AUDITING LOCAL ATTENTION FIDELITY AS A PROXY FOR GENERATION QUALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

During inference in Transformer-based large language models, key–value (KV) cache compression is widely used to reduce memory overhead. Yet accurate local attention reconstruction need not imply reliable generation after compression. We test this assumption by separately measuring fitting-query residuals, operator errors at future full-model queries, and task scores from compressed generation. We use calibrated local merging (CALM) as a tractable test construction and compare it with eviction and controlled attention-fitting methods. A classical conditional Kullback–Leibler (KL) identity yields a local output bound; real-state measurements show that its global form is loose, while a retrospective group certificate reduces median prompt-level bound/error ratios from thousands to 6.40–7.95 at 40% retention. Across three near-4B model families, twelve LongBench tasks, 64 examples per task, and 8K/16K caps, we evaluate 27,648 predictions. On 24 matched Phi texts, CALM has lower operator error than ChunkKV throughout, yet lower task scores on ten. Allocation controls do not establish a geometry-specific task benefit, and compression before the first output reveals protocol sensitivity. Genuine-length inputs save 50.16–59.70% of all persistent cache state, but eager decoding is slower. These results delimit what local fidelity can support as a proxy without implying that fidelity and generation quality are generally unrelated.

1 INTRODUCTION

Local attention reconstruction is a convenient diagnostic for key–value cache compression: approximate and full attention can be compared at the same query without running an entire compressed generation. Whether this diagnostic supports conclusions about generated answers is less clear. A compact cache may match its fitting queries, behave differently at later queries, and alter the hidden states that produce subsequent tokens. Storage and latency introduce another distinction: fewer persistent entries need not accelerate a particular backend.

These distinctions matter as context grows. Autoregressive attention stores earlier keys and values; grouped-query attention (GQA) and efficient kernels reduce some costs but do not eliminate sequence-dependent cache storage (Vaswani et al., 2017; Ainslie et al., 2023; Dao et al., 2022). Hybrid decoders also retain recurrent state. An evaluation therefore needs to distinguish the representation being approximated, the queries on which it is tested, and the resources actually saved.

We examine local fidelity at three measurement levels: fitting-query reconstruction, operator error at future full-model queries, and task scores from compressed autoregressive generation. Calibrated local merging (CALM) serves as a tractable test construction alongside eviction and controlled attention-fitting comparators. Its fixed-anchor identity makes explicit what local reconstruction preserves. Mass calibration and the conditional-KL identity are established (Tian et al., 2026; Zweiger et al., 2026; Banerjee et al., 2005); here they expose approximation terms that can be measured on real states.

We contribute an empirical assessment of local bounds and their slack, matched operator/task comparisons on the same texts, and controlled tests of allocation and evaluation-protocol sensitivity. We also derive a group-adaptive retrospective certificate that refines the bound while retaining its dependence on raw groups and future queries. Our evaluation covers three near-4B families, twelve

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.

LongBench tasks, 64 examples per task, and 8K/16K caps, with cohort-wise estimates and a prompt-construction sensitivity analysis.

Across these measurements, CALM has lower future operator error than ChunkKV on all 24 matched Phi texts but lower task scores on ten. Global bounds remain quantitatively loose, and stronger allocation controls do not establish a task benefit from directional geometry. Physical storage savings coexist with protocol sensitivity and slower eager decoding. Together, these findings constrain the use of local fidelity as a proxy within the tested setting; they do not imply universal independence between reconstruction and task quality.

2 Related Work

Attention-based deletion retains entries selected by observed attention. Heavy-Hitter Oracle (H_2O) combines heavy hitters with recent tokens, Scissorhands examines persistent importance, and StreamingLLM preserves attention sinks (Zhang et al., 2023; Liu et al., 2023; Xiao et al., 2024). SnapKV uses an observation suffix to select prompt positions; ChunkKV retains nearby groups; RocketKV combines coarse eviction with finer sparse attention (Li et al., 2024; Liu et al., 2025; Behnam et al., 2025). Our eviction comparisons use equal physical entry budgets and a common decoding controller, rather than reproducing each system's optimized implementation.

Other methods alter precision, depth sharing, channels, or allocation. KIVI and KVQuant quantize the cache, MiniCache shares information across depth, ThinK prunes key channels, and Pyramid-Infer allocates capacity according to inference behavior (Liu et al., 2024c; Hooper et al., 2024; Liu et al., 2024a; Xu et al., 2025; Yang et al., 2024). GraphKV considers graph structure (Li et al., 2025). These directions are complementary; we do not claim measured gains from combining them with local merging.

KeepKV and Attention Matching are closer to the present representation (Tian et al., 2026; Zweiger et al., 2026). A port of KeepKV's released centroid/vote mechanism is included in a secondary 32-example comparison, with its adaptations disclosed. The main comparison implements the Attention Matching (AM) HighestAttnKeys-basic configuration: selected keys, bounded scalar mass fitting, and least-squares values. It uses the same final 64 prompt positions as CALM, corresponding to $64G$ query rows per physical KV head, but omits the published system's on-policy query collection. A fixed-ridge value fit tests numerical stability separately. KVSlimmer and SemantiCache study asymmetric merging and semantic grouping with proportional attention correction, respectively (Liu et al., 2026; Wu et al., 2026); they remain unimplemented comparators. None of these controlled comparisons ranks all features of the published systems.

The information-bottleneck perspective motivates retention of task-relevant distinctions (Tishby et al., 1999), but our address-entropy argument does not bound information in continuous cache vectors. Query visibility also matters (Luo et al., 2026): the benchmark question is included in the observed prompt, so this is query-aware post-prefill compression, not a reusable question-independent document cache. Evidence position can change long-context performance (Liu et al., 2024b). Prompt construction can also affect the evaluated protocol, as demonstrated by our prompt-control experiment, motivating input hashes, actual-length reporting, and the explicit prompt and first-output controls.

3 Audit Framework and Test Construction

We use one-anchor merging to make local fidelity analytically explicit. CALM supplies a fixed, training-free construction for the audit: its representations, allocation controls, and error terms can be varied or measured separately. The identities below concern this representation; whether they support claims about generated answers is an empirical question.

3.1 Cache representation and exact anchor matching

Consider one physical KV head with keys $k_i \in \mathbb{R}^d$ after rotary position embedding (RoPE) and values $v_i \in \mathbb{R}^d$, $i = 1, \ldots, N$. Write $x = sq$ for the scaled query, where s is the model's attention

scale. The attention output is

$$o(x) = \frac{\sum_i e^{x^\top k_i} v_i}{\sum_i e^{x^\top k_i}}. \tag{1}$$

A compressed entry stores $(\bar{k}_C, \bar{v}_C, b_C)$ for a group C, contributing the logit $x^\top \bar{k}_C + b_C$. At an anchor x_0, define

$$\begin{aligned} Z_C^0 &= \sum_{i \in C} e^{x_0^\top k_i}, & w_i^0 &= e^{x_0^\top k_i} / Z_C^0, \\ \bar{k}_C &= \sum_{i \in C} w_i^0 k_i, & \bar{v}_C &= \sum_{i \in C} w_i^0 v_i, \\ b_C &= \log Z_C^0 - x_0^\top \bar{k}_C. \end{aligned} \tag{2}$$

At x_0, the representative has mass Z_C^0 and numerator $Z_C^0 \bar{v}_C$. Replacing disjoint groups therefore preserves the entire attention output in real arithmetic. The key mean also matches the first derivative of the group's log-partition at the anchor. This is a local identity; it does not state that the representative is exact for every query.

The operation is closed under recursive merging at a fixed anchor. For two current entries a, b, let $\ell_a = x_0^\top k_a + b_a$, $\ell_b = x_0^\top k_b + b_b$, and $w = \mathrm{sigmoid}(\ell_a - \ell_b)$. The new key and value are $wk_a + (1-w)k_b$ and $wv_a + (1-w)v_b$, and its bias is

$$b_{ab} = \mathrm{logaddexp}(\ell_a, \ell_b) - x_0^\top \bar{k}_{ab}. \tag{3}$$

Existing biases must be included in the child logits. Raw tokens begin with bias zero. Unmerged entries remain valid alongside merged representatives and newly appended tokens.

3.2 Fixed allocation and observed queries

CALM pools queries at the last 64 real prompt positions across grouped query heads into an anchor and diagonal covariance D; this gives $64G$ queries per physical KV head shared by G query heads. Adjacent pairs are ranked by observed salience, key dispersion, and normalized value distortion (Eq. 15). The first four and last 64 entries are protected, with target $B = \min(N, \max(69, \lfloor fN \rfloor))$ and $f = 0.4$. Statistics use the completed prompt, including the question; no answer is consulted. The anchor identity holds at the pooled mean, not at every observed query. Appendix C gives the exact scoring, round schedule, and local surrogate.

3.3 A conditional-information interpretation

Let $\delta = x - x_0$ and $r_i = \delta^\top (k_i - \bar{k}_C)$. Since $\mathbb{E}_{w^0} r_i = 0$, the exact group partition can be written as

$$Z_C(x) = e^{x^\top \bar{k}_C + b_C}\, e^{\psi_C(x)}, \qquad \psi_C(x) = \log \mathbb{E}_{w^0} e^{r_i}. \tag{4}$$

The conditional attention distribution inside C is $w_i(x) = w_i^0 e^{r_i - \psi_C(x)}$. Consequently,

$$\boxed{\mathrm{KL}\big(w^0 \| w(x)\big) = \psi_C(x).} \tag{5}$$

This is the classical exponential-family Bregman identity (Banerjee et al., 2005): for $A_C(x) = \log Z_C(x)$, $\psi_C = A_C(x) - A_C(x_0) - \nabla A_C(x_0)^\top (x - x_0)$. Its use here identifies the error from freezing a KV group's conditional distribution; the identity itself is not new.

3.4 A local output bound and its scope

Proposition 1. *For every group and member, assume $|r_i| \le \epsilon$, $\|\bar{v}_C\|_2 \le V$, and $\|v_i - \bar{v}_C\|_2 \le \rho$. Let $\widetilde{o}(x)$ use the representatives in Eq. 2. In real arithmetic,*

$$\|o(x) - \widetilde{o}(x)\|_2 \le V\epsilon^2/4 + \rho\epsilon. \tag{6}$$

Hoeffding's lemma gives $0 \le \psi_C \le \eta = \epsilon^2/2$ (Hoeffding, 1963). The exact group distribution is the compressed group distribution reweighted by e^{ψ_C}. A second application of the lemma, followed

by Pinsker's inequality, bounds group reweighting by $V\eta/2$. Conditional value drift contributes at most $\rho\sqrt{2\eta}$. The full proof is in Appendix A.

The bound concerns one attention operator at a specified query. It does not propagate through layer projections, recurrent updates, or autoregressive hidden states, and it does not bound a benchmark score. It can also be loose when the within-group query radius is large. A two-key example in the appendix shows why value error can be first order even when partition error is second order. Bfloat16 (BF16) storage and bias casting introduce additional rounding error beyond the identity.

4 Controlled Evaluation

4.1 Physical compaction and model families

Actual cache tensors are shortened after full prefill. Each stored scalar bias is added to its representative's attention logit, with zero bias for new entries. Position identifiers (IDs) continue from the logical prompt length. All compressed methods use the same scaled dot-product attention (SDPA) controller, including allocated zero biases for eviction, and retain no hidden full KV cache. Merge arithmetic uses 32-bit floating point (FP32), stored keys/values use BF16, and FP32 biases are cast to the query dtype. Full-prefill computation and transient memory are not reduced.

The first output token in the main protocol comes from common full-prefill logits; subsequent tokens use the compressed cache. A separate experiment moves compression before the final prompt-token forward, allowing the first output to change. Main runs allow ordinary cache growth. A dynamic experiment instead recompacts every 128 decoding steps to the original prompt budget, using the current grouped-head queries as the new anchor. Its covariance describes cross-head dispersion, not a stored temporal history; it is zero when $G = 1$. Re-anchoring preserves the current surrogate, not the original raw cache.

The checkpoints are Qwen3.5-4B, Phi-4-mini-instruct, and Ministral-3-3B-Instruct-2512-BF16 (Qwen Team, 2026; Microsoft, 2025; Mistral AI, 2025). Loaded parameter counts are 4.539, 3.836, and 3.849 billion, including vision parameters where present; only text inputs are used. Qwen has eight full-attention and 24 recurrent layers, whereas Phi and Ministral have 32 and 26 dense attention layers. This selection spans distinct model families and architectures, but is not intended to represent every recent reasoning variant.

4.2 Sampling, prompt construction, and comparisons

LongBench v1 supplies official prompts and metrics for NarrativeQA, Qasper, MultiFieldQA, HotpotQA, 2WikiMQA, MuSiQue, GovReport, QMSum, MultiNews, Text REtrieval Conference (TREC) classification, TriviaQA, and SAMSum (Bai et al., 2024). Answer F1, the longest-common-subsequence variant of Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L), or classification accuracy is scaled to 0–100. Within each task, seed 2026 shuffles row indices: eight examples are reserved for development and the following 64 form the evaluation set. This gives 768 distinct source examples per model, evaluated at both caps. For a cohort-wise robustness check, we also report the first and second 32-example halves separately; their membership was fixed by the shuffled indices before evaluation. These estimates are descriptive rather than preregistered confirmatory tests.

Long inputs retain their first and last token halves; short inputs are not padded. Greedy generation uses official task limits and identical input IDs across methods. For Qwen on SAMSum, we place the benchmark's unchanged few-shot content in the model's native chat template and disable thinking. Without this wrapper, generation stops at a two-token thinking marker on 30/32 and 31/32 full-cache outputs at 8K and 16K. Appendix K reports this prompt sensitivity and an eleven-task analysis that excludes SAMSum. All main comparisons, including component controls, use the chat-wrapped prompt.

SnapKV uses a 64-position observation suffix and width-five average pooling. ChunkKV uses 20-token chunks with a protected suffix and token remainder to meet the exact budget. AM-basic selects keys by root-mean-square (RMS) observed attention, fits bounded mass weights with the released least-squares initialization and two projected steps, then fits values by FP32 least squares. It pools

Table 1: LongBench macro scores on 64 examples per task and twelve tasks. Every compressed method retains 40% of prompt entries, subject to the common minimum. AM variants use queries from the controlled final 64 prompt positions ($64G$ rows per physical KV head). Caps are limits, not padding targets. Higher is better.

