, which carries the $O(H)$ slack of Step 3. But the quantity the gate computes is a *first moment of a pairwise average*, and a first-moment analysis needs neither exact recovery nor the union/intersection detour.

E′.2 Model and selection probabilities Fix a layer and suppress its index. There are T key positions and H heads; head h has a relevant set $\mathcal{R}_h$ with $|\mathcal{R}_h| = k$, matching the deployed $k = 32 \sim |\mathcal{R}|$. Logits are $z_{h,i} = m \cdot \mathbf{1}[i \in \mathcal{R}_h] + \eta_{h,i}$ with $\eta_{h,i}$ i.i.d. continuous, symmetric, of scale σ, independent across (h, i), and $\kappa := m/\sigma$ as before. Head h selects $S_h = \mathrm{top}k(z_{h,\cdot})$, $|S_h| = k$. Define the per-head marginal selection probabilities

$$p_R := \Pr[i \in S_h \mid i \in \mathcal{R}_h], \qquad p_N := \Pr[i \in S_h \mid i \notin \mathcal{R}_h]. \tag{19}$$

Both depend only on (k, T, κ) and on no other head. The fixed-budget identity $kp_R + (T-k)p_N = k$ pins $p_N = k(1 - p_R)/(T - k)$. At $\kappa = 0$ signal and noise logits are identically distributed, so $p_R = p_N = k/T$; raising m raises every relevant logit pointwise and top-k is monotone in the logits, so $p_R - p_N$ is nondecreasing in κ. Hence $p_R > p_N$ for every $\kappa > 0$, with $p_R - p_N \to 0$ as $\kappa \to 0$. Nothing here assumes exact recovery.

For two heads write $s := |\mathcal{R}_h \cap \mathcal{R}_{h'}|$ for the shared relevant count. The operative assumption is a single ordering, which replaces (A3):

Assumption 1 (Early-overlap ordering, A3′′)**.** Capacity-bound inputs have strictly lower mean pairwise early-layer relevant-set overlap than dilution-prone inputs, $s^{\mathcal{C}}_{\mathrm{early}} < s^{\mathcal{D}}_{\mathrm{early}}$.

No assumption is made about the late bin. In particular the retrieval idealisation $s^{\mathcal{C}}_{\mathrm{late}} = k$ of (A3) is *not* assumed; the finite-κ agreement below reproduces the measured sub-unit late agreement instead.

E′.3 Expected agreement is exactly linear in overlap Let $X := |S_h \cap S_{h'}|$. Since $|S_h| = |S_{h'}| = k$, the union is deterministic given X, $|S_h \cup S_{h'}| = 2k - X$, so the pairwise Jaccard $J = X/(2k - X)$ is a fixed increasing function of X.

Lemma 2 (Expected intersection, exact)**.** *Under the model of E′.2,*

$$\mathbb{E}[X] \;=\; s\,p_R^2 \;+\; 2(k-s)\,p_R p_N \;+\; (T - 2k + s)\,p_N^2, \tag{20}$$

and consequently $\mathrm{d}\mathbb{E}[X]/\mathrm{d}s = (p_R - p_N)^2 > 0$*, a constant independent of s and of the regime.*

Proof. The two heads select independently, so $\mathbb{E}[X] = \sum_i \Pr[i \in S_h] \Pr[i \in S_{h'}]$. Partition the T positions by membership: s shared-relevant positions contribute p_R^2 each, $2(k - s)$ singly-relevant positions contribute $p_R p_N$, and the remaining $T - 2k + s$ contribute p_N^2. Differentiating in s gives $p_R^2 - 2p_R p_N + p_N^2 = (p_R - p_N)^2$, which does not depend on s because p_R and p_N are single-head marginals. □

Since p_R, p_N are per-head constants shared by both regimes, the only regime-dependent input to Eq. (20) is the overlap s. Write $\bar{a} := \mathbb{E}[X]/(2k - \mathbb{E}[X])$ for the pooled agreement, a ratio of expectations that tracks the mean-of-ratios a_ℓ up to Jensen slack and may be adopted as the target quantity to make the statement exact. Then $\bar{a}$ is strictly increasing in $\mathbb{E}[X]$, hence in s, with slope $\partial\bar{a}/\partial\mathbb{E}[X] \in [1/(2k), 2/k]$ on $\mathbb{E}[X] \in [0, k]$.

Note what is absent from Lemma 2: the slope $(p_R - p_N)^2$ contains no factor of H, because p_R and p_N are single-head marginals. The average over the $\binom{H}{2}$ pairs is a mean of per-pair expectations each governed by that same H-free slope, so the operative constant is $O(1)$ in H. The $O(H)$ slack of Eq. (18) is an artifact of the union/intersection detour and does not arise here.

