OVAL*: Output-Aware Local Page Bases for KV Cache Retrieval

Anonymous authors
Paper under double-blind review

Abstract

Long context inference with large language models becomes increasingly expensive as attention must operate over an ever growing KV cache. Page sparse attention reduces this cost by representing each KV page compactly and retrieving only a subset for each query. Existing retrieval methods are designed to estimate attention scores or page relevance, but their objectives do not directly account for how approximation errors affect the resulting value weighted attention output. We introduce OVAL, an output aware page encoding derived from the joint structure of keys and values while preserving the key information needed for accurate retrieval. OVAL is training free and requires no additional value dependent statistics at inference time. Once constructed, its stored representation has the same size and decode time scoring cost as a key only spectral representation. Across long reasoning, long context understanding, and long generation benchmarks, OVAL consistently improves over the key only spectral baseline and performs competitively with recent KV cache compression and retrieval methods. On long reasoning benchmarks, it achieves strong avg@k performance across model benchmark pairs, while matching or surpassing leading baselines on several long context understanding and generation settings with modest decoding overhead. Code is available at `https://anonymous.4open.science/r/oval-kv-4424/`.

1 Introduction

Large language models rely on self attention to retrieve information from increasingly long sequences, enabling applications from long document understanding and multi turn dialogue to code generation and extended reasoning (Vaswani et al., 2017; Bai et al., 2024; 2025; Wu et al., 2025). This flexibility comes at substantial computational cost. For a sequence of length n, standard self attention compares every query with every key, giving quadratic complexity during full sequence processing (Vaswani et al., 2017; Dao, 2024). For contexts spanning tens or hundreds of thousands of tokens, attention therefore incurs substantial storage and bandwidth costs (Fu, 2024; Bai et al., 2025).

Autoregressive inference avoids recomputing previous representations by maintaining a key value cache, or KV cache. At each layer, keys and values from previous tokens are stored and reused during generation. For a new query q, the model scores the cached keys, normalizes the logits into attention weights, and forms the output as a weighted combination of the corresponding values. Although this removes redundant computation, the KV cache grows linearly with sequence length, number of layers, batch size, and number of KV heads (Ainslie et al., 2023; Zhang et al., 2023). Each decoding step must also read and score the cached history, making long context inference increasingly dominated by KV access costs (Tang et al., 2024; Sun et al., 2025; Lee et al., 2024).

In practice, this history is not managed as an unstructured collection of tokens. Modern inference systems group consecutive KV entries into fixed size blocks, or *pages*, which form the natural unit of cache management and sparse retrieval (Kwon et al., 2023). Page retrieval

*OVAL has also been used in prior work on unrelated topics; here it refers specifically to output-aware local page bases for KV cache retrieval.

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.

methods first identify a small subset of relevant pages and then perform exact attention over the tokens they contain. To do so efficiently, they maintain a compact summary for each page and use it to estimate relevance to the current query (Tang et al., 2024; Sun et al., 2025). The effectiveness of this pipeline therefore depends critically on what information the page summary preserves.

Existing methods answer this question primarily through the keys. They estimate page importance using extrema, low dimensional key representations, geometric statistics, or spectral summaries (Tang et al., 2024; Chen et al., 2024; Sun et al., 2025; Lee et al., 2024; Hwang, 2026; Deng et al., 2026). This is natural because keys determine the attention logits. However, preserving logits is not the final objective of an attention layer. What is propagated to the rest of the network is the attention output. Let $o(q) = \sum_i \alpha_i(q) v_i$, where $\alpha(q) = \mathrm{softmax}(Kq/\sqrt{d})$. The effect of an error in the attention distribution therefore depends on the associated values. A large logit error may have little effect when attention is redistributed among tokens with similar values, while a small error along a weak key direction may substantially change the output if it shifts probability between very different values. Thus, a high variance key direction can be unimportant to the output, while a lower variance direction can be critical. Any criterion based only on key geometry cannot distinguish these cases. This motivates the central question: *what information should a page preserve to maintain the attention output?*

Our key observation is that not all attention score errors matter equally. An error matters only when it redistributes attention between tokens whose values are meaningfully different. Therefore, the importance of a key direction cannot be determined from the keys alone; it depends on how keys and values vary together within the page. Guided by this observation, we construct each page summary jointly from its keys and values, prioritizing the directions that matter most to the resulting attention output rather than simply preserving the dominant variation in key space. Output fidelity, however, is only one role of a page summary. Before a page can contribute to the output, it must first be retrieved, which requires an accurate estimate of its attention mass. We show that these two objectives have different local sensitivities: conditional attention output error is generically first order in discarded logit directions, while page log mass error is locally second order. This motivates a hybrid objective that combines output sensitivity with the conventional key covariance. The resulting basis is obtained through a small page local eigenproblem. At one endpoint, it reduces to ordinary page local key PCA; moving away from this endpoint introduces value information according to its effect on the attention output.