Model	Cap	Full	SnapKV	ChunkKV	AM-basic	AM-ridge	CALM
Qwen3.5-4B	8K	45.65	45.55	45.48	40.19	40.70	45.81
Qwen3.5-4B	16K	46.87	46.79	46.95	41.48	41.71	47.22
Phi-4-mini	8K	43.93	43.10	43.59	29.45	37.75	42.79
Phi-4-mini	16K	45.37	44.81	44.82	30.34	38.92	44.22
Ministral-3-3B	8K	40.72	40.27	40.44	32.64	32.97	38.98
Ministral-3-3B	16K	42.89	42.52	42.52	33.52	33.82	41.36

the $64G$ real queries per physical KV head without a causal fitting mask. It has no extra protected entries or on-policy layer rerun. These are material adaptations of the released basic configuration, detailed in Appendix E.

Because AM-basic produces finite but very large values in some diagnostic fits, AM-ridge changes the value-fitting stage to the objective $\|XC - Y\|_F^2 + 10^{-6}\|C\|_F^2$, solved by 64-bit floating-point (FP64) dual Cholesky before BF16 storage. The coefficient is the upstream fallback value, applied to every head without task-score tuning. This sensitivity changes regularization, precision, and solver together. We report both variants. Six methods at two caps on all 64 examples yield 27,648 predictions, covering every case in the completed grid.

4.3 Estimands and measurement units

Macro scores weight the twelve task means equally. Unless stated otherwise, paired task-stratified percentile bootstrap intervals use 10,000 resamples (Dror et al., 2018); both caps of a source text are resampled together when pooled. Intervals are descriptive, without multiplicity correction or a non-inferiority margin. A zero-containing interval does not establish equivalence. Separate diagnostics distinguish fitted queries, future full-model queries, and compressed autoregressive generation.

Persistent memory counts actual tensor allocations, scalar metadata, and recurrent state. Peak PyTorch allocation is separate. Matched-token timing uses one reserved GovReport example per model that reaches 16K, both caps, and the same 63 teacher-forced decode forwards. After per-method warmup, three measured repetitions rotate method order on a graphics processing unit (GPU) without other active project workers. Co-tenant benchmark timings are operational logs, not speed estimates. The environment is eight NVIDIA B200 GPUs, CUDA 13.2, NVIDIA PyTorch 2.12.0a0, and Transformers 5.16.1.

5 Task-Level Evidence and Allocation Controls

Table 1 shows model-dependent effects. CALM-minus-full differences are $+0.16$ $[-0.43, 0.82]$ and $+0.35$ $[-0.20, 0.96]$ for Qwen at 8K/16K; -1.14 $[-1.69, -0.62]$ and -1.15 $[-1.79, -0.55]$ for Phi; and -1.74 $[-2.61, -0.92]$ and -1.52 $[-2.37, -0.73]$ for Ministral. CALM is below ChunkKV in both Phi and both Ministral conditions, with negative paired intervals. For Qwen, the CALM-minus-SnapKV and CALM-minus-ChunkKV intervals include zero at both caps. Conversely, CALM exceeds both adapted AM variants in all six conditions, with positive paired intervals. Ridge improves AM's Phi scores substantially, but does not recover full-cache performance. These comparisons favor neither a general merging advantage nor uniformly negligible loss; complete paired intervals are in Appendix H.

Appendix H reports the complete 64-example intervals, separate estimates for the second 32-example cohort, and per-task CALM-minus-full differences. The second cohort preserves the qualitative pattern: Qwen's CALM-minus-full intervals include zero, whereas Phi and Ministral have negative intervals at both caps. Ministral's 8K change is -2.50 $[-3.94, -1.22]$, larger than its full-set decline. The wider interval and changed point estimate illustrate the instability of conclusions

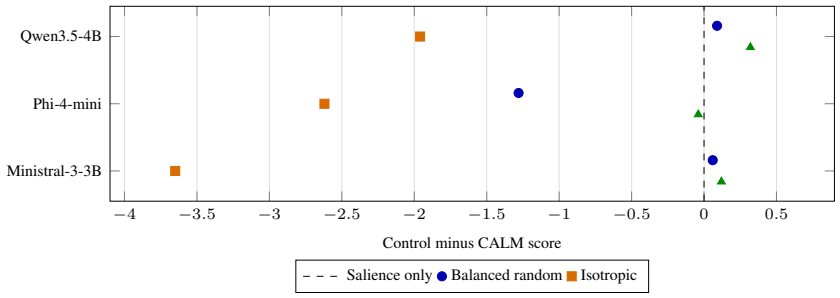

Figure 1: Allocation-control minus CALM score differences, averaged over both caps on 16 examples per task. Positive values favor the control. Candidate construction, protection, update formula, and removal counts are shared. Selection changes the realized groups and tree depths. Complete paired intervals are in Appendix I.

drawn from either half alone. The Qwen SAMSum chat protocol applies to all six methods and both cohorts. A separate five-method analysis of the unwrapped prompt is reported only as a prompt sensitivity and is not pooled with the main estimates.

At the 16K cap, 167/768 Qwen, 81/768 Phi, and 178/768 Ministral inputs actually reach 16,384 tokens; medians are 9,179, 8,678, and 9,408. At 8K, the corresponding at-cap counts are 419, 397, and 432. Thus the full score grid is a cap-based study, not a set of uniformly long inputs. Model-specific tokenization changes membership in the actual-at-cap subsets. Those subsets contain fewer tasks and estimate a different macro average; they cannot replace the twelve-task result. Complete length distributions and paired actual-at-cap estimates are in Appendix G.

Salience-only ranks pairs by $p_a + p_b$. Balanced random selects one candidate from each approximately equal contiguous bin of current pair indices; it balances current representative positions, not raw-token span. Isotropic geometry replaces D by $(\operatorname{tr} D/d)I$, retaining the other terms in Eq. 15. All three preserve adjacent candidate construction, protected entries, per-round merge counts, and weighted updates. The isotropic control isolates directional anisotropy; salience-only removes the entire within-pair distortion weighting. No control can keep the final merge tree fixed while changing which pairs merge.

Balanced random loses 1.96, 2.62, and 3.65 points relative to CALM on Qwen, Phi, and Ministral; all three paired intervals are negative. This supports nonrandom allocation beyond the left-to-right control. However, salience-only changes scores by $+0.09$, -1.28, and $+0.06$, while isotropic geometry changes them by $+0.32$, -0.04, and $+0.12$; every corresponding interval includes zero. Thus neither the full distortion ranking nor directional query geometry has an established task advantage over these stronger controls. The Phi salience-only upper endpoint rounds to $+0.00$ but is positive before rounding. Additional no-mass, fixed-order, no-value, and zero-anchor controls use the same Qwen SAMSum chat prompt. Their complete intervals and precise interventions appear in Appendix I. These comparisons test components of one local construction; they do not establish a global clustering optimum.

6 Auditing Local Fidelity

6.1 Worst-case bounds on real states

We audit the local bound on six reserved-development texts, each evaluated under both context caps and all three models. The first, middle, and last eligible attention layers are sampled, with retained fractions 0.2, 0.4, and 0.5 and at most eight full-model future queries before the end-of-sequence (EOS) token. All 324 layer/fraction trace checks reproduce CALM's deployed tensors exactly; 39,960 correlated head/query conditions satisfy the ideal-arithmetic bound within the stated FP64 tolerance. The measured radii are nevertheless too large for a tight global guarantee. After averaging bounds and errors separately within each prompt/cap unit, the median global-bound/error ratios at fraction 0.4 are 3,695, 23,643, 5,184 for Qwen, Phi, and Ministral, respectively (Table 3).

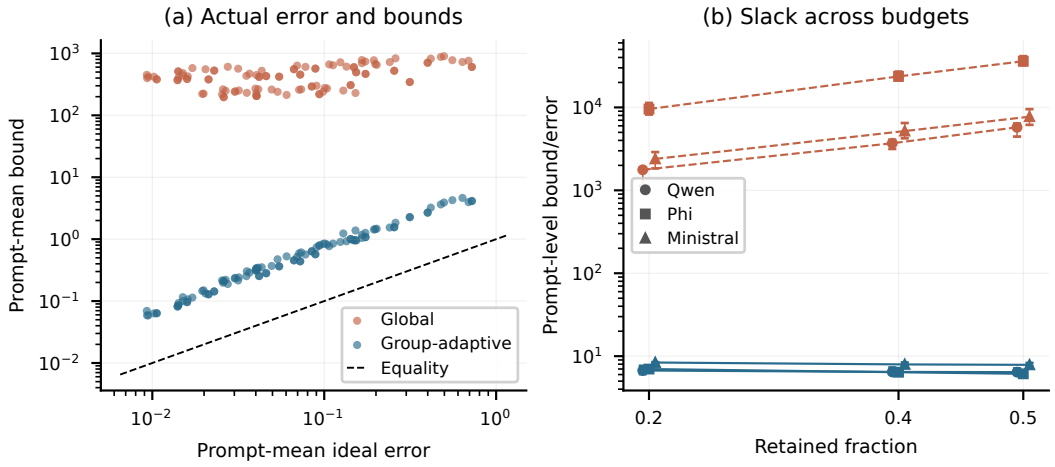

Figure 2: Retrospective bound diagnostics. (a) Each point is a prompt/cap mean at one model and retained fraction, pooling the three fractions. (b) Median prompt-level bound/error ratios with interquartile ranges over twelve units; model markers are offset slightly for readability. Orange dashed curves denote the worst-case global bound and blue solid curves the group-adaptive certificate. Repeated caps and model views share source texts.

Retaining group-specific radii and the exact group-mass drift in the same proof gives a retrospective certificate. With exact and representative group probabilities $P_C, \widetilde{P}_C$ and value radius $\rho_C = \max_{i \in C} \|v_i - \bar{v}_C\|_2$,

$$\|o - \widetilde{o}\|_2 \le \left\| \sum_C (P_C - \widetilde{P}_C) \bar{v}_C \right\|_2 + \sum_C P_C \rho_C \min\{2, \sqrt{2\psi_C}\} = \mathcal{B}_{\mathrm{KL}}(x). \tag{7}$$

Its corresponding median ratios are 6.46, 6.40, 7.95. This refinement needs the original groups and future query, so it supplies neither a free online guarantee nor a new general KL identity. As retention increases, measured error falls more rapidly than the global radius bound (Figure 2). The refined certificate tracks these changes more closely, but remains several times the measured error. These descriptive associations concern six texts with repeated caps, not independent head samples or validated task predictions. BF16 deployment error is reported separately because numerical rounding is outside the theorem. Appendix B gives the proof, distributions, and precision checks.

6.2 Fitted queries, future operators, and generated answers

A designated diagnostic uses two examples from every task: 24 source texts per model at both caps. For each prompt, the first, middle, and last eligible layers are evaluated on at most eight actual future full-model queries, stopping at EOS or the task's stopping condition. Each method receives the same queries and fresh, uncompressed generated KV entries. Relative errors are averaged within a prompt before comparisons; heads, layers, steps, and the two cap conditions are not independent examples. This tests out-of-fit operator behavior, not accumulated compressed-generation query drift.

Figure 3 reports both means and medians. CALM's mean relative errors are 0.0452, 0.0208, and 0.0610 for Qwen, Phi, and Ministral, compared with SnapKV's 0.0533, 0.0606, and 0.0810 and ChunkKV's 0.0602, 0.0619, and 0.0900. AM-ridge gives 0.2189, 0.0919, and 0.1386 despite small stored-reference residuals. AM-basic has a Phi mean of 148.493 but median 0.201: finite ill-conditioned fits create an extreme tail. Ridge removes that tail without removing the gap between fitted and future queries. Appendix I reports fitting residuals and future errors as separate estimands.

The 720 matched operator/task records comprise 24 source texts, three models, five compressed methods, and two caps; they are paired records, not 720 independent texts. They expose disagreements on the same texts. For Phi, CALM has lower operator error than ChunkKV on all 24 sources, yet its task score is lower on ten, tied on twelve, and higher on two after averaging caps. CALM's descriptive Spearman correlations between operator error and task loss are 0.07, -0.09, and -0.11

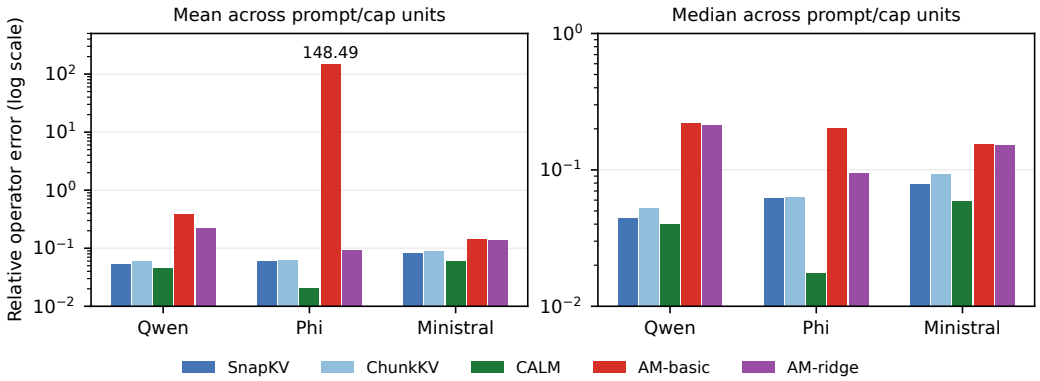

Figure 3: Future-query relative operator error on 24 designated diagnostic texts per model, each at two caps. Means and medians use prompt/cap units after within-prompt averaging. Both panels retain every method, including the AM-basic Phi tail; logarithmic scales differ.

across the three models. These small-sample coefficients do not establish an absence of association. Appendix I gives every comparator's rankings, scatter points, and paired uncertainty. A small fitted-query residual therefore answers a narrower question than a small future-query error, and neither alone certifies a generated answer. This is consistent with the local theorem: it holds at a specified query and contains no bound on later hidden states, discrete token decisions, or task metrics. Solver stability is another distinct issue; the AM-basic tail is reported alongside the fixed-ridge sensitivity rather than attributed to attention fitting as a general principle.