Theorem 1 (Class separation at finite margin)**.** *With $\bar{a}_\ell$ the pooled agreement in bin ℓ, the class gap decomposes exactly as*

$$\mathbb{E}[D \mid \mathcal{D}] - \mathbb{E}[D \mid \mathcal{C}] \;=\; \big(\bar{a}^{\mathcal{D}}_{\text{early}} - \bar{a}^{\mathcal{C}}_{\text{early}}\big) \;+\; \big(\bar{a}^{\mathcal{C}}_{\text{late}} - \bar{a}^{\mathcal{D}}_{\text{late}}\big), \tag{21}$$

and each bracket is a Lemma 2 image of the corresponding overlap difference. Under Assumption 1 the first bracket is strictly positive for every $\kappa > 0$ and carries the gap.

Proof. Expand $\mathbb{E}[D \mid \cdot] = \bar{a}_{\text{early}} - \bar{a}_{\text{late}}$ and subtract; Eq. (21) is then an identity. By the mean-value theorem and Lemma 2, each bracket equals $g'(p_R - p_N)^2\,\Delta s$ for the corresponding overlap difference Δs and some $g' \in [1/(2k), 2/k]$. The early bracket is positive by Assumption 1 together with $p_R > p_N$, which holds at every $\kappa > 0$. No step invokes exact recovery or a union bound, so the separation remains informative precisely where Lemma 1 fails. □

The sign of the late bracket is not fixed, and we do not claim it is: measured per-task decompositions on Qwen2.5-1.5B 4K give an early term that is positive and large in every case ($+0.062$ to $+0.198$) against a late term that is small and negative on three of four tasks. The separation is an early-layer effect.

E′.4 The sign of D on capacity-bound inputs The exact-recovery route of Appendix F.2 forced $a^{\mathcal{C}}_{\text{late}} = 1$, giving $D \le a_{\text{early}} - 1 \le 0$, which Table 26 contradicts in five of seven cells. Under the finite-κ model the observed behaviour is what the model predicts. Because $p_R < 1$, agreement never reaches 1 at any depth, so a task whose heads specialise early and stay specialised has both $\bar{a}_{\text{early}}$ and $\bar{a}_{\text{late}}$ low, and

$$\mathbb{E}[D \mid \mathcal{C}] \;=\; \bar{a}^{\mathcal{C}}_{\text{early}} - \bar{a}^{\mathcal{C}}_{\text{late}} \tag{22}$$

is small, with a sign set by the residual early-minus-late difference rather than pinned to be non-positive. This is the flat, low profile the per-layer logs show on NIAH-MK3 (agreement near 0.3 at every depth, Appendix A.3.3), not a near-unit late agreement perturbed downward. The relaxed $\Pr[D \le o(1)]$ form in Appendix F.2 *allows* a small positive drop on capacity-bound inputs, and the analysis here *derives* it. The qualitative claim that survives is a separation of means, $\mathbb{E}[D \mid \mathcal{D}] > \mathbb{E}[D \mid \mathcal{C}]$, carried by the early bracket, rather than a sign test on D.

E′.5 What is measured rather than fitted The analysis above needs no fitted constant. We attempted to estimate p_R and p_N from the released per-pair logs and report that they are *not* point-identifiable: within a single layer the per-pair mean Jaccard ranges from 0.05 to 0.55, so a single global (p_R, p_N) cannot fit heads that heterogeneous, and the per-pair overlap s is latent because we have no ground-truth relevant sets. The separation claims rest instead on the qualitative monotonicity of Lemma 2 and on directly measured agreement.

E′.6 Per-input separability (leading order) The gate thresholds D per input, so per-input behaviour depends on the dispersion of D and not only its mean. Writing $X = \sum_i B^h_i B^{h'}_i$ with the two heads independent, fixed-k selection makes $\{B^h_i\}$ negatively associated within a head, so the products inherit non-positive covariance and $\mathrm{Var}(X) \le \mathbb{E}[X]$. A single head pair therefore separates the regimes at signal-to-noise $O(1)$, which is weak. But D averages over $\binom{H}{2}$ pairs per layer with H independent head units, so $\mathrm{Var}(\bar{a}_\ell)$ shrinks like $1/H$ and the per-input separability improves accordingly. We flag this as a leading-order estimate using a Gaussian approximation and a $1/H$ variance heuristic; a rigorous concentration bound over correlated head pairs is open.

One structural point falls out. The head count H enters the two analyses with opposite sign: it inflates the union bound of the exact-recovery route, which is what makes that route uninformative, and it reduces the variance of the averaged statistic, which is what makes the per-input gate usable. The $O(LH^2k)$ cost reported in Section 6 as a deployment burden is, statistically, that variance reduction.