We call the resulting method OVAL. OVAL uses both keys and values when constructing each page basis, but stores no value statistics and requires no value dependent computation during retrieval. Values are accessed only when a completed page summary is constructed. At decode time, the summary has the same dimensionality and scoring complexity as a key only spectral page summary. Across long reasoning, long context understanding, long generation, and efficiency benchmarks (Hendrycks et al., 2021; Bai et al., 2025; Wu et al., 2025; Lee et al., 2024), the output aware component consistently improves over the key PCA endpoint on long reasoning tasks and yields strong performance against contemporary KV cache compression and retrieval methods.

Our main contributions are:

- We identify a mismatch between key driven page retrieval objectives and the attention output that ultimately matters, and derive the exact sensitivity of the conditional attention output to within page logit perturbations.
- We derive a query free output aware rank r page basis governed by centered key value coupling, and combine it with key fidelity through a single normalized parameter η, yielding a closed form spectral construction.
- With the best evaluated η, OVAL achieves the highest avg@k among compressed methods in all long reasoning settings (+3.7–3.9 points) and the best LongBench v2 Overall score on two of three models.
- We study practical selection of η using a disjoint calibration set and evaluate how well a single calibrated value transfers across tasks relative to the per setting $\eta^\star$.

2 Related Work

KV cache compression. Several complementary approaches reduce the KV cache footprint. Grouped query attention shares KV heads across queries (Ainslie et al., 2023), quantization lowers the precision of cached entries (Hooper et al., 2024), and low-rank methods such as STAR-KV project the cache onto an adaptively ranked subspace (Bhatnagar et al., 2026). Eviction methods keep only part of the history based on attention importance, recency, or prompt observations (Zhang et al., 2023; Li et al., 2024; Tang et al., 2025; Xiao et al., 2025). CriticalKV grounds eviction in attention output perturbation and shows that value states matter beyond attention weights (Feng et al., 2026). In all of these, evicted information cannot be recovered if its relevance changes later in generation.

Sparse attention and KV retrieval. Retrieval methods instead estimate which cached regions matter for the current query and attend only to the selected subset (Tang et al., 2024; Xiao et al., 2025; Chen et al., 2024; Sun et al., 2025; Liu et al., 2026). At page granularity, Quest bounds page scores with coordinatewise key minima and maxima (Tang et al., 2024), ArkVale summarizes pages with compact geometric representations (Chen et al., 2024), and UNIQUE uses key mean and dispersion statistics (Tang et al., 2024). At token granularity, several methods score keys cheaply in a reduced space. SparQ uses a subset of query components (Ribar et al., 2024), Loki uses a PCA basis of keys (Singhania et al., 2024), ShadowKV reconstructs selected entries from a low-rank key cache (Sun et al., 2025), and InfiniGen prefetches entries predicted from partial key information (Xiao et al., 2025). Other methods borrow from vector search. PQCache retrieves tokens by maximum inner product search over product-quantized keys (Zhang et al., 2025), RaBitQCache estimates attention weights with rotated binary quantization to enable adaptive top-p selection (Li et al., 2026), and ParisKV combines collision-based candidate selection with quantized inner-product reranking (Qi et al., 2026). MagicPIG argues that top-k selection degrades when attention is not sparse and instead samples keys with locality-sensitive hashing (Chen et al., 2025). Still other methods predict important tokens or blocks using learned predictors, clustering, or structured sparsity (Akhauri et al., 2026; Yang et al., 2025; Liu et al., 2025; Gao et al., 2025; Yuan et al., 2025; Lu et al., 2025; Wan et al., 2023).

Page level spectral retrieval. Closest to our setting, LOCKS builds a page-local spectral basis from keys to approximate within-page logits and estimate page attention mass (Hwang, 2026). We keep this spectral summary structure but change the basis objective. Rather than preserving directions by key energy alone, we use paired key–value geometry to prioritize the directions that matter to the attention output. This choice is consistent with evidence that value states help identify critical entries (Feng et al., 2026).

3 Method

Key PCA retains directions with the most key energy. OVAL combines key energy with the sensitivity of the conditional attention output to choose the basis. Its retrieval summary has the same format as key PCA. Throughout, keys and queries include the model's positional transformations, including RoPE. Proofs, remainder bounds, and extensions to nonisotropic queries and nonuniform reference attention are given in Appendix A.