7 First-Output and Systems Sensitivity

7.1 Shared first output versus compression before it

The main protocol fixes the first generated ID from full-prefill logits. We stratify its predictions by full-cache output length, excluding the union of tokenizer-special, decoder-special, and EOS IDs; raw-ID lengths including EOS are also retained. Membership is fixed by full cache for every method. A one-content-token answer is not guaranteed unchanged: the next compressed forward can continue instead of predicting EOS. A one-raw-ID run stopped immediately would be invariant under the shared stopping policy. Buckets also differ in task composition, so their differences are not causal length effects.

All 148 one-content-token condition pairs have unchanged CALM scores, but Qwen 8K has only 33/34 identical ID sequences. There are no one-raw-ID full-cache runs. For Phi outputs of at least 17 content tokens, CALM-minus-full differences are -3.08 points $[-4.55, -1.71]$ at 8K and -2.69 $[-3.98, -1.41]$ at 16K. Ministral instead has larger observed losses in some shorter strata, and Qwen's long-output intervals include zero. Thus the protocol and task mixture matter; these strata do not establish a universal length-dependent degradation law.

The first-output experiment prefills $N - 1$ prompt tokens, compresses that prefix, and consumes the final prompt token at its original position before selecting the first output. Its full-cache baseline uses the same split prefill. Four examples from every task, both caps, and all three models give 48 distinct source examples per model. The missing final token changes the observed query set, and the prefix budget plus that new entry can exceed the main compressed count by one. This is an operational sensitivity, not an otherwise identical intervention on a single logit. Full split-versus-monolithic changes and paired differences-in-differences are reported in Appendix J.

The both-cap CALM differences-in-differences are $+0.55$ $[-0.13, 1.36]$, -0.18 $[-1.25, 0.78]$, and -0.60 $[-1.33, -0.03]$ for Qwen, Phi, and Ministral. The Ministral loss increases even though no first output ID changes in its 96 conditions. Full-cache split and monolithic sequences themselves disagree in 25.0–41.7% of cases across the six model–cap rows. These results reveal model-dependent protocol sensitivity; they neither isolate the first token nor establish protocol invariance.

Table 2: CALM storage and warmed latency on genuine-length GovReport inputs. Persistent bytes include scalar metadata and recurrent state. Speed below one is slower than full cache. All six methods' timing and peak allocation appear in Appendix M.

Model	Cap	Full MiB	CALM MiB	Saved %	Merge ms	Decode ×
Qwen3.5-4B	8K	305.5	152.3	50.16	22.2	0.99
Qwen3.5-4B	16K	561.5	255.1	54.57	32.7	0.98
Phi-4-mini	8K	1024.0	412.7	59.70	95.8	0.78
Phi-4-mini	16K	2048.0	825.5	59.69	118.9	0.83
Ministral-3-3B	8K	832.0	335.3	59.70	77.8	0.85
Ministral-3-3B	16K	1664.0	670.7	59.69	115.7	0.85

7.2 Persistent memory, timing, and refresh

An entry stores $4d$ BF16 KV bytes plus four FP32 bias bytes, giving an approximate compressed-to-full ratio of $f(1 + 1/d)$ for KV tensors plus bias metadata. At $f = 0.4$, this corresponds to 59.7–59.9% savings in those allocations. Qwen's unchanged 49.5 mebibyte (MiB) recurrent state reduces the savings across all persistent cache state to 50.16% at 8K and 54.57% at 16K; dense-model savings across all persistent state are about 59.7%. Shorter Qwen prompts can save less than half of all persistent state. CALM decodes at 0.78–0.99 times full-cache speed, with 22–119 ms of initial compaction and total latency 1.02–1.40 times full cache. Thus the storage reduction again provides no decoding speedup in this backend. These eager, batch-one implementations include bias handling, cache updates, and numerical diagnostics; they do not establish optimized serving performance.

Recompression every 128 steps keeps the initial representative budget plus bounded growth between events, but loses additional information. On four examples from each of three summary tasks, mean refresh-minus-once differences range from -0.19 to -0.87 points; the Ministral 8K interval is $[-1.71, -0.13]$. Total differences from full cache range from -1.18 to -3.36. Most other refresh intervals include zero on this small subset. Original tokens cannot be recovered by changing the anchor. Appendix N reports every row, refresh cost, and matching-prefix checks.

8 Limitations and Conclusion

The evidence concerns three near-4B families, twelve English tasks, fixed source subsets, and two caps. It does not cover multilingual or code tasks, unseen-question document reuse, or long interactive conversations. The modern comparator is an explicitly adapted basic configuration, not the full on-policy Attention Matching system; other recent compaction systems remain untested. The ridge variant is a numerical-stability sensitivity motivated by the operator diagnostic, not a preregistered comparison. Cohort and prompt sensitivities are reported separately from the main estimates.

The theory is local, assumes real arithmetic, and does not propagate through generation. The tighter certificate needs raw groups and future queries. Diagonal query statistics, local pairing, salience estimates, and BF16 storage remain approximations. Operator diagnostics use short full-model trajectories, and output-length strata change task composition. Benchmark contamination and wider workload variability are not resolved by source hashes or paired confidence intervals.

This audit separates local reconstruction from the claims it is asked to support. Exact anchor matching admits a conditional-error decomposition, but the resulting global bound is loose on measured states. A retrospective refinement reduces that slack without becoming an online certificate. Matched texts exhibit lower operator error alongside lower task scores, and allocation controls leave geometry-specific benefits unresolved. First-output and systems measurements reveal further protocol and backend dependence. Local fidelity is consequently useful diagnostic evidence whose relevance to generation must be tested, with task outcomes and resource costs measured separately. The present comparisons establish those limits for the tested constructions and queries, rather than a universal ranking of cache-compression methods.

AI USE STATEMENT

We used generative AI tools for research planning, literature-search assistance, code development, and manuscript review. We verified reported benchmark measurements against saved outputs from public data and pretrained checkpoints; the numerical results were not generated as illustrative values. We also checked source and input hashes, scoring, cache accounting, mathematical identities, document compilation, and publication metadata. We reviewed all AI-assisted work and take responsibility for the final text, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

Appendix D documents the implementation and baseline adaptations, Appendix F specifies models, data, sampling, software, and analysis seeds, and Appendix A gives complete proofs. The supplementary materials include pinned dependencies, source and input hashes, run manifests, saved predictions, raw timing records, and scripts that regenerate the reported tables and figures.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005. URL http://jmlr.org/papers/v6/banerjee05b.html.

Payman Behnam, Yaosheng Fu, Ritchie Zhao, Po-An Tsai, Zhiding Yu, and Alexey Tumanov. RocketKV: Accelerating Long-Context LLM Inference via Two-Stage KV Cache Compression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 3358–3392, 2025. URL https://proceedings.mlr.press/v267/behnam25a.html.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. doi: 10.52202/068431-1189. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/67d57c32e20fd0a7a302cb81d36e40d5-Paper-Conference.pdf.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084, 2019. doi: 10.18653/v1/P19-1102. URL https://aclanthology.org/P19-1102/.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, 2019. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409/.

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

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

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1270–1303, 2024. doi: 10.52202/079017-0040. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/028fcbcf85435d39a40c4d61b42c99a4-Paper-Conference.pdf.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient Attentions for Long Document Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1419–1436, 2021. doi: 10.18653/v1/2021.naacl-main.112. URL https://aclanthology.org/2021.naacl-main.112/.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

Xuelin Li, Xiangqi Jin, and Linfeng Zhang. GraphKV: Breaking the Static Selection Paradigm with Graph-Based KV Cache Eviction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21899–21909, 2025. doi: 10.18653/v1/2025.emnlp-main.1112. URL https://aclanthology.org/2025.emnlp-main.1112/.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM Knows What You are Looking for Before Generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22947–22970, 2024. doi: 10.52202/079017-0722. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/28ab418242603e0f7323e54185d19bde-Paper-Conference.pdf.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004. URL https://aclanthology.org/W04-1013/.

Akide Liu, Jing Liu, Zizheng Pan, Yefei He, Gholamreza Haffari, and Bohan Zhuang. MiniCache: KV Cache Compression in Depth Dimension for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139997–140031, 2024a. doi: 10.52202/079017-4443. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fd0705710bf01b88a60a3d479ea341d9-Paper-Conference.pdf.

Lianjun Liu, Hongli An, Weiqi Yan, Xin Du, Shengchuan Zhang, Huazhong Liu, and Yunshan Zhong. KVSlimmer: Theoretical Insights and Practical Optimizations for Asymmetric KV Merging. *arXiv preprint arXiv:2603.00907*, 2026. URL https://arxiv.org/abs/2603.00907.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, and Xiaowen Chu. ChunkKV: Semantic-Preserving KV Cache Compression for Efficient Long-Context LLM Inference. In *Advances in Neural Information Processing Systems*, volume 38, pp. 28728–28778, 2025. doi: 10.52202/085713-0966. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2987f911151b39cd3a1761e212319e8e-Paper-Conference.pdf.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the Persistence of Importance Hypothesis for LLM KV Cache Compression at Test Time. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52342–52364, 2023. doi: 10.52202/075280-2279. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a452a7c6c463e4ae8fbdc614c6e983e6-Paper-Conference.pdf.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32332–32344, 2024c. URL https://proceedings.mlr.press/v235/liu24bz.html.

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

Microsoft. Phi-4-mini-instruct Model Card, 2025. URL https://huggingface.co/microsoft/Phi-4-mini-instruct.

Mistral AI. Ministral-3-3B-Instruct-2512-BF16 Model Card, 2025. URL https://huggingface.co/mistralai/Ministral-3-3B-Instruct-2512-BF16.

Qwen Team. Qwen3.5-4B Model Card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-4B.

Yuxuan Tian, Zihan Wang, Yebo Peng, Aomufei Yuan, Zhiming Wang, Bairen Yi, Xin Liu, Yong Cui, and Tong Yang. KeepKV: Achieving Periodic Lossless KV Cache Compression for Efficient LLM Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33259–33267, 2026. doi: 10.1609/aaai.v40i39.40611. URL https://ojs.aaai.org/index.php/AAAI/article/view/40611.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The Information Bottleneck Method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. ♪ MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Shunlong Wu, Hai Lin, Shaoshen Chen, Tingwei Lu, Yongqin Zeng, Shaoxiong Zhan, Hai-Tao Zheng, and Hong-Gee Kim. SemantiCache: Efficient KV Cache Compression via Semantic Chunking and Clustered Merging. *arXiv preprint arXiv:2603.14303*, 2026. URL https://arxiv.org/abs/2603.14303.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5e5fd18f863cbe6d8ae392a93fd271c9-Paper-Conference.pdf.

Yuhui Xu, Zhanming Jie, Hanze Dong, Lei Wang, Xudong Lu, Aojun Zhou, Amrita Saha, Caiming Xiong, and Doyen Sahoo. ThinK: Thinner Key Cache by Query-Driven Pruning. In *International Conference on Learning Representations*, volume 2025, pp. 56691–56709, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8edb116d5b288b6a9bba4c16ab647702-Paper-Conference.pdf.

Dongjie Yang, Xiaodong Han, Yan Gao, Yao Hu, Shilin Zhang, and Hai Zhao. PyramidInfer: Pyramid KV Cache Compression for High-throughput LLM Inference. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3258–3270, 2024. doi: 10.18653/v1/2024.findings-acl.195. URL https://aclanthology.org/2024.findings-acl.195/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang "Atlas" Wang, and Beidi Chen. H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710, 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6ceefa7b15572587b78ecfcebb2827f8-Paper-Conference.pdf.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, 2021. doi: 10.18653/v1/2021.naacl-main.472. URL https://aclanthology.org/2021.naacl-main.472/.

Adam Zweiger, Xinghong Fu, Han Guo, and Yoon Kim. Fast KV Compaction via Attention Matching. *arXiv preprint arXiv:2602.16284*, 2026. URL https://arxiv.org/abs/2602.16284.

A Proofs and Interpretation

A.1 Anchor identity and recursive closure

For a group C, Eq. 2 gives $\exp(x_0^\top \bar{k}_C + b_C) = Z_C^0$ and $\exp(x_0^\top \bar{k}_C + b_C)\bar{v}_C = \sum_{i\in C} \exp(x_0^\top k_i) v_i$. Summing these identities across disjoint groups reproduces the denominator and numerator of Eq. 1. A singleton has $b_C = 0$ and is unchanged. For recursive merging, child masses are e^{ℓ_a} and e^{ℓ_b}, so $w = e^{\ell_a}/(e^{\ell_a} + e^{\ell_b})$. Equation 3 preserves their summed mass and weighted value numerator. Induction establishes closure for any sequence of disjoint merges at a fixed anchor. The candidate tree affects the final partition. For a fixed final group and anchor, its representative is independent of merge parenthesization in real arithmetic.

A.2 Conditional KL and the group distribution

For $\delta = x - x_0$, expand $x^\top k_i = x_0^\top k_i + \delta^\top \bar{k}_C + r_i$. Factoring out the first two group-level terms gives Eq. 4. Since $\sum_i w_i^0 r_i = 0$,

$$\mathrm{KL}(w^0 \| w(x)) = \sum_{i\in C} w_i^0 \log \frac{w_i^0}{w_i^0 e^{r_i - \psi_C}} = \psi_C - \sum_i w_i^0 r_i = \psi_C. \tag{8}$$

This is a reverse conditional divergence. It compares the frozen anchor distribution with the actual within-group distribution; it is not a KL divergence between complete model output distributions.

Let $\widetilde{P}_C$ be the compressed attention probability of group C. The exact probability is

$$P_C = \frac{\widetilde{P}_C e^{\psi_C}}{\sum_D \widetilde{P}_D e^{\psi_D}}. \tag{9}$$

The true conditional mean value is $\mu_C(x) = \sum_i w_i(x) v_i$, so $o(x) = \sum_C P_C \mu_C(x)$ and $\widetilde{o}(x) = \sum_C \widetilde{P}_C \bar{v}_C$. This decomposition separates changed group mass from changed values inside a group.