F SHARPENED DEFERRED PROOFS

This section upgrades three results that Appendices C and D left as sketches: (i) the leading constant and a two-sided bound for the scaling formula (4); (ii) the necessary (converse) direction of the D-to-dilution bridge; and (iii) the status of the useful-set identification (A4). Each subsection ends with an explicit "What this does and does not establish" paragraph so the paper's claims can be scoped exactly. Throughout we keep the notation of Appendices C–D: $\alpha_{\mathcal{R}}$ is the pre-eviction signal mass, γ the noise-retention ratio, $\rho_{\mathcal{R}}, \rho_{\mathcal{N}}$ the retained mass fractions, δ_t Bui's dilution (6), D the head-agreement drop (3), σ_z the pre-softmax logit-noise scale, m the logit margin, H, L, T the head, layer and length counts, and $\hat{A}(\cdot)$ the strictly increasing, $[0,1]$-valued, unit-saturating accuracy surrogate.

F.1 ITEM 1: THE EXACT LEADING CONSTANT AND A TWO-SIDED SCALING BOUND

The sketch in Appendix C obtains $\rho_{\mathrm{KV}} \le K\,(1 - A_{\mathrm{full}})\,p_{\mathrm{recov}} + K/4$ by a union bound over the K-point budget grid $\mathcal{B} = \{b_1, \dots, b_K\}$ followed by a covariance residual. We show that the grid factor K is an artifact of an unnecessary union step, that the correct leading constant is exactly 1, and that a distribution-free *two-sided* bound holds with explicit endpoints; the only genuinely unresolved quantity is a covariance, for which we give matching, attainable bounds.

Events We work inside the surrogate of Appendix C, where the per-input improvement event at a single budget b factorises (their displayed identity) into a headroom indicator and a recovery indicator. Fix the prompt distribution and, for a random prompt x, define the two Bernoulli random variables

$$\begin{aligned} A(x) &:= \mathbf{1}\big[\hat{A}(\mathrm{SNR}(b_{\max}, x)) < 1\big] && \text{(headroom)},\\ B(x) &:= \mathbf{1}\Big[\exists\, b \in \mathcal{B} :\ \mathrm{SNR}_{\mathrm{post}}(b, x) > \mathrm{SNR}_{\mathrm{pre}}(x)\Big] && \text{(recovery)}. \end{aligned} \tag{23}$$

The operational recovery rate is $\rho_{\mathrm{KV}} = \Pr_x[A(b^*(x), x) > A(b_{\max}, x)]$ with $b^*(x) = \arg\max_b A(b, x)$ (Eq. (1)). Because b^* is the argmax over the grid, "improvement at b^*" is exactly "improvement at some $b \in \mathcal{B}$," and the single-budget surrogate factorisation gives

$$\big\{A(b^*(x), x) > A(b_{\max}, x)\big\} \;=\; \bigcup_{b \in \mathcal{B}} \big(A(x) \cap B_b(x)\big) \;=\; A(x) \cap \Big(\bigcup_{b \in \mathcal{B}} B_b(x)\Big) \;=\; A(x) \cap B(x), \tag{24}$$

where $B_b(x) = \{\mathrm{SNR}_{\mathrm{post}}(b, x) > \mathrm{SNR}_{\mathrm{pre}}(x)\}$ and $B = \bigcup_b B_b$. The headroom event does not depend on b, so it pulls out of the union: the union over budgets lives *inside* B. We set $p_{\mathrm{recov}} := \Pr_x[B]$ (recovery at the best budget) and note $\Pr_x[A] = 1 - A_{\mathrm{full}}$ by definition of the surrogate headroom.

Theorem 2 (Exact leading constant, two-sided scaling bound)**.** *With the definitions in* (23)–(24) *and* $p_{\mathrm{recov}} := \Pr_x[B]$*, the surrogate per-input recovery rate satisfies the* exact *identity*

$$\rho_{\mathrm{KV}} \;=\; (1 - A_{\mathrm{full}})\,p_{\mathrm{recov}} \;+\; \mathrm{Cov}_x\big(A, B\big). \tag{25}$$

What matters here is not the identity (25)*, which is merely the definition of covariance for two Bernoullis, but the observation that once* p_{recov} *is defined as the* union *probability* $\Pr_x[\bigcup_b B_b]$ *(the quantity* b^* *actually selects) rather than the single-budget* $\max_b \Pr_x[B_b]$*, the budget-grid factor* K *of Appendix C is absorbed into* p_{recov} *rather than multiplying it. This factor* K *reappears inside* p_{recov} *whenever the recovery budgets are near-disjoint (Corollary 1(ii)). The covariance is bounded, by*

$$0 \;\le\; \mathrm{Cov}_x(A, B) \;\le\; \min\{1 - A_{\mathrm{full}},\, p_{\mathrm{recov}}\} - (1 - A_{\mathrm{full}})\,p_{\mathrm{recov}}, \tag{26}$$

where the lower bound requires the association hypothesis of Lemma 3 below, and the upper bound (Fréchet) is unconditional. Equivalently,

$$\boxed{(1 - A_{\mathrm{full}})\,p_{\mathrm{recov}} \;\le\; \rho_{\mathrm{KV}} \;\le\; \min\{\,1 - A_{\mathrm{full}},\; p_{\mathrm{recov}}\,\}.} \tag{27}$$

The lower bound is attained iff $A \perp B$*. The upper bound is attained iff* A *and* B *are Fréchet–Hoeffding comonotone (one event a.s. contains the other).*

Proof. Identity (25) is the definition of covariance applied to the two Bernoulli variables in (23), using $\rho_{\mathrm{KV}} = \Pr_x[A \cap B] = \mathbb{E}_x[AB]$ from (24) and $\mathbb{E}_x[A] = 1 - A_{\mathrm{full}}$, $\mathbb{E}_x[B] = p_{\mathrm{recov}}$. For (26), the upper endpoint is the Fréchet–Hoeffding inequality $\Pr[A \cap B] \le \min\{\Pr[A], \Pr[B]\}$, valid for any two events; subtracting the product gives the stated covariance bound and the right side of (27). The lower endpoint $\mathrm{Cov}_x(A, B) \ge 0$ is Lemma 3. Attainment: independence gives equality in the lower bound by construction; comonotonicity ($A \subseteq B$ or $B \subseteq A$ a.s.) gives $\Pr[A \cap B] = \min\{\Pr[A], \Pr[B]\}$, the upper endpoint. □

Lemma 3 (Association of headroom and recovery)**.** *Suppose the noise-isotropy assumption (A1) of Appendix C holds, so that both A and B are non-increasing functions of the single scalar driver $\alpha_{\mathcal{R}}(x)$ (equivalently, $\mathbb{E}[A \mid \alpha_{\mathcal{R}}]$ and $\mathbb{E}[B \mid \alpha_{\mathcal{R}}]$ are non-increasing in $\alpha_{\mathcal{R}}$ and monotone in the same direction). Then $\mathrm{Cov}_x(A, B) \ge 0$.*

Proof. If $A = f(\alpha_{\mathcal{R}})$ and $B = g(\alpha_{\mathcal{R}})$ with f, g both non-increasing, Chebyshev's association (sum) inequality gives $\mathbb{E}[fg] \ge \mathbb{E}[f]\mathbb{E}[g]$, i.e. $\mathrm{Cov} \ge 0$. If A depends on an auxiliary driver ζ independent of $\alpha_{\mathcal{R}}$ (e.g. a parametric-knowledge covariate), write $\mathrm{Cov}(A, B) = \mathrm{Cov}(\mathbb{E}[A \mid \alpha_{\mathcal{R}}], B)$ by the tower rule and $B = g(\alpha_{\mathcal{R}})$; the same monotone-in-a-common-scalar argument applies to the conditional mean, so the sign is preserved. □

The monotonicity of A in $\alpha_{\mathcal{R}}$ is immediate (larger signal mass ⇒ larger $\mathrm{SNR}_{\mathrm{pre}}$ ⇒ smaller headroom). The monotonicity of B follows from Eq. (10): the attainable SNR gain from removing noise mass shrinks as $\alpha_{\mathcal{R}} \to 1$ (less noise to remove), so recovery is less likely at larger $\alpha_{\mathcal{R}}$.

Where the grid factor K comes from, and why it is not real The K flagged in Appendix C comes from the step $\Pr_x[\bigcup_b B_b] \le \sum_b \Pr_x[B_b] \le K \max_b \Pr_x[B_b]$ followed by identifying $\max_b \Pr_x[B_b]$ with p_{recov}. Defining p_{recov} as the union probability $\Pr_x[\bigcup_b B_b]$ instead, which is the quantity the operational $b^*(x)$ actually selects, removes the union step entirely. The overcount incurred by the discarded step is, exactly,

$$\sum_{b \in \mathcal{B}} \Pr_x[B_b] \;-\; \Pr_x\Big[\bigcup_b B_b\Big] \;=\; \sum_{j \ge 2} (-1)^{j} \sum_{|S| = j, S \subseteq \mathcal{B}} \Pr_x\Big[\bigcap_{b \in S} B_b\Big] \;\ge\; 0, \tag{28}$$

the inclusion–exclusion tail. Its size is what makes the union bound tight or loose:

Corollary 1 (Tightness characterisation of the budget-grid bound)**.** *Let $p^\sharp := \max_b \Pr_x[B_b]$ be the single-budget recovery probability.*

(i) (Monotone gain profile ⇒ no K factor.) *If the SNR gain $g(b, x) := \rho_{\mathcal{R}}(b, x)^2/\gamma(b, x)$ is monotone in b for a.e. x (as it is for a calibrated scorer, since keeping more tokens raises both $\rho_{\mathcal{R}} \to 1$ and $\gamma \to 1$ so $g \downarrow 1$), then the events $\{B_b\}$ are nested and $p_{\mathrm{recov}} = \Pr_x[\bigcup_b B_b] = p^\sharp$. The genuine leading factor is 1; the union bound $K\,p^\sharp$ overcounts by $(K-1)p^\sharp$.*

(ii) (Disjoint gain profile ⇒ K is real.) *If instead the $\{B_b\}$ are pairwise disjoint (each input recovers at exactly one budget) and rare ($Kp^\sharp \ll 1$), then $p_{\mathrm{recov}} = \sum_b \Pr_x[B_b]$ is genuinely $\Theta(K)$ times larger than $p^\sharp$; recovery at the best budget really is K-fold more frequent than at a fixed budget, and the factor K is not slack but signal.*

In both cases Theorem 2 holds with constant 1 once p_{recov} is the union probability; the two regimes differ only in how much larger p_{recov} is than the single-budget $p^\sharp$.