A.3 Proof of the output bound

For a centered random variable supported on an interval of width 2ϵ, Hoeffding's lemma gives $\log \mathbb{E} e^r \le (2\epsilon)^2/8 = \epsilon^2/2$ (Hoeffding, 1963). Jensen's inequality supplies the lower bound zero. Hence $\psi_C \in [0, \eta]$ with $\eta = \epsilon^2/2$. Using Eq. 9,

$$\mathrm{KL}(\widetilde{P} \| P) = \log \mathbb{E}_{\widetilde{P}} e^{\psi_C} - \mathbb{E}_{\widetilde{P}} \psi_C \le \eta^2/8. \tag{10}$$

Pinsker's inequality in the convention $\|p - q\|_1 \le \sqrt{2\,\mathrm{KL}(p\|q)}$ therefore gives $\|P - \widetilde{P}\|_1 \le \eta/2$. For each conditional distribution, $\|w(x) - w^0\|_1 \le \sqrt{2\psi_C} \le \sqrt{2\eta}$. Because the probability differences sum to zero,

$$\|\mu_C(x) - \bar{v}_C\|_2 = \left\| \sum_i (w_i(x) - w_i^0)(v_i - \bar{v}_C) \right\|_2 \le \rho\sqrt{2\eta}. \tag{11}$$

Finally,

$$\begin{aligned} \|o - \widetilde{o}\|_2 &\le \left\| \sum_C P_C(\mu_C - \bar{v}_C) \right\|_2 + \left\| \sum_C (P_C - \widetilde{P}_C)\bar{v}_C \right\|_2 \\ &\le \rho\sqrt{2\eta} + V\eta/2 = \rho\epsilon + V\epsilon^2/4. \end{aligned} \tag{12}$$

No independence across keys or groups is assumed. The bounded-radius assumptions apply to the particular query under consideration. A large ϵ can make the bound uninformative without making it invalid.

A.4 Linearized pair criterion and a counterexample

For a pair with anchor weight w, its conditional value at a displaced query is $\mu_v(x_0+\delta) = v_b + \mathrm{sigmoid}(\mathrm{logit}\, w + \delta^\top \Delta k)\Delta v$. The linearized change is $w(1-w)(\delta^\top \Delta k)\Delta v$. Under the local query model in Appendix C, its normalized RMS is exactly $w(1-w)\sqrt{ag_v}$. Equation 16 follows from $2\sqrt{ag_v} \le a + g_v$. The expected quadratic log-partition term is $\frac{1}{2}w(1-w)a$. This explains the implemented sum as a local surrogate. It supplies no approximation ratio for greedy clustering, and it does not replace the radius assumptions of Proposition 1.

A scalar analytic example illustrates the limitation. Take $k_1 = 1$, $k_2 = -1$, $v_1 = 1$, $v_2 = -1$, and $x_0 = 0$. The representative has key and value zero and bias $\log 2$. At $x = t$, the exact partition is $2\cosh t$, the missing log-partition is $\log\cosh t$, and the exact output is $\tanh t$. The compressed output is zero. Partition error is second order near zero, while value error is first order. This is a mathematical example, not a synthetic benchmark or a reported empirical data point.

A.5 Recompression and finite precision

When the anchor changes, Eq. 3 is applied to current representative keys, values, and stored biases. The identity then concerns that current surrogate cache. It does not concern a hypothetical reconstruction of the original raw token groups at the new anchor. Earlier losses remain. In the code, merges use FP32 arithmetic, keys and values are cast back to BF16, and attention biases are cast to the query dtype. Rounding is outside the real-arithmetic proof. Unit tests cover irregular cache lengths, pooled GQA anchors, preserved prefix/suffix entries, and a new-anchor recompression of an existing surrogate. The largest observed anchor discrepancy in these numerical identity tests is 4.01×10^{-8}; their test tensors are excluded from all benchmark analyses.

B Real-State Bounds and a Retrospective Group Certificate

B.1 A refinement of the output-bound argument

The conditional-KL identity is the classical exponential-family identity. The following result retains group-specific radii and the exact group-mass term in the same output-error decomposition. It is a refinement for auditing the slack in a worst-case bound, rather than a new general divergence theorem.

Proposition 2 (Retrospective group certificate)**.** *Fix a query x, anchor x_0, and partition of the raw cache, and construct the anchor representatives in real arithmetic. Let P_C and $\widetilde{P}_C$ be the exact and representative group probabilities. Define $\rho_C = \max_{i\in C}\|v_i - \bar{v}_C\|_2$, $\epsilon_C = \max_{i\in C}|(x - x_0)^\top(k_i - \bar{k}_C)|$, $\psi_C = \mathrm{KL}(w_C(x_0)\|w_C(x))$, and*

$$\mathcal{M}_{\mathrm{mass}}(x) = \left\| \sum_C (P_C - \widetilde{P}_C)\bar{v}_C \right\|_2 .$$

With $V = \max_C \|\bar{v}_C\|_2$, $\rho = \max_C \rho_C$, and $\epsilon = \max_C \epsilon_C$, we have

$$\begin{aligned} \|o - \widetilde{o}\|_2 &\le \mathcal{M}_{\mathrm{mass}}(x) + \sum_C P_C \rho_C \min\{2, \sqrt{2\psi_C}\} =: \mathcal{B}_{\mathrm{KL}}(x) \\ &\le \mathcal{M}_{\mathrm{mass}}(x) + \sum_C P_C \rho_C \min\{2, \epsilon_C\} =: \mathcal{B}_{\mathrm{rad}}(x) \\ &\le V\epsilon^2/4 + \rho\epsilon =: \mathcal{B}_{\mathrm{glob}}(x). \end{aligned} \tag{13}$$

Proof. Set $\mu_C(x) = \sum_{i\in C} w_i(x) v_i$. Then

$$o - \widetilde{o} = \sum_C P_C(\mu_C - \bar{v}_C) + \sum_C (P_C - \widetilde{P}_C)\bar{v}_C.$$

The conditional probability differences sum to zero, so $\|\mu_C - \bar{v}_C\|_2 \le \rho_C \|w_C(x) - w_C(x_0)\|_1$. Pinsker's inequality bounds this ℓ_1 distance by $\sqrt{2\psi_C}$, and two is also an upper bound for any pair of probability distributions. The triangle inequality proves the first line. Hoeffding's lemma gives

Table 3: Bound diagnostics at retained fraction 0.4. Radii are medians of prompt/cap means. Ratios are medians [25th, 75th percentiles] of $\bar{B}_u/\bar{E}_u$ over twelve units per model; "Adaptive" uses $\mathcal{B}_{\mathrm{KL}}$. These are not ratios of displayed medians.

Model	ϵ	ρ	V	Global/error	Adaptive/error
Qwen	6.57	27.99	39.59	3,695 [3,156, 4,083]	6.46 [6.10, 7.29]
Phi	10.45	6.12	9.92	23,643 [21,082, 27,025]	6.40 [6.05, 6.68]
Ministral	5.26	12.86	12.24	5,184 [4,242, 6,482]	7.95 [7.65, 8.39]

$\psi_C \le \epsilon_C^2/2$, proving the second. For the final line, put $\eta = \epsilon^2/2$. Since $P_C \propto \widetilde{P}_C e^{\psi_C}$ and $0 \le \psi_C \le \eta$, Hoeffding's lemma on an interval of width η gives $\mathrm{KL}(\widetilde{P}\|P) \le \eta^2/8$. Pinsker then gives $\|P - \widetilde{P}\|_1 \le \eta/2$, hence $\mathcal{M}_{\mathrm{mass}}(x) \le V\epsilon^2/4$. Finally, $\sum_C P_C \rho_C \min\{2, \epsilon_C\} \le \rho\epsilon$, because $\sum_C P_C = 1$. □

The certificate is retrospective: computing its exact group probabilities, conditional divergences, and radii requires the original raw groups and a specified future query. It is not evaluated by the deployed compressor and does not constitute a free online guarantee, a task-quality predictor, or an end-to-end autoregressive bound. Newly appended entries are singleton groups with zero conditional radius; their value norms still enter V.

B.2 Fixed diagnostic protocol and aggregation

The diagnostic reuses two reserved-development examples from each of Qasper, HotpotQA, and GovReport, selected by seed 2026 and offset zero. The six source texts are reused across caps 8,192 and 16,384 and all three model families; short texts are not padded. Fractions are fixed at 0.2, 0.4, and 0.5. The selected layer indices are Qwen [3, 19, 31], Phi [0, 16, 31], and Ministral [0, 13, 25]: the first, middle, and last eligible layers in each model. All grouped query heads are included. Queries and appended KV entries come from unchanged full-model generation, with at most eight decode forwards per prompt before EOS. Actual future-step totals are 53, 51, 74 in model order. This gives 36 model/prompt/cap units and 39,960 head/query/fraction records. Input hashes and EOS-limited query counts agree across independently stored diagnostic records for every unit. This is an operator study, not a task benchmark.

For record j in prompt/cap unit u, let E_{uj} be the absolute ideal attention-output error and B_{uj} the corresponding bound. We first compute $\bar{E}_u = |J_u|^{-1} \sum_{j \in J_u} E_{uj}$ and $\bar{B}_u = |J_u|^{-1} \sum_{j \in J_u} B_{uj}$ at a fixed model and fraction. The main ratios are $R_u = \bar{B}_u/\bar{E}_u$, summarized by the median and interquartile range across twelve prompt/cap units. This gives equal weight to short and long pre-EOS continuations. It is distinct from both a pooled median over B_{uj}/E_{uj} and $\mathrm{median}_u \bar{B}_u / \mathrm{median}_u \bar{E}_u$. The radii columns likewise report medians of prompt/cap means of the head/query-specific maxima. Neither caps nor heads are independent samples.

B.3 Measured slack, layers, and retained fraction

Table 3 shows that the empirical radii are generally outside a small-radius regime. The global bound is far above the observed error. The refinement removes much of this worst-case slack but still exceeds error by several times. It is not described as tight. Table 4 reports all fractions using the same prompt-level unit; errors and bounds are absolute ℓ_2 norms.

As retention increases, the actual error decreases much more rapidly than the global bound, so the global bound/error ratio increases. The effect cannot be inferred from a small attention error alone. Layer differences are also substantial: at fraction 0.4, pooled head/query median global/error ratios range from 493 to 23,643 across the selected layers, whereas adaptive/error ratios range from 5.02 to 10.00. These layer summaries use a different unit from the prompt-level table and are explicitly descriptive.

Table 4: Absolute attention-output errors and bound slack across retained fractions. Error and right-hand-side columns are medians of prompt/cap means; ratio columns are medians [25th, 75th percentiles] of prompt-level ratios over twelve units per model.

Model	Fraction	Error median	Global bound median	Global/error	Adaptive/error
Qwen	0.2	0.4872	716.31	1,768 [1,130, 1,874]	6.72 [6.08, 7.64]
Qwen	0.4	0.1962	626.06	3,695 [3,156, 4,083]	6.46 [6.10, 7.29]
Qwen	0.5	0.1183	589.59	5,750 [4,448, 6,399]	6.42 [6.24, 7.11]
Phi	0.2	0.0490	458.69	9,559 [8,200, 11,321]	6.99 [6.66, 7.51]
Phi	0.4	0.0187	433.91	23,643 [21,082, 27,025]	6.40 [6.05, 6.68]
Phi	0.5	0.0123	420.22	36,027 [32,197, 41,487]	6.13 [6.00, 6.35]
Ministral	0.2	0.0996	260.82	2,393 [1,837, 2,892]	8.38 [8.13, 8.47]
Ministral	0.4	0.0409	241.88	5,184 [4,242, 6,482]	7.95 [7.65, 8.39]
Ministral	0.5	0.0264	227.45	7,727 [6,161, 9,528]	7.86 [7.53, 8.28]

Across the twelve prompt/cap mean errors at fraction 0.4, Spearman correlations are 0.440, 0.490, 0.351 for the global bound and 0.998, 0.900, 0.977 for the group-adaptive certificate, in Qwen/Phi/Ministral order. These retrospective associations on observed raw groups carry no independent-sample significance test or held-out predictive claim. The refined certificate's dependence on exact group probabilities partly explains its closer association with measured error. Per-head distributions and each prompt's raw records are retained for inspection.

B.4 Implementation audit and finite precision

The diagnostic traces the contiguous final groups while reproducing the deployed CALM floating-point operations and pair choices. All 324 layer/fraction comparisons have identical keys, values, scalar biases, and diagnostics to the deployed compressor. Every trace partitions the uncompressed prefix. Ideal representatives are reconstructed in FP64 from the corresponding stored BF16 states and the traced partition. The anchor is the computed FP32 CALM anchor promoted to FP64, avoiding a silent change of anchor. FP64 numerically evaluates the real-arithmetic formula; it is not interval arithmetic or an exact floating-point certificate.

No global-bound, refined-certificate, conditional-KL, or anchor-identity check failed. The numerical comparisons use absolute and relative tolerances of 10^{-9}. Maximum anchor discrepancy is 4.28×10^{-14} and maximum conditional-KL identity discrepancy is 2.82×10^{-12}. A failure would terminate the run and save the failing case; no head or prompt was excluded based on its bound quality. Mathematical self-tests are separate from these measurements.

Deployed BF16 output is evaluated through the same SDPA wrapper, including FP32-stored bias cast to query dtype. Each record separately contains its error relative to full FP64 attention, its displacement from the ideal representative, and full-BF16 versus full-FP64 error. At fraction 0.4, pooled head/query median deployment errors are 0.11387, 0.01145, 0.01613; median displacements from ideal representatives are 0.01533, 0.00216, 0.00238, in model order. These pooled medians are not the prompt-level values above. The triangle extension

$$\|o - o_{\mathrm{BF16}}\|_2 \le \mathcal{B}_{\mathrm{glob}} + \|\widetilde{o} - o_{\mathrm{BF16}}\|_2$$

passes every check. Numerical effects are not folded into the real-arithmetic theorem.

Only three layers per model and a short full-model query horizon are sampled. Compression-induced query drift, discrete task decisions, and serving latency are not bounded or measured here. The global bound remains valid but quantitatively loose on this subset. The refined certificate diagnoses where probability weighting and local radii reduce slack; it supplies no general guarantee that future compressed generation will preserve accuracy.