Proof. (i) If $g(\cdot, x)$ is monotone then $\{b : g(b, x) > 1\}$ is an interval anchored at the aggressive end of the grid, so $B_b(x) \subseteq B_{b'}(x)$ whenever b' is more aggressive than b; the B_b are nested and their union equals the loosest single event, of probability $p^\sharp$. (ii) For disjoint events $\Pr[\bigcup_b B_b] = \sum_b \Pr[B_b]$ exactly; rareness makes this $\approx Kp^\sharp$. □

Because our scorers are calibrated (the top-k retention curve is monotone in b), the operative regime is Corollary 1(i) and the constant is 1. This is the statement Appendix C defers to; the K and $K/4$ terms there serve only to isolate the step that introduces the slack.

What this does and does not establish *Establishes:* inside the deterministic-SNR surrogate, the scaling law holds as the *exact* identity (25) with leading constant precisely 1 (no budget-grid inflation), and the two-sided bound (27) with fully explicit, distribution-free endpoints; the only free quantity is the covariance $\mathrm{Cov}_x(A, B) \in [0,\ \min\{\Pr A, \Pr B\} - \Pr A \Pr B]$, and both endpoints are attainable, so no tighter distribution-free constant exists. *Does not establish:* (a) the surrogate itself, that accuracy is a deterministic (or fixed-conditional) function of head SNR, is assumed, not proven, so the constant "1" is exact only relative to that surrogate; (b) the covariance is pinned only to an interval, not a value, because it depends on the joint law of headroom and recovery, which is task- and model-specific; Lemma 3 fixes its *sign* (non-negative) under single-driver monotonicity but not its magnitude; (c) when headroom is driven partly by an $\alpha_{\mathcal{R}}$-independent factor (parametric-knowledge failures, e.g. the QA cells in Table 1), recovery collapses ($p_{\mathrm{recov}} \to 0$) so both bounds pinch $\rho_{\mathrm{KV}} \to 0$ despite headroom, consistent with the observed QA exceptions, but a prediction of the sign structure, not a proof about those tasks. The empirically observed ratio band $[0.25, 0.55]$ is thus an estimate of $p_{\mathrm{recov}} + \mathrm{Cov}/(1 - A_{\mathrm{full}})$, not of a universal constant. A fourth point concerns the leading constant itself. The value 1 rests on Corollary 1(i), which requires the gain $g(b, x) = \rho_{\mathcal{R}}^2/\gamma$ to be monotone in the budget b, and this is a property of the gain rather than of the retention curve, which is monotone by construction. Where the accuracy profile has an interior optimum in b, the recovery events are not nested and the constant is the $\Theta(K)$ of Corollary 1(ii). We report the constant 1 under the monotone case that our calibrated scorers exhibit, and treat the monotonicity of g as an empirical property of those scorers rather than a derived one.

F.2 Item 2: why the margin route to the necessary direction fails

Appendix D proves the sufficient direction (large D forces large δ_t) and states that the converse "capacity-bound $\Rightarrow$ small D" needs "a margin condition we have not verified." We verified it. It is false at our own parameters, and it fails in a way that rules out the whole exact-recovery route rather than merely loosening a constant. We report that measurement here because it is what motivates the finite-margin analysis of Appendix E, which reaches the necessary direction without any margin condition.

What the margin route would require Recall $z_{h,i}^{(\ell)} = (q_h^{(\ell)})^\top k_i + \eta_{h,i}^{(\ell)}$ with σ_z-sub-Gaussian noise (A2), and the logit margin $m = m_+ - m_-$ (A1). Define the *margin-to-noise ratio* $\kappa := m/\sigma_z$. To force late-layer heads onto a common top-k set, and hence $a_{\mathrm{late}} \to 1$ and $D \le 0$ on capacity-bound inputs, one needs every common relevant position to beat every noise position at the late bin with high probability jointly, which is the threshold of Lemma 1:

$$\kappa \ \ge\ 2\sqrt{\log\bigl(2H^2LT^2/\beta\bigr)}. \tag{29}$$

At our measured cell's $(H, L, T) = (12, 28, \approx 3956)$ this is $\kappa \ge 10.11$ for $\beta < 1$. Note that the context length is not what drives the requirement: $\sqrt{\log T} \approx 2.88$ at $T = 4096$, so the threshold comes from the H^2L head-and-layer multiplicity inside the logarithm, i.e. from the union bound over $2H^2LT^2$ tuples rather than from the retrieval problem itself.

The margin-to-noise ratio, measured On Qwen2.5-1.5B at RULER 4K, NIAH-MK3, over 20 inputs and roughly 10,744 (layer, head, query) triples per input, we take the top-k pre-softmax logits at each site as the relevant set $\mathcal{R}$ and the remainder as noise $\mathcal{N}$, and report $\kappa = (\overline{\mathcal{R}} - \overline{\mathcal{N}})/\mathrm{sd}(\mathcal{N})$. The measured median is $\kappa = 3.33$, roughly $3\times$ too small, though individual heads reach 77.66.

The requirement is therefore *uninformative at the paper's own parameters*, not merely at parameters we consider implausible. The conclusion is robust to the choice of split: sweeping $k \in \{8, 16, 32, 64\}$ gives median $\kappa = 4.01, 3.68, 3.33, 2.99$, monotone decreasing in k, so no choice of split in this range rescues it, and the largest value seen is $0.40\times$ the requirement. It also holds across architectures: repeating at the deployed $k = 32$ on a second architecture gives $\kappa = 3.71$ (Mistral-7B). The magnitude of the shortfall is not marginal, since a factor of three in κ is a factor of $e^{(10.11^2 - 3.33^2)/4} > 10^9$ in β.

Why κ is not the route to the separation On a dilution-prone task the same measurement gives $\kappa = 3.40$ (VT), close to the 3.33 on the capacity-bound one. On this pair the margin-to-noise ratio therefore does not separate the two classes, so it is not the quantity that carries the partition,

whatever threshold it is held to. In the surrogate κ governs the top-k recovery step rather than the class separation, and the near-equal values are consistent with that role. An analysis of the partition routed through a margin condition on κ is therefore not the right route to the separation, independently of the scale gap.

A second, independent check points the same way. The exact-recovery route predicts $D \leq 0$ on capacity-bound inputs. Measured D on NIAH-MK3 is mildly *positive* in five of seven cells ($+0.040$, $+0.060$, $+0.053$, $+0.019$, $+0.073$) and negative only on Qwen2.5-3B at both contexts (-0.004, -0.021), as reported in Table 26. Relaxing the idealisation that late heads share one common set, to agreement on an $O(1)$ neighbourhood, weakens the conclusion to $\Pr[D \leq o(1)]$, which the measurements satisfy but which no longer distinguishes the classes, since it permits a small positive drop on either side.

What we do instead We therefore do not state a converse in the margin form. Appendix E gives the necessary direction as a monotonicity statement at finite κ: expected agreement is a strictly increasing function of head-set overlap with slope $(p_R - p_N)^2 > 0$ at *every* $\kappa > 0$ (Lemma 2). Because agreement is strictly increasing in head-set overlap, the early-layer overlap difference that separates the classes carries the same sign as the early-agreement difference that D reads, and under the measured early-overlap ordering this term is strictly positive at every $\kappa > 0$ (Theorem 1). The same finite-κ model leaves D on capacity-bound inputs small with a sign not pinned to be non-positive (Appendix E), which is consistent with the small positive values in Table 26 rather than contradicted by them. The class separation the gate relies on remains empirical (Table 26) and stands on its own.

F.3 Item 3: the A4 useful-set identification

Assumption (A4) of Appendix D identifies Bui's useful set U_t (a residual-stream, decoding-step object) with the head-common signal set $\mathcal{I}^{(\ell^*)} = \bigcap_h \mathcal{R}_h^{(\ell^*)}$ (a single-layer attention object), and the text flags it as "a modelling choice, not a derived equality." We show that a clean exact equality cannot be proven, but a *quantitative approximate* identification can: under a named consensus–usefulness assumption plus ε-closeness of heads, the two sets induce dilution values that agree up to $O(\varepsilon)$, with an explicit constant. We therefore replace the "derived equality" language with a proved approximation plus a named residual assumption.

Set-up At the divergence layer ℓ^* write the head-private top-k sets $\mathcal{R}_h := \mathcal{R}_h^{(\ell^*)}$, their intersection $\mathcal{I} := \bigcap_h \mathcal{R}_h$ and union $\mathcal{U} := \bigcup_h \mathcal{R}_h$. For a set $\mathcal{S}$ define its induced dilution $\delta(\mathcal{S}) := 1 - \sum_{i \in \mathcal{S}} \alpha_{t,i}$, so Bui's is $\delta^{(U)} := \delta(U_t)$ and ours is $\delta^{(I)} := \delta(\mathcal{I})$. Let $\bar{\alpha}_{\max} := \max_i \alpha_{t,i}$.