Artifacts. The measurement source and its hash, partition intervals, input and query hashes, all raw records, detailed logs, and environment metadata are stored in the supplementary artifact manifest. The figures and numeric inserts are regenerated by `code/scripts/build_bound_`

`artifacts.py` from the saved summary and analysis files in JavaScript Object Notation (JSON). No new model experiment is needed to rebuild them.

C Allocation Details of the Test Construction

C.1 Probe statistics and merge selection

Prefill supplies the queries at the last $W = 64$ real prompt positions. For a KV head shared by G query heads, the $n_{\text{probe}} = GW$ scaled queries are pooled into an anchor $x_0 = n_{\text{probe}}^{-1} \sum_m x_m$ and a diagonal covariance estimate D. The anchor identity applies at this pooled mean, not simultaneously at every query head. Initial importance is computed by averaging individual probe distributions,

$$p_i = \frac{1}{n_{\text{probe}}} \sum_m \frac{\exp(x_m^\top k_i + b_i)}{\sum_j \exp(x_m^\top k_j + b_j)}. \tag{14}$$

These probes are evaluated over the completed prompt. They are not the causal attention matrices that were used during prefill. Every prompt token is available when compression is applied, including the benchmark question, and no answer is consulted.

The first four entries and the last 64 entries are protected. Disjoint adjacent pairs in the remaining cache are considered in each round. For a pair, define $\Delta k = k_a - k_b$, $\Delta v = v_a - v_b$, and let σ_v^2 be the mean squared value magnitude over the current mergeable cache for that head. The ranking score is

$$c_{ab} = (p_a + p_b)w(1-w)\left[\Delta k^\top D \Delta k + \lambda \frac{\|\Delta v\|_2^2}{d\sigma_v^2 + 10^{-8} d}\right], \qquad \lambda = 1. \tag{15}$$

The lowest-cost pairs are merged, and their importances are added. At most $\max(1, \lfloor \lfloor L/2 \rfloor / 2 \rfloor)$ pairs are selected from an interior of length L in one round. Thus approximately one quarter of the available entries can be removed at a time. This restriction avoids forcing every original pair to merge when the requested budget is below one half. Pairing is reconstructed after each round, and the final round removes only the entries needed to reach the budget.

The target is $B = \min(N, \max(69, \lfloor fN \rfloor))$, with $f = 0.4$ in the main study. Selection is head-specific; all physical KV heads retain the same number of entries. The procedure is more than top-k truncation: discarded addresses contribute to weighted keys, values, and mass. Nevertheless, the local tree and diagonal query model are approximations, and global optimality is not claimed.

C.2 Address interpretation and local surrogate

If a sampled token address I is mapped deterministically to one of B representatives, its group address $C = g(I)$ satisfies $H(C) \le \log B$ and $I(I; C) = H(C)$. This bounds the address variable, not information in the continuous representative vectors or mutual information between a document and its cache. Reducing the address set while limiting Eq. 5 provides an operational representation-bottleneck interpretation. It is not a solution to a global information-bottleneck objective.

For a two-child pair, the second-order expansion is $\psi_{ab}(x) = \frac{1}{2} w(1-w)(\delta^\top \Delta k)^2 + O(\|\delta\|^3)$. Under the local query model $\mathbb{E}\delta = 0$, $\mathrm{Cov}(\delta) = D$, set $a = \Delta k^\top D \Delta k$ and $g_v = \|\Delta v\|^2/(d\sigma_v^2)$, assuming $\sigma_v^2 > 0$ and omitting the numerical stabilizer. The normalized RMS of the linearized conditional value change is

$$w(1-w)\sqrt{a g_v} \;\le\; \tfrac{1}{2} w(1-w)(a + g_v). \tag{16}$$

Thus the sum in Eq. 15 contains the modeled KL curvature and controls a linearized value-error term. Off-diagonal covariance, higher-order effects, changing group probability, and future query shift are not controlled by this argument. Multiplication by observed salience remains an allocation heuristic.

D Implementation and Baseline Adaptations

Execution order. The full prompt is tokenized and truncated before inference. Full prefill is performed with the native SDPA backend while real post-RoPE queries are captured. Cache tensors

are compacted only after all layers finish prefill. Query probes are released after compression. The first token is selected from the already computed prefill logits; subsequent forwards receive explicit positions $N, N+1, \ldots$. These positions refer to the tokenized, truncated prompt, not to removed portions of the original document. Every cached representative is causally valid at a one-token decode step, so its additive bias is used without reinterpreting shortened storage indices as positions.

Layer discovery and state. The controller recognizes attention modules with either separate query projections or fused query–key–value (QKV) projections. An absent supported layer or a no-op compaction raises an error. Qwen's linear-attention layers are left intact. The general backend also excludes short sliding-window caches. Phi's configured window exceeds the evaluated lengths, so its 32 layers are compressed. Tensor storage is deduplicated by allocation and recurrent state is traversed recursively, including tensors nested in lists. Bias arrays are included even when all their entries are zero. Python object overhead and CUDA memory not tracked by PyTorch are not presented as cache tensor bytes.

SnapKV. The last 64 prefill queries are used with a causal observation-window mask. Attention is averaged over GQA query heads, summed over observation positions, and width-five average pooling is applied to older-token scores. The suffix is protected and the remaining budget selects the largest pooled scores. Unlike CALM probes over the completed prefix, these scores follow the causal observation convention. Both methods have access to the same completed prompt, but their score estimators differ.

ChunkKV. Recent-window scores are aggregated across physical heads and grouped into 20-token chunks, following the public kvpress implementation. The observation suffix is reserved. Complete chunks are retained and a small remainder is filled by the best unselected token scores, ensuring an identical total token budget. This suffix/remainder rule is an explicit adaptation; original fixed-chunk budgets need not produce exactly the same entry count.

KeepKV-public. The port follows the released `modify_llama_oursk.py` centroid/vote mechanism at commit `a25a2b5ef5d924347195786ee019439ea956e6a6`. Recent-window unnormalized exponential scores are averaged across grouped heads. Important entries and the recent suffix are retained; other entries are assigned to the most similar retained key when cosine similarity exceeds 0.5. Weighted key/value centroids and log vote counts are stored. Unlike the archived source, this port does not explicitly reserve sink tokens and omits its recent-suffix causal-exposure correction, using a uniform window average. These changes affect selection and centroid weights; the results apply to this variant. Retained entries are explicitly preserved when duplicate keys tie. The implementation is not the paper's radial Zero Inference-Perturbation Merging (ZIP-Merging) equation, and it does not keep a host copy for periodic reload. The original source and provenance are archived. An investigated radial variant is excluded from the main method set.

Local merge procedure. For each eligible layer and KV head: (1) estimate the pooled anchor, diagonal query covariance, and averaged probe importance; (2) separate the four sink entries and the 64-entry suffix; (3) form disjoint adjacent interior pairs; (4) evaluate Eq. 15, choose the permitted lowest-cost pairs, and apply Eq. 3; (5) preserve order, add child importances, and repeat until the target is met; (6) concatenate protected entries and representatives, cast keys/values to BF16, and install the shortened tensors. The variance estimate remains fixed within a compaction; the value scale is recomputed from the current mergeable cache each round. Candidate pairing and ordering are independent for each physical KV head. Batch size one is the evaluated interface.

Stopping and refresh. Generation stops on the union of model-config and tokenizer EOS IDs, or at the official task limit. SAMSum additionally stops once a newline has been decoded after at least two output tokens. Its scored prediction, like TREC and TriviaQA, follows the official first-line preprocessing. At a refresh event, only the current decode query per query head is available; no rolling 64-query decode buffer is implied. Recompression returns to the original prompt budget. Decode steps are zero-indexed: the refresh follows the forward at step 128, which has already produced output token 130. Thus output tokens 1–130 match the no-refresh CALM run, and token 131 is the earliest output that could be affected. Full-cache controls match their main-run outputs exactly under the same stopping policy.

Cost and precision. Probe scoring costs $O(MNd)$ per KV head with $O(MN)$ temporary scores. Successive interior lengths shrink geometrically; sorting retained positions gives a conservative $O(N \log N + Nd)$ merge-phase cost. Merge arithmetic is FP32 and stored K/V are BF16, with FP32 bias metadata cast to query dtype. These arithmetic costs are not latency guarantees. The KeepKV port above is evaluated only on the designated 32-example subset.

E Attention-Matching Comparator and Numerical Sensitivity

The comparator is a controlled port of AM-HighestAttnKeys-basic from the official Attention Matching repository (Zweiger et al., 2026), commit `2020f6d37940fe013e7e7f1ebf409e5d107c5670`. The MIT-licensed source, configuration, paper version, and per-file hashes are archived. This is not the on-policy default, orthogonal matching pursuit (OMP) key selection, or a reproduction of the original evaluation. The shared short query collection is a substantial adaptation: real queries from the final 64 prompt positions are flattened across G query heads into $64G$ rows per physical KV head. The completed prompt, including its question, is visible; generated answer labels are not used.

For query matrix Q and key matrix K, let the scaled logits be $L = QK^\top s$. Define $P = \mathrm{softmax}(L)$ and select the budgeted columns by $\sqrt{\mathrm{mean}_m P_{mi}^2}$. No additional sink or suffix entry is forced into that selection. Set $c_m = \max_i L_{mi}$, $M_{mj} = \exp(L_{m,j} - c_m)$ for selected keys, and $y_m = \sum_i \exp(L_{mi} - c_m)$. The mass fit minimizes a row-rescaled residual. Following the released basic solver, FP32 `gels` initializes a least-squares solution, which is clamped to $[e^{-3}, e^3]$. Three power iterations estimate a step size; two projected steps refine the bounded fit. The scalar bias is the logarithm of the fitted weight. This is not claimed to solve bounded least squares exactly or to be an entirely closed-form algorithm.

For values, let $X = \mathrm{softmax}(L_{\text{selected}} + \beta)$, with β rounded to the query dtype as in deployment, and let $Y = PV$. AM-basic computes the FP32 least-squares solution of $XC \approx Y$ with `gels`. It retains the upstream exception/nonfinite fallback at 10^{-6} regularization. Finite but poorly conditioned solutions do not trigger that fallback. Cached keys and fitted values are stored in BF16; scalar biases are FP32. The local solver random generator is reset to seed 2026 per head. Entries are sorted back into position order only after fitting, an attention-invariant permutation.

The operator diagnostic reveals very large finite values in some basic fits. We therefore define AM-ridge as a separately labeled numerical sensitivity. It preserves key selection, mass fitting, query collection, and physical budgets, and changes every value objective to $\|XC - Y\|_F^2 + 10^{-6}\|C\|_F^2$. The fixed coefficient is the archived upstream fallback value, not selected by a task-score sweep. Its solution is

$$C = X^\top (XX^\top + 10^{-6} I)^{-1} Y. \tag{17}$$

The small dual Gram matrix and Cholesky solve use FP64 so that the regularizer is not lost to FP32 Gram rounding; the final values remain BF16. There is no case-specific fallback, clipping, or exclusion. An exception is a failed case, not a replacement prediction. Effective numerical rank, fitted residual, value norm, and extrema are logged. These diagnostics add runtime overhead and are part of the measured eager implementation.

Independent checks compare the basic solver with the archived official class in overdetermined and underdetermined regimes, yielding identical selected keys, biases, and values with matched FP32 inputs and solver seeds. Independent NumPy least-squares discrepancies are below 4×10^{-7}. Ridge solutions agree with FP64 direct and singular value decomposition (SVD) formulas within 5.73×10^{-11} relative error, including rank-deficient cases. Keys and mass biases are identical across the two variants. Physical budget, GQA association, allocation accounting, and fraction-one identity are also checked. These mathematical test tensors never enter benchmark estimates.

On the fixed real Phi 16K development state, a head can have 192 reference rows but 6,553 retained columns. A small reference residual therefore does not identify behavior at future queries. In the operator diagnostic, the largest basic prompt-mean relative error is 2,929.657 on Phi; all recorded values remain finite. The corresponding ridge maximum across Phi prompts is 0.179243. The basic mean is tail-sensitive, so the median and extrema are retained alongside it. This numerical result concerns the disclosed short-query port and its solver conditions, not a claim that the full Attention Matching system is intrinsically unstable.

No model weights are updated by either comparator. The two projected mass iterations and numerical linear solves optimize cache representatives only. No repeat-prefill, self-study generation, on-policy layer rerun, head-budget optimization, or question-independent cache reuse is implemented. Those omissions limit comparison with the original method; they are not hidden behind the method name.

F Models, Data, and Reproduction Details

The model revisions are pinned in the download manifest:

- Qwen/Qwen3.5-4B: `851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a`.
- microsoft/Phi-4-mini-instruct: `cfbefacb99257ffa30c83adab238a50856ac3083`.
- mistralai/Ministral-3-3B-Instruct-2512-BF16: `b6d637bef2393152b3da2b2fde72eecdee30557e`.

Qwen has four KV heads of dimension 256 in each full-attention layer. Phi and Ministral have eight KV heads of dimension 128. The loaded checkpoints contain 4,539,265,536; 3,836,021,760; and 3,849,090,048 parameters, respectively. Published model names and total loaded parameter counts are both reported to avoid treating effective-parameter labels as total parameter counts. A Gemma E4B checkpoint was considered during preparation but excluded because its total parameter count is substantially larger; it contributes no reported comparison.

The LongBench dataset revision is `5e628be450b7e67fb7ae6e201bd6d8f7056f7672`. The upstream prompt/scoring code is pinned to revision `2e00731f8d0bff23dc4325161044d0ed8af94c1e`. Downloads use validated Hypertext Transfer Protocol (HTTP) range requests and published Git Large File Storage (Git LFS) checksums based on the 256-bit Secure Hash Algorithm (SHA-256). The downloaded archive, extracted public JSON Lines (JSONL) files, exact split indices, input hashes, and source-row hashes are retained. There are no mocked benchmark examples.

NarrativeQA evaluates narrative comprehension (Kočiský et al., 2018); Qasper uses questions about scientific papers (Dasigi et al., 2021); HotpotQA, 2WikiMQA, and MuSiQue cover multi-hop question answering (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). GovReport, QMSum, and MultiNews provide long-document or multi-document summarization settings (Huang et al., 2021; Zhong et al., 2021; Fabbri et al., 2019). TriviaQA and SAMSum provide trivia question answering and dialogue summarization (Joshi et al., 2017; Gliwa et al., 2019). MultiFieldQA and the few-shot TREC setup follow LongBench. ROUGE-L is evaluated with the implementation released with LongBench (Lin, 2004; Bai et al., 2024).

Sampling and analysis. Each dataset uses Python `random.Random(2026)` to shuffle source indices. Development takes [0:8] and evaluation takes [8:72]; the latter is divided into cohort A [8:40] and cohort B [40:72] for robustness checks. Sorting is only for execution. Controls take [8:24], prefix sensitivity [8:12], and the operator study [40:42]. Source texts are shared across caps and models. Main bootstrap seed is 7321; joint-cap selection/component seed is 40117; output-length seed is 19427; prefix and refresh seeds are recorded in their analysis scripts. Intervals use 10,000 resamples unless specified otherwise; the supplementary operator-ranking analysis uses 20,000. No failed prediction or unfavorable dataset is removed from a completed grid.

Environment and source audit. The observed PyTorch version is `2.12.0a0+5aff3928d8.nv26.05`, corresponding to NVIDIA's 26.05 framework build. Python is 3.12.3, Transformers is 5.16.1, flash-linear-attention is 0.5.2, and causal-conv1d is 1.7.0. Package locks and a framework-container reproduction route are supplied. All benchmark code is frozen in an archived source snapshot; formatting and docstring-only changes have identical executable abstract syntax trees. Worker shards operate on disjoint case keys, and completed cases are reused only after their hashes are checked. Scheduling changes and intentional process exits are logged; comparative timing uses the separate GPU-isolated procedure.

Table 5: Official task metrics and generation limits used in the experiment. Every score is scaled to 0–100 and maximized over the official reference answers where applicable.

Task	Metric	Maximum new tokens
NarrativeQA	Answer F1	128
Qasper	Answer F1	128
MultiFieldQA	Answer F1	64
HotpotQA	Answer F1	32
2WikiMQA	Answer F1	32
MuSiQue	Answer F1	32
GovReport	ROUGE-L	512
QMSum	ROUGE-L	512
MultiNews	ROUGE-L	512
TREC	Classification accuracy	64
TriviaQA	Answer F1	32
SAMSum	ROUGE-L	128

Reproduction entry points. From the project root, source `code/env.sh`, preserve the documented NVIDIA framework, and install the locked dependencies. The reproduction entry point documented in `code/README.md` runs the experiment suite on one dedicated GPU and checks complete saved cases before resuming. Its analysis-only mode recomputes scores, audits, numerical tables, and figures. The README also records downloads, pinned sources, every phase, and both language builds. The consolidated `code/experimental_results/results.json` contains the main results and content-hashed links to auxiliary records. There is no training loss because model weights are never trained. Archived source snapshots preserve experiment provenance; the supplied style files are unchanged.

G Actual Input Lengths

The nominal 8K and 16K conditions impose upper bounds on tokenized input length; they do not make every prompt equally long. The LongBench tasks have different source-length distributions, and each model's tokenizer can assign a different length to the same source example. Consequently, an example may reach the cap for one model but remain below it for another. We therefore describe the complete 768-example length distribution for every model–cap condition before considering the subset that actually reaches the stated limit. Table 6 reports both the central tendency and the amount of truncation, separating the nominal experimental setting from the context lengths realized by the models.

Table 6: Actual tokenized lengths in the 64-example-per-task evaluation grid. Each row contains 768 source examples; no prompt is padded.

Model	Cap	n	At cap	Median	Max.
Qwen3.5-4B	8K	768	419	8,192	8192
Qwen3.5-4B	16K	768	167	9,179	16384
Phi-4-mini	8K	768	397	8,192	8192
Phi-4-mini	16K	768	81	8,678	16384
Ministral-3-3B	8K	768	432	8,192	8192
Ministral-3-3B	16K	768	178	9,408	16384

At 8K, between 397 and 432 of the 768 inputs reach the cap for each model, and all twelve tasks remain represented. At 16K, only 81–178 inputs reach the cap, spanning six to nine tasks depending on the tokenizer. Figure 4 therefore provides a conditional stress-case analysis rather than a replacement for the complete twelve-task result. Qwen's intervals include zero at both caps, whereas the Phi and Ministral subsets show negative CALM-minus-full differences. Because subset membership

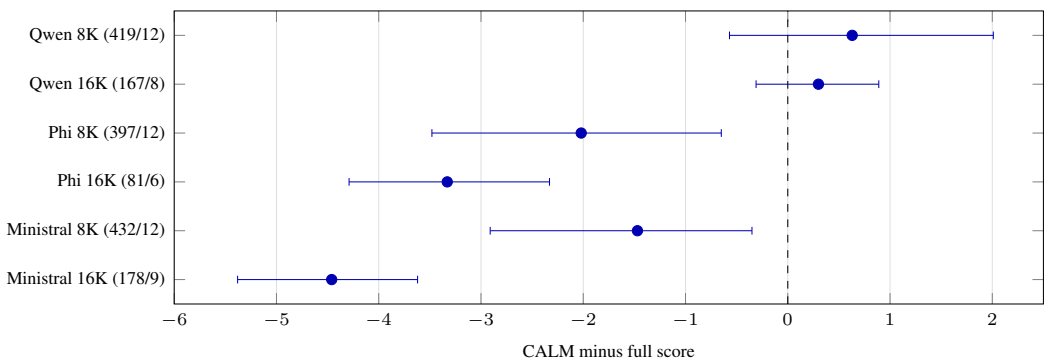

Figure 4: CALM-minus-full estimates restricted to inputs actually at the stated cap. Points show paired means and horizontal bars show 95% confidence intervals (CIs); labels give the number of inputs and represented tasks. Only represented tasks receive macro weight, so these task mixtures differ from the full twelve-task estimand.

and task composition change across models and caps, these contrasts should not be interpreted as a controlled estimate of how performance changes with length alone.

Table 7: Full-set and cohort B differences from full cache, with paired 95% task-stratified bootstrap intervals. Full-set estimates use 64 examples per task and 768 paired sources per row; cohort B estimates use 32 examples per task and 384 paired sources per row. Cohort B is contained within the full set and is reported as a descriptive robustness check, not an independent replication.

			Full set (64/task)		Cohort B (32/task)	
Model	Cap	Method	Δ	95% CI	Δ	95% CI
Qwen3.5-4B	8K	AM-basic	-5.46	[-6.75, -4.20]	-5.88	[-7.86, -3.94]
Qwen3.5-4B	8K	AM-ridge	-4.95	[-6.17, -3.75]	-5.34	[-7.17, -3.51]
Qwen3.5-4B	8K	CALM	+0.16	[-0.43, +0.82]	+0.39	[-0.54, +1.45]
Qwen3.5-4B	16K	AM-basic	-5.39	[-6.66, -4.19]	-6.08	[-8.02, -4.20]
Qwen3.5-4B	16K	AM-ridge	-5.16	[-6.42, -3.96]	-5.74	[-7.70, -3.88]
Qwen3.5-4B	16K	CALM	+0.35	[-0.20, +0.96]	+0.50	[-0.36, +1.54]
Phi-4-mini	8K	AM-basic	-14.48	[-16.28, -12.64]	-16.13	[-18.86, -13.39]
Phi-4-mini	8K	AM-ridge	-6.18	[-7.48, -4.93]	-6.97	[-8.86, -5.17]
Phi-4-mini	8K	CALM	-1.14	[-1.69, -0.62]	-1.47	[-2.24, -0.80]
Phi-4-mini	16K	AM-basic	-15.02	[-16.93, -13.17]	-17.31	[-20.21, -14.49]
Phi-4-mini	16K	AM-ridge	-6.45	[-7.82, -5.15]	-7.48	[-9.48, -5.56]
Phi-4-mini	16K	CALM	-1.15	[-1.79, -0.55]	-1.24	[-2.23, -0.33]
Ministral-3-3B	8K	AM-basic	-8.08	[-9.69, -6.50]	-9.92	[-12.35, -7.53]
Ministral-3-3B	8K	AM-ridge	-7.75	[-9.31, -6.23]	-9.85	[-12.26, -7.53]
Ministral-3-3B	8K	CALM	-1.74	[-2.61, -0.92]	-2.50	[-3.94, -1.22]
Ministral-3-3B	16K	AM-basic	-9.36	[-11.08, -7.73]	-11.32	[-13.80, -8.87]
Ministral-3-3B	16K	AM-ridge	-9.07	[-10.72, -7.46]	-11.25	[-13.76, -8.83]
Ministral-3-3B	16K	CALM	-1.52	[-2.37, -0.73]	-1.62	[-2.83, -0.51]

H Sampling Cohorts and Paired Comparisons

The main grid has 64 examples per task, two caps, three model families, and six methods. Every method uses identical input IDs within each model, task, cap, and source example. Both AM variants are evaluated on all 64 examples. KeepKV-public appears only in the five-method, 32-example comparison in Appendix L; it is not represented as a 64-example result.

Cohorts A and B were determined by seed and index before their scores were summarized. Their separate estimates are robustness checks, not preregistered confirmatory hypothesis tests. The cache method and budget are identical across cohorts. The fixed value-fitting stability sensitivity has no task-score parameter sweep. Main predictions, prompt controls, component contrasts, prefix-protocol runs, and operator measurements remain separate artifact collections.

The eviction contrasts are model-dependent: Qwen intervals include zero, while most Phi and all Ministral intervals favor SnapKV or ChunkKV. In contrast, CALM exceeds both controlled AM variants in every model–cap condition, with all intervals bounded away from zero. Figure 5 preserves the cap-specific paired comparisons while making the two regimes visually explicit.

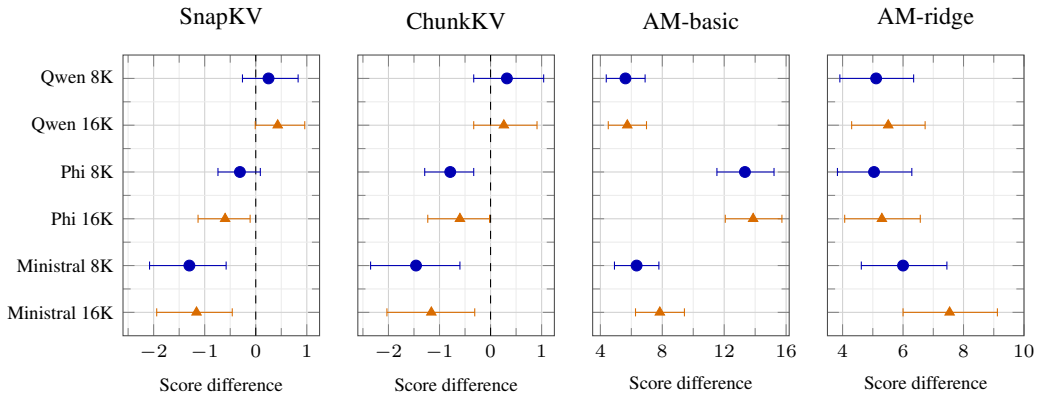

Figure 5: CALM-minus-comparator score differences on all 64 examples per task. Points show paired means and horizontal bars show 95% task-stratified bootstrap intervals; circles denote 8K and triangles denote 16K. Positive values favor CALM. Dashed lines mark zero in the eviction panels. Intervals are not multiplicity-adjusted.

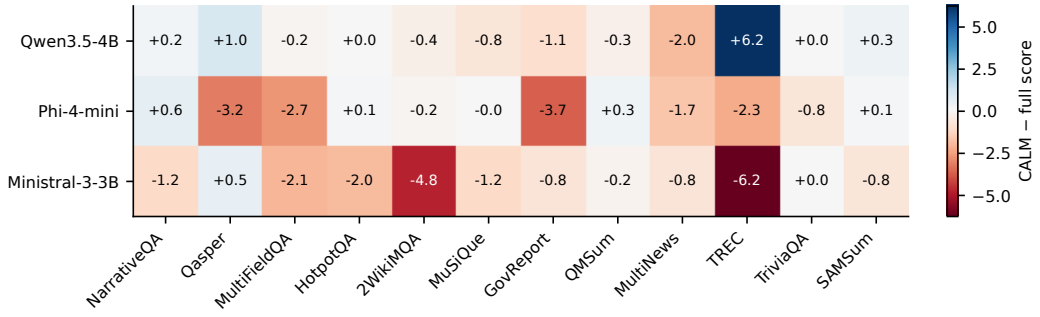

Figure 6: CALM-minus-full task differences on 64 source examples, averaged over the two cap conditions. Each cell contains 64 distinct texts, not 128 independent texts.

I Allocation Controls and Future-Query Diagnostics

Each allocation contrast uses the first 16 evaluation selections per task, with Qwen SAMSum under the chat-wrapped prompt. Four additional component interventions are included: no mass removes scalar corrections after otherwise identical merges; fixed order selects left-to-right; no value removes the normalized value term; zero anchor changes both representatives and pair weights while retaining multiplicity correction and observed dispersion. The latter is not merely a bias ablation.

The three selection controls change only selection scoring or sampling. Salience-only uses $p_a + p_b$, removing $w(1-w)$ as well as key/value distortion weighting. Isotropic geometry uses $(\mathrm{tr}\, D/d)I$. Balanced random chooses one uniform candidate per approximately equal bin of current representative-pair indices. Its local seed is the first 64 bits of the input-ID SHA-256, interpreted little-endian, combined with 2026 by exclusive OR (XOR), and reduced modulo $2^{63}-1$. The generator restarts at each eligible layer: equally shaped layers can share draws, and the experiment supplies one deterministic random realization per input. It does not estimate variation across random seeds.

All controls preserve the protected prefix and suffix, construction of disjoint adjacent pairs, per-round removal counts, total number of rounds, representative update arithmetic, and final budget. Actual selected groups and realized tree depths change. Balanced current indices need not balance original token spans. Salience-only and random controls still compute unused intermediate statistics in this instrumented implementation, so no latency advantage is inferred from their simpler selection rule.