Assumption 2 (ε-closeness of heads)**.** The late-layer heads are ε-close in their top-k sets: $\max_{h,h'} |\mathcal{R}_h \triangle \mathcal{R}_{h'}| \leq \varepsilon k$, equivalently pairwise Jaccard $\geq (1 - \varepsilon/2)/(1 + \varepsilon/2) = 1 - O(\varepsilon)$.

Assumption 3 (Consensus–usefulness, CUA)**.** The useful set is sandwiched between the head-consensus and head-union sets: $\mathcal{I} \subseteq U_t \subseteq \mathcal{U}$. The right inclusion ($U_t \subseteq \mathcal{U}$) holds whenever routing to the residual stream is through attention only, so a position no head attends to cannot affect the next-token margin. The left inclusion ($\mathcal{I} \subseteq U_t$) is the substantive content: a position every head attends to is margin-preserving.

Proposition 1 (Quantitative approximate identification)**.** *Under Assumptions 2–3,*

$$\left| \delta_t^{(U)} - \delta_t^{(I)} \right| \;\leq\; \sum_{i \in \mathcal{U} \setminus \mathcal{I}} \alpha_{t,i} \;\leq\; H(H-1)\,\varepsilon\, k\, \bar{\alpha}_{\max} \;=\; O\big(H^2 \varepsilon\big), \tag{30}$$

where the last step uses $\bar{\alpha}_{\max} = \Theta(1/k)$ in the high-SNR softmax. In particular, as $\varepsilon \to 0$ (heads perfectly agree) the two dilution values coincide: $\delta_t^{(U)} \to \delta_t^{(I)}$.

Proof. By Assumption 3, $U_t \triangle \mathcal{I} \subseteq \mathcal{U} \setminus \mathcal{I}$, because $\mathcal{I} \subseteq U_t \subseteq \mathcal{U}$ forces every element of $U_t \triangle \mathcal{I} = U_t \setminus \mathcal{I}$ to lie in $\mathcal{U} \setminus \mathcal{I}$. Since $\delta(\mathcal{S})$ is a signed sum of attention masses over $\mathcal{S}$, $|\delta_t^{(U)} - \delta_t^{(I)}| = \left|\sum_{i \in U_t} \alpha_{t,i} - \sum_{i \in \mathcal{I}} \alpha_{t,i}\right| \leq \sum_{i \in U_t \triangle \mathcal{I}} \alpha_{t,i} \leq \sum_{i \in \mathcal{U} \setminus \mathcal{I}} \alpha_{t,i}$, the first inequality of (30). For the cardinality bound, for each head h, $|\mathcal{R}_h \setminus \mathcal{I}| \leq \sum_{h' \neq h} |\mathcal{R}_h \setminus \mathcal{R}_{h'}| \leq \sum_{h' \neq h} |\mathcal{R}_h \triangle \mathcal{R}_{h'}| \leq (H-1)\varepsilon k$

by Assumption 2. Since $\mathcal{U} \setminus \mathcal{I} = \bigcup_h (\mathcal{R}_h \setminus \mathcal{I})$, $|\mathcal{U} \setminus \mathcal{I}| \le \sum_h |\mathcal{R}_h \setminus \mathcal{I}| \le H(H-1)\varepsilon k$. Bounding each mass by $\bar{\alpha}_{\max}$ gives the middle inequality; $\bar{\alpha}_{\max} = \Theta(1/k)$ gives the $O(H^2\varepsilon)$ order. □

Remark 1 (When the identification fails). The left inclusion of Assumption 3 ($\mathcal{I} \subseteq U_t$) can fail, and this is exactly the case the equality cannot cover: if a *single* specialised head carries the decisive value for the next-token margin, then a position in $\mathcal{R}_{h_0} \setminus \mathcal{I}$ (attended by that one head, absent from the consensus) can be margin-critical, so $U_t \not\subseteq \mathcal{U} \setminus (\mathcal{R}_{h_0} \setminus \mathcal{I})$ and $U_t \neq \mathcal{I}$ even at $\varepsilon = 0$ of the *other* heads. This is the single-decisive-head regime, precisely a capacity-bound, single-needle retrieval situation, where one retrieval head dominates. That regime is not hypothetical: it is the *retrieval-head* phenomenon documented by Wu et al. (2025), who show that a small, sparse and largely input-independent set of heads carries long-context factual recall, and that ablating those heads destroys retrieval while leaving other capabilities intact. Their analysis and ours agree on the mechanism and differ in what is done with it: they identify *which* heads matter, whereas we read a consensus statistic over all heads and treat its collapse as evidence that no subset of the cache is safe to drop. Their result also predicts our failure mode, since a single decisive head is invisible to a pairwise-agreement average. The approximation of Proposition 1 is therefore reliable in the dilution-prone regime (many heads, diffuse consensus, no single decisive head) and degrades exactly where the paper already declines to apply the bridge (capacity-bound).