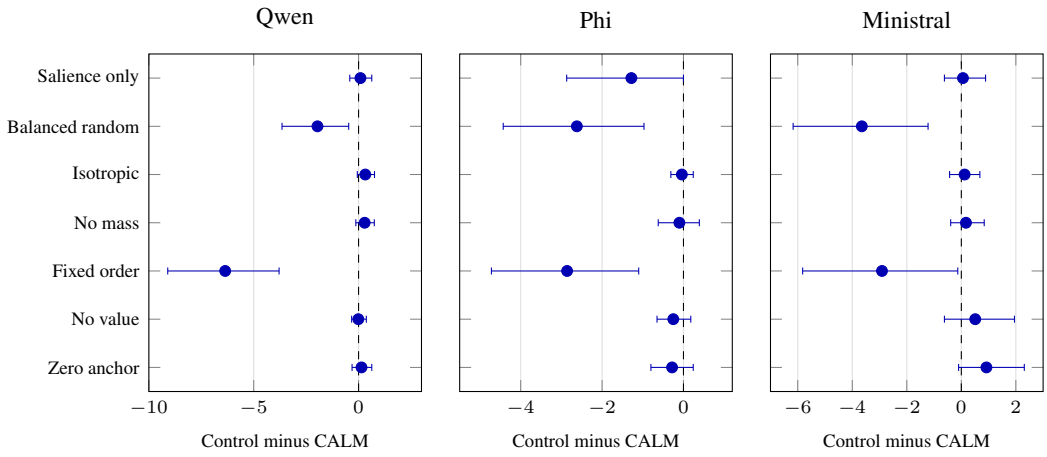

Figure 7: Control-minus-CALM differences on 192 distinct sources per model. Points show paired means and horizontal bars show 95% task-stratified bootstrap intervals; positive values favor the control. Both caps are averaged within each source before resampling. The Phi salience-only upper endpoint is positive before rounding.

Future-query cohort. Two examples from every task are selected from the beginning of the second 32-example cohort. At both caps, this gives 48 prompt/cap units from 24 source texts per model and 144 units across models. The same source texts are reused across model families. First/middle/last eligible layer indices are Qwen [3,19,31], Phi [0,16,31], and Ministral [0,13,25]. At most eight real full-model decode forwards are observed; EOS and SAMSum stopping are respected. No artificial queries or padded long contexts are used.

All methods use identical actual queries and fresh generated KV entries, appended uncompressed. Each attention product and softmax in this measurement is evaluated in FP32 over the stored BF16 K/V; scalar biases first undergo the controller's query-dtype rounding. This measures the stored representation using a common numerical evaluator, not every rounding operation of deployed SDPA. Deployment-versus-ideal precision is measured separately in Appendix B. Relative error divides the Frobenius output difference by the full output norm, with a 10^{-8} floor. Available steps and layers are averaged within each prompt before means, medians, or extrema are computed. A zero-step prompt would have a missing mean, not an invented zero.

The five-method diagnostic has 14,850 layer/step/method records. Across its components, all 144 input/source hashes, selected layers, observed continuations, next tokens, and stopping rules match. The mass-fitting diagnostics also match for all 2,880 AM head fits. These repeated measurements are not thousands of independent documents. Full-task runs on the same sources match the diagnostic continuations exactly.

The corresponding future-query means and medians are visualized in Figure 3. They use 48 prompt/cap units per model, representing 24 source texts, and retain the AM-basic tails.

Fitting and evaluation are different estimands. The stored-reference residual is measured per physical KV head on its $64G$ fitting queries. Future error aggregates the full query-head/output tensor at a real future layer/step, then averages within prompts. These ratios use different query populations and aggregation units, so their quotient is not a calibrated generalization multiplier. AM-ridge's 95th-percentile stored-reference residuals are 0.001794, 0.006016, and 0.009596 in model order, while its future prompt-mean errors are 0.218905, 0.091874, and 0.138567. Small reference residuals thus do not ensure small future error in this fixed-query implementation. This observation does not establish how the full on-policy system would behave.

Linkage to complete task outcomes. The diagnostic is joined to task predictions only by identical model/task/index/cap and input hash. Both caps are then averaged within each source example. Figure 8 uses 24 points per model/method; correlations are descriptive Spearman coefficients

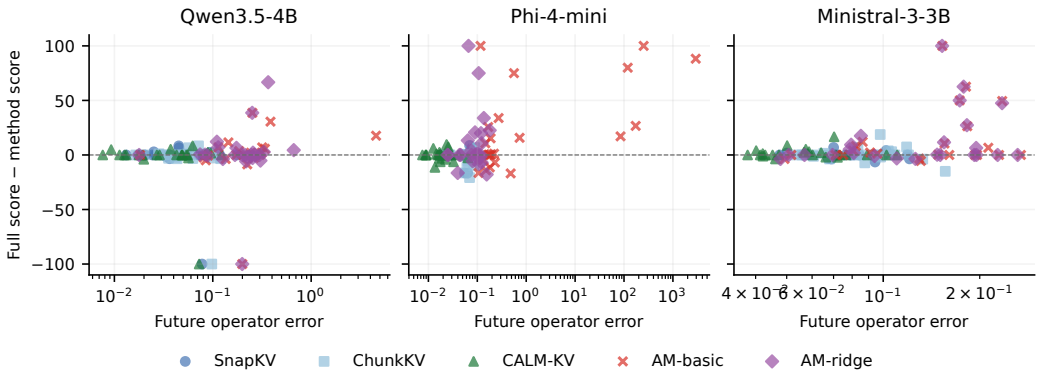

Figure 8: Future operator error versus paired task loss, after averaging both caps within each source text. Positive task loss means a lower compressed score. Each method/model has 24 source points; all methods share those texts. Log axes retain the AM-basic tail.

Table 8: Descriptive paired uncertainty on 24 source texts per model, two per task, after averaging caps. ΔE is CALM-minus-comparator mean relative operator error. $p_{\downarrow E,\downarrow S}$ is the percentage of all 24 texts with both lower CALM error and lower CALM task score; ties remain in the denominator. Limits use 20,000 paired bootstrap draws within the twelve fixed tasks. With only two texts per task, their empirical support is limited and population coverage is not guaranteed. The stability comparison focuses on AM-ridge; Figure 3 also reports AM-basic.

Model	Comparator	ΔE	95% limits	$p_{\downarrow E,\downarrow S}$	95% limits
Qwen	SnapKV	-0.0081	[-0.0117, -0.0045]	16.7%	[8.3, 25.0]%
Qwen	ChunkKV	-0.0149	[-0.0188, -0.0111]	25.0%	[12.5, 37.5]%
Qwen	AM-ridge	-0.1737	[-0.2057, -0.1416]	20.8%	[8.3, 33.3]%
Phi	SnapKV	-0.0397	[-0.0439, -0.0357]	25.0%	[12.5, 37.5]%
Phi	ChunkKV	-0.0411	[-0.0450, -0.0372]	41.7%	[29.2, 54.2]%
Phi	AM-ridge	-0.0711	[-0.0765, -0.0655]	20.8%	[8.3, 33.3]%
Ministral	SnapKV	-0.0200	[-0.0224, -0.0176]	37.5%	[29.2, 45.8]%
Ministral	ChunkKV	-0.0290	[-0.0330, -0.0250]	29.2%	[16.7, 41.7]%
Ministral	AM-ridge	-0.0776	[-0.0877, -0.0674]	12.5%	[4.2, 20.8]%

with average ranks for ties and no independent-sample significance test. Task metrics differ across datasets, and operator queries come from the full model. Associations therefore cannot identify a causal mechanism or measure accumulated compressed-generation query shift.

In Qwen/Phi/Ministral order, descriptive Spearman coefficients between operator error and paired task loss are $(0.22, -0.05, -0.18)$ for SnapKV, $(-0.08, -0.09, -0.07)$ for ChunkKV, $(0.07, -0.09, -0.11)$ for CALM, $(0.46, 0.50, 0.25)$ for AM-basic, and $(0.08, 0.12, 0.31)$ for AM-ridge. Thus the association varies across methods; weak CALM coefficients do not establish that operator error is unrelated to task outcomes.

The same source-text draws are used across methods and models. Conditional lower-score fractions among lower-error cases, symmetric strict ranking disagreements, and undefined-denominator counts are retained in the JSON. A degenerate interval reflects limited observed support, not certainty.

J Output Length and Compression Before the First Output

Length strata are determined from full-cache generated IDs before compressed outcomes are inspected. The primary count removes the union of `tokenizer.all_special_ids`, decoder-

added IDs marked special, and effective model/tokenizer EOS IDs. The union matters for Phi, whose EOS 200020 is skipped by the decoder but absent from the tokenizer's listed special IDs. Whitespace, newline tokens, and non-special thinking markers remain counted. This is neither word count nor length after benchmark first-line preprocessing. Raw-ID lengths including EOS are also stored. Zero-content outputs form a separate explicit bucket.

Table 9: CALM-minus-full results by full-defined content-token length on the 64-example evaluation set. Each panel reports paired examples/tasks, score difference and interval, exact ID/text equality, and full terminal-EOS count.

(a) Qwen3.5-4B

Length	Cap	n / tasks	Δ [95% CI]	IDs = (%)	Text = (%)	EOS n
1	8K	34 / 6	+0.00 [+0.00, +0.00]	97.1	97.1	34
2–4	8K	143 / 7	-0.44 [-1.11, +0.00]	96.5	96.5	143
5–16	8K	133 / 8	+0.06 [-0.84, +1.02]	91.7	91.7	133
17+	8K	458 / 12	+0.25 [-0.42, +0.97]	22.5	22.5	233
1	16K	32 / 6	+0.00 [+0.00, +0.00]	100.0	100.0	32
2–4	16K	153 / 7	+0.00 [+0.00, +0.00]	98.0	98.0	153
5–16	16K	126 / 8	+0.13 [-0.57, +1.04]	92.9	92.9	126
17+	16K	457 / 12	+0.37 [-0.27, +1.05]	22.8	22.8	226

(b) Phi-4-mini

Length	Cap	n / tasks	Δ [95% CI]	IDs = (%)	Text = (%)	EOS n
1	8K	29 / 6	+0.00 [+0.00, +0.00]	100.0	100.0	29
2–4	8K	155 / 6	-0.13 [-1.12, +0.82]	92.3	92.3	155
5–16	8K	180 / 9	-0.37 [-1.03, +0.24]	69.4	69.4	167
17+	8K	404 / 12	-3.08 [-4.55, -1.71]	14.1	14.1	172
1	16K	26 / 6	+0.00 [+0.00, +0.00]	100.0	100.0	26
2–4	16K	159 / 6	-0.19 [-1.18, +0.65]	91.8	91.8	159
5–16	16K	174 / 9	-0.20 [-0.98, +0.65]	68.4	68.4	162
17+	16K	409 / 12	-2.69 [-3.98, -1.41]	13.2	13.2	176

(c) Ministral-3-3B

Length	Cap	n / tasks	Δ [95% CI]	IDs = (%)	Text = (%)	EOS n
1	8K	14 / 5	+0.00 [+0.00, +0.00]	100.0	100.0	14
2–4	8K	170 / 7	-2.01 [-3.98, -0.37]	87.6	87.6	170
5–16	8K	171 / 8	-1.42 [-2.69, -0.25]	67.8	67.8	158
17+	8K	413 / 12	-0.71 [-2.49, +1.21]	13.1	13.1	134
1	16K	13 / 5	+0.00 [+0.00, +0.00]	100.0	100.0	13
2–4	16K	162 / 7	-0.91 [-2.43, +0.33]	90.1	90.1	162
5–16	16K	169 / 8	-1.41 [-2.88, -0.15]	68.6	68.6	156
17+	16K	424 / 12	-0.64 [-2.49, +1.29]	16.5	16.5	127

Every method within a bucket uses the same full-defined members. Paired score differences are averaged equally over the tasks represented in that bucket; identity rates instead weight examples equally. Small strata and zero-width intervals are flagged in the JSON. Different buckets contain different task mixtures, so neither a causal length trend nor the overall twelve-task macro average follows by weighting buckets by their sizes. A one-content-token full answer can change after the fixed first output because the next compressed forward can predict another token rather than EOS. The original Qwen NarrativeQA example 105 illustrates this: full cache outputs `Nothing` then EOS; CALM continues `Nothing; her brother was sent to school.` Both receive zero score, which does not establish identical generation.

Matched split-prefill protocol. The first $N-1$ prompt tokens are prefilled and compressed, then the final prompt token is forwarded at position $N-1$. The first generated ID comes from that compressed-state forward; later positions are $N, N+1, \ldots$. Full cache uses the same split schedule. At the first output, the cache has $B(N-1)+1$ entries, which differs from the main $B(N)$ by

zero or one. Logged prefix memory is measured before the last prompt-token append. The anchor and observations also omit that last prompt token. These jointly changed conditions make this an operational protocol sensitivity, not a pure first-token intervention.

Table 10: Full split-versus-monolithic differences. First-ID and complete-sequence disagreement expose numerical/execution-schedule effects that must be distinguished from compression.

Model	Cap	n	Full drift [CI]	First $\neq$	IDs $\neq$	Text $\neq$	Score $\neq$
qwen	8K	48	-0.48 [-1.10, +0.05]	2.1	27.1	27.1	22.9
qwen	16K	48	-0.25 [-0.57, +0.05]	2.1	25.0	25.0	25.0
phi	8K	48	-0.33 [-0.89, +0.25]	0.0	35.4	35.4	29.2
phi	16K	48	+0.27 [-0.30, +0.85]	0.0	41.7	41.7	33.3
ministral	8K	48	-0.00 [-0.28, +0.29]	0.0	33.3	33.3	22.9
ministral	16K	48	-0.16 [-0.54, +0.17]	0.0	31.2	31.2	22.9

The cohort has four fixed evaluation selections per task: 48 distinct texts per model at two caps. Full, SnapKV, CALM, AM-basic, and AM-ridge give 1,440 measured predictions. For each source, let Δ_{main} be method minus full under completed prefill and Δ_{split} the corresponding difference under split prefill. The paired difference in differences (DiD) is $\Delta_{\text{split}} - \Delta_{\text{main}}$. Bootstrap draws preserve all four outcomes and both caps when pooled, and resample sources within tasks. Full split and full monolithic outputs are compared directly rather than assumed equal.

Table 11: Both-cap prefix sensitivity, with caps averaged within each of 48 source texts before task-stratified resampling. DiD is the paired change in compression loss between protocols, not an isolated first-token causal effect.

Model	Method	n	Primary Δ	Split Δ [CI]	DiD [CI]
qwen	SnapKV	48	+0.07	+0.83 [-0.41, +2.04]	+0.76 [+0.29, +1.28]
qwen	CALM	48	-0.62	-0.08 [-0.69, +0.55]	+0.55 [-0.13, +1.36]
qwen	AM-basic	48	-5.41	-6.43 [-11.35, -2.46]	-1.02 [-3.35, +0.95]
qwen	AM-ridge	48	-6.30	-7.00 [-11.96, -2.84]	-0.71 [-2.76, +0.77]
phi	SnapKV	48	-0.36	-0.58 [-1.42, +0.20]	-0.22 [-0.79, +0.36]
phi	CALM	48	-0.75	-0.93 [-2.06, +0.09]	-0.18 [-1.25, +0.78]
phi	AM-basic	48	-12.65	-12.50 [-16.88, -8.14]	+0.15 [-1.81, +1.98]
phi	AM-ridge	48	-5.36	-5.11 [-9.01, -1.79]	+0.25 [-0.34, +0.84]
ministral	SnapKV	48	-2.17	-1.61 [-2.73, -0.57]	+0.56 [-0.07, +1.20]
ministral	CALM	48	-1.02	-1.62 [-2.61, -0.69]	-0.60 [-1.33, -0.03]
ministral	AM-basic	48	-7.43	-5.91 [-9.76, -2.52]	+1.52 [+0.69, +2.32]
ministral	AM-ridge	48	-6.99	-6.62 [-10.64, -2.97]	+0.37 [-0.82, +1.47]

With both caps averaged, CALM-minus-split-full scores are -0.08 $[-0.69, 0.55]$, -0.93 $[-2.06, 0.09]$, and -1.62 $[-2.61, -0.69]$. The corresponding first-ID disagreements with split full are 2/96, 3/96, and 0/96 condition pairs; relative to CALM under completed prefill they are 0/96, 3/96, and 0/96. Counts over conditions are not 96 independent texts. Qwen's two disagreements arise because split full changes its first ID while CALM does not. The negative Ministral difference-in-differences does not require a changed first ID: the protocols also change cached representations and subsequent computation. This small panel is a sensitivity check, not a replacement for the 64-example task study.

K Prompt-Construction Sensitivity

The main grid uses the native chat wrapper for every Qwen SAMSum method. This section isolates the effect of that choice on a fixed 32-example subset and reports a companion analysis that excludes SAMSum.

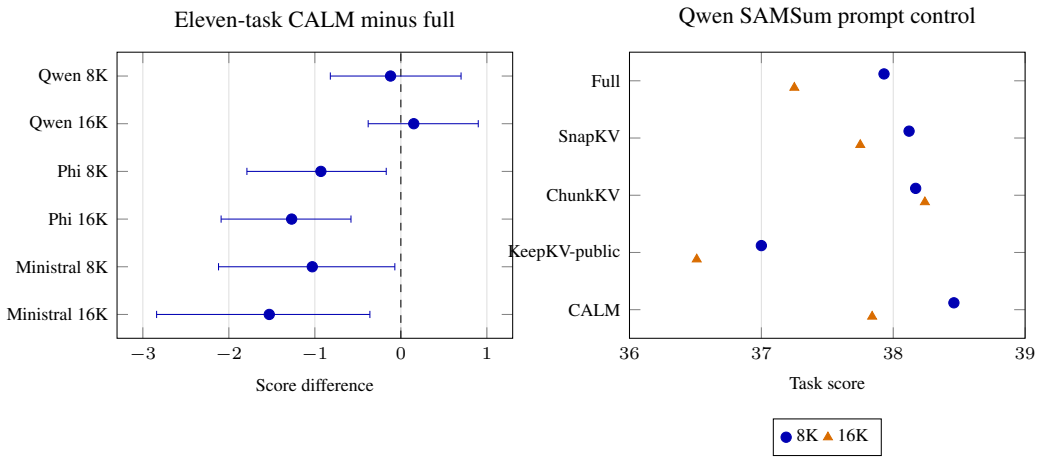

Figure 9: Prompt-construction sensitivity on the fixed 32-example subset. Left: eleven-task CALM-minus-full differences after excluding SAMSum; points show paired means and bars show 95% confidence intervals, with 352 sources per row. Right: absolute Qwen SAMSum scores under the native chat wrapper, with 32 sources per method and cap. The task mixtures and estimands differ, so neither panel replaces the main twelve-task comparison.

When SAMSum's official few-shot prompt is passed to Qwen without a chat wrapper, 30 of 32 full-cache predictions at 8K and 31 of 32 at 16K contain only a two-token thinking marker before the newline stopping rule terminates generation. The same failure counts occur for all five methods. These zero scores do not establish compression robustness on dialogue summarization, so the unwrapped condition is treated only as a prompt sensitivity.

The eleven-task sensitivity analysis retains negative CALM-minus-full intervals for Phi and Ministral and intervals containing zero for Qwen. Thus the main qualitative limitation is not removed by excluding the failed task. Exclusion is reported as a diagnostic, not as a favorable replacement benchmark.

The chat-wrapped control has no two-token thinking-marker failures. CALM-minus-full differences are $+0.53$ with interval $[-0.60, 1.73]$ at 8K and $+0.59$ with interval $[-0.85, 2.35]$ at 16K. The control confirms the prompt failure and supplies a functioning within-protocol comparison, but does not establish a reliable compression gain. Inputs retain their actual lengths: 23/32 reach 8,192 tokens and 2/32 reach 16,384 tokens; caps do not imply that every prompt has that length. No algorithm or budget was tuned on these scores.

On this 32-example subset, using the chat-wrapped SAMSum condition in the twelve-task aggregate gives CALM-minus-full differences of -0.07 $[-0.72, 0.68]$ at 8K and $+0.19$ $[-0.33, 0.88]$ at 16K. The wrapper source hash, separate run manifests, raw predictions, official re-scoring, independently recomputed tokenization hashes, and paired estimates are supplied under the `prompt_control` and `prompt_sensitivity` artifacts.

L Secondary 32-Example Comparator Study

This five-method comparison uses cohort A and the unwrapped Qwen SAMSum prompt. It is reported separately from the 64-example main estimates because it includes KeepKV-public, the disclosed centroid/vote port, and serves as a prompt-sensitivity record.

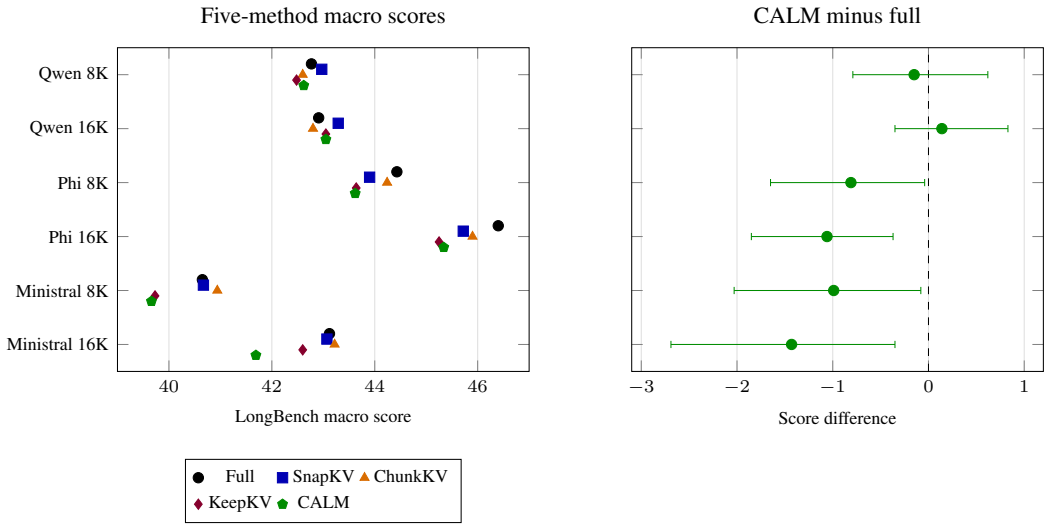

Figure 10: Secondary comparator results on cohort A with 32 examples per task and the unwrapped Qwen SAMSum prompt. Left: five-method LongBench macro scores. Right: paired CALM-minus-full means and 95% confidence intervals.

M Complete Timing Measurements

Tables 12–14 report the complete timing grid under the matched-token protocol. Means use three repetitions after warmup and 63 identical teacher-forced decode forwards. Total time includes prefill, compaction, and decoding; peak is PyTorch allocated memory in gibibytes (GiB), including weights and temporaries.

The timing example is chosen from reserved GovReport data solely because its tokenized prompt reaches 16K. Its full-model continuation supplies a fixed sequence, so differences in generated answer length do not contaminate the comparison. The initial repetition for each method and cap is excluded; subsequent orders rotate. Three repetitions on one example per model/cap do not characterize production workload variability. All raw repetitions are available. Comparisons concern this shared backend implementation, including zero bias arrays for eviction methods, and should not be interpreted as the original systems' optimized runtime claims.

Table 12: Qwen3.5-4B timing measurements under the matched-token protocol.

Cap	Method	Prefill s	Fit ms	Token/s	Decode ×	Total ×	Peak GiB
8K	Full	0.154	0.1	39.0	1.00	1.00	9.79
8K	SnapKV	0.134	4.0	39.8	1.02	0.98	9.79
8K	ChunkKV	0.122	5.9	38.1	0.98	1.00	9.79
8K	AM-basic	0.121	136.3	42.8	1.12	0.96	9.79
8K	AM-ridge	0.121	196.3	41.6	1.09	1.02	9.79
8K	CALM	0.161	22.2	38.7	0.99	1.02	9.79
16K	Full	0.252	0.1	40.8	1.00	1.00	11.04
16K	SnapKV	0.253	9.5	38.9	0.96	1.05	11.04
16K	ChunkKV	0.252	8.9	40.9	1.00	1.00	11.04
16K	AM-basic	0.252	174.5	41.3	1.02	1.08	11.04
16K	AM-ridge	0.252	220.9	38.8	0.95	1.17	11.04
16K	CALM	0.292	32.7	39.7	0.98	1.06	11.04

Table 13: Phi-4-mini timing measurements under the matched-token protocol.

Cap	Method	Prefill s	Fit ms	Token/s	Decode ×	Total ×	Peak GiB
8K	Full	0.100	0.1	77.0	1.00	1.00	8.88
8K	SnapKV	0.114	22.7	61.6	0.80	1.26	8.88
8K	ChunkKV	0.103	22.8	59.8	0.77	1.29	8.88
8K	AM-basic	0.099	930.7	57.5	0.75	2.32	8.88
8K	AM-ridge	0.197	1358.0	61.4	0.80	2.81	8.88
8K	CALM	0.137	95.8	59.9	0.78	1.40	8.88
16K	Full	0.254	0.1	79.7	1.00	1.00	10.57
16K	SnapKV	0.254	30.7	66.0	0.83	1.19	10.57
16K	ChunkKV	0.288	35.7	66.0	0.83	1.22	10.57
16K	AM-basic	0.208	1135.3	65.0	0.82	2.21	10.57
16K	AM-ridge	0.301	1444.9	65.5	0.82	2.59	10.57
16K	CALM	0.298	118.9	65.9	0.83	1.31	10.57

Table 14: Ministral-3-3B timing measurements under the matched-token protocol.

Cap	Method	Prefill s	Fit ms	Token/s	Decode ×	Total ×	Peak GiB
8K	Full	0.083	0.1	79.9	1.00	1.00	8.65
8K	SnapKV	0.145	21.6	68.2	0.85	1.25	8.65
8K	ChunkKV	0.157	34.8	68.8	0.86	1.27	8.65
8K	AM-basic	0.175	880.2	68.2	0.85	2.27	8.65
8K	AM-ridge	0.082	1302.0	68.1	0.85	2.65	8.65
8K	CALM	0.082	77.8	68.0	0.85	1.25	8.65
16K	Full	0.179	0.1	81.1	1.00	1.00	10.07
16K	SnapKV	0.222	31.2	66.8	0.82	1.25	10.07
16K	ChunkKV	0.180	35.0	66.4	0.82	1.22	10.07
16K	AM-basic	0.198	1116.2	68.2	0.84	2.34	10.07
16K	AM-ridge	0.179	1440.1	68.8	0.85	2.65	10.07
16K	CALM	0.180	115.7	69.0	0.85	1.26	10.07

N Dynamic Refresh and Audit Scope

The dynamic comparison uses 12 examples per model/cap, four from each of GovReport, QMSum, and MultiNews. Relative to the identical one-time CALM runs, mean changes and 95% intervals are: Qwen 8K -0.65 $[-1.92, 0.52]$, Qwen 16K -0.87 $[-2.09, 0.34]$; Phi 8K -0.52 $[-1.83, 0.79]$, Phi 16K -0.19 $[-1.59, 0.88]$; Ministral 8K -0.86 $[-1.71, -0.13]$, Ministral 16K -0.44 $[-0.83, 0.00]$ (the unrounded upper endpoint is positive). These small-sample intervals are descriptive. Most include zero, although every mean change is negative. The result does not justify an assumption that repeated re-anchoring is harmless.

The final audit independently recomputes every main, ablation, and dynamic score from saved predictions and public reference answers. It checks exact sample/method/cap grids, source hashes, paired input hashes, token counts, physical compaction, and cache-byte arithmetic. The timing audit checks identical input and continuation tokens, repetition grids, and positive measured durations. Native generation agrees token-for-token with the custom full-cache loop and fraction-one identity path on one reserved 8K example per family when identical EOS rules are supplied. Mathematical identity tests additionally cover GQA pooling and surrogate recompression. These checks establish the reported implementation and accounting properties; they do not establish general accuracy guarantees.