What this does and does not establish *Establishes:* the identification of U_t with $\mathcal{I}^{(\ell^*)}$ is not an exact derived equality, but under the named Assumptions 2–3 the two sets induce *dilution values* that agree up to $O(H^2\varepsilon)$ in attention mass (Proposition 1), with an explicit constant $H(H-1)k\bar{\alpha}_{\max}$; so wherever heads are ε-close, using $\mathcal{I}^{(\ell^*)}$ in place of U_t perturbs Bui's δ_t by a controlled $O(\varepsilon)$ term rather than by an uncontrolled amount. *Does not establish:* (a) the substantive left inclusion $\mathcal{I} \subseteq U_t$ (CUA); it is named as Assumption 3, justified by the additive-routing picture, and shown in Remark 1 to fail precisely in the single-decisive-head (capacity-bound) regime, so it is an assumption about the model's read-out, not a theorem; (b) any control of the H^2 prefactor, which is a worst-case union count and is loose when residuals overlap across heads (the same $O(H)$-slack issue as Step 3 of Appendix D); (c) a bound when ε is $\Theta(1)$ (genuinely divergent heads, the dilution-prone late layers themselves), there the approximation is uninformative and $\mathcal{I}$ and U_t are related only through the CUA sandwich, not quantitatively. The measured late-layer agreement is itself $\Theta(1)$ rather than close to one: on Qwen2.5-1.5B the late-bin agreement on NIAH-MK3 is $a_{\text{late}} \approx 0.28$ (Appendix A.3.3), so the implied ε is of order one and the $O(H^2\varepsilon)$ term does not bind at the values we measure. The approximate identification is therefore applicable only where late-layer heads are close in their top-k sets, a condition the measurements do not meet. The honest scoping is thus: **A4 is a named modelling assumption (CUA), under which we prove an $O(H^2\varepsilon)$ approximate identification of the dilution values**, not a derived equality; the word "derived equality" in Appendix D should be read as this approximation.

G Reproduction notes

This section is the reproduction anchor cited by the Reproducibility Statement, and documents the implementation details that matter for exact reproduction, all of which the released code implements. Draw-to-draw and code-path/hardware variation are bounded separately in Appendices A.6.3 and A.3.6. The cache slicing uses explicit `position_ids` to preserve RoPE alignment, since the naive attention-mask trick produces a silent position shift on evicted positions. Long-context (16K) runs use SDPA for prefill and switch to eager for the $w \times T$ observation-window scoring pass to avoid the $O(T^2)$ attention-matrix allocation on 80GB. Head agreement uses Jaccard similarity of top-32 key sets per head pair, averaged over pairs. The deployed gate takes the early bin as the first $\lfloor L/3 \rfloor$ layers and the late bin as the *last* $\lfloor L/3 \rfloor$, i.e. $[0, \lfloor L/3 \rfloor)$ and $[L - \lfloor L/3 \rfloor, L)$, so the two bins are equal in size and, when $3 \nmid L$, the middle band is one layer wider than a strict partition into thirds. Table 26 instead reports the strict-thirds variant $[\lfloor 2L/3 \rfloor, L)$ for the late bin. The two agree whenever $3 \mid L$ (Qwen2.5-3B, $L = 36$) and differ by at most 0.004 in mean D elsewhere (Qwen2.5-1.5B $L = 28$: 0.044 vs 0.040; Llama-3.1-8B $L = 32$: 0.071 vs 0.073), which changes no gate decision at $\tau = 0.07$. Every gated number in the paper uses the equal-size convention that the released code implements. Decoding is greedy throughout, so the gated outcome of any input is exactly reproducible from its recorded head-agreement drop plus the plain-eviction and full-KV outcomes. We use this identity to evaluate gate thresholds post-hoc, and one matrix cell (Qwen2.5-

3B 4K SnapKV, executed at $\tau = 0.04$) is reported at $\tau = 0.07$ via this exact re-evaluation. The threshold variants this identity was used to compare appear in Table 8, and the calibration recipe that selects τ is given in Appendix A.3.4.

Two separate audits check the budget 1.0 boundary at different levels, and neither subsumes the other. At the level of scored outcomes, Appendix A.6 verifies on all 16 matrix cells and the four 16K cells that the gated arm keeps the whole cache and matches plain eviction on every one of the $5{,}120$ rows. At the level of raw generated text, a pipeline audit on six examples verifies that eviction at budget 1.0 matches clean prefill-decode token for token, with a trailing-period mismatch on one of the six. The first establishes that $r(x) \geq 0$ holds as an identity, and the second checks the decode path itself on a small sample. The released logs include the audit script.

H USE OF LARGE LANGUAGE MODELS

Large language models were used as general-purpose assistants during this project: for literature search and triage, for drafting and copy-editing prose, and as coding aids when writing and analyzing the experiment scripts. All research ideas, hypotheses, experimental designs, and claims are the authors' own. Every reported number was produced by the released code on the stated models and data, and the authors verified the experiments and their reporting. LLMs were not used as a source of scientific facts or citations without verification against primary sources.