3.1 Setup and objectives

Consider one attention head and a completed page of $m \geq 1$ key–value pairs. Let $K \in \mathbb{R}^{m\times d}$ and $V \in \mathbb{R}^{m\times d_v}$ have rows $k_i^\top$ and $v_i^\top$. With $\mathbf{1} \in \mathbb{R}^m$ the all-ones vector, define the page means and centered matrices by

$$\mu = \frac{1}{m}K^\top \mathbf{1}, \qquad X = K - \mathbf{1}\mu^\top, \qquad \nu = \frac{1}{m}V^\top \mathbf{1}, \qquad Y = V - \mathbf{1}\nu^\top.$$

Thus $\mathbf{1}^\top X = 0$ and $\mathbf{1}^\top Y = 0$. For a query $q \in \mathbb{R}^d$, set $z(q) = Xq/\sqrt{d}$ and define $\mathrm{lse}(z) = \log\big(\sum_i e^{z_i}\big)$. The log of the page's unnormalized attention mass is

$$\ell(q) = \log\left(\sum_{i=1}^{m} e^{k_i^\top q/\sqrt{d}}\right) = \frac{\mu^\top q}{\sqrt{d}} + \mathrm{lse}(z(q)).$$

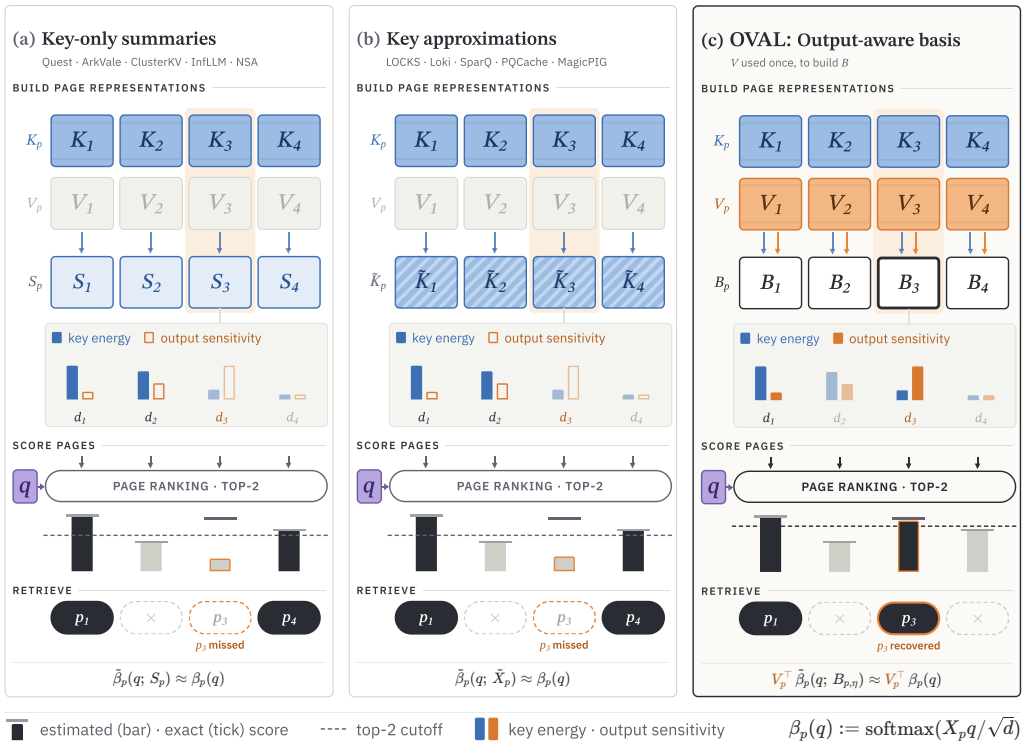


Figure 1: Overview of page representations for sparse KV retrieval. (a) Key only summaries and (b) key approximations construct retrieval representations from key information alone and therefore do not explicitly account for output sensitivity. Consequently, an output important page such as p_3 can fall below the top 2 retrieval cutoff. (c) OVAL instead uses keys and values when constructing an output aware basis, preserving directions important to both key fidelity and output sensitivity.

For a fixed query, the global attention denominator is shared by all pages, so ranking by $\ell(q)$ ranks their total attention probability. Let I_n denote the $n \times n$ identity matrix and define $\mathrm{St}(d, r) = \{B \in \mathbb{R}^{d\times r} : B^\top B = I_r\}$ for $1 \le r < d$. An orthonormal basis $B \in \mathrm{St}(d, r)$ of an r-dimensional subspace defines the projector $P_B = BB^\top$. We retain μ exactly and approximate the centered keys by XP_B, giving

$$\widehat{z}_B(q) = \frac{XP_Bq}{\sqrt{d}}, \quad e_B(q) = z(q) - \widehat{z}_B(q) = \frac{X(I_d - P_B)q}{\sqrt{d}}, \quad \widehat{\ell}_B(q) = \frac{\mu^\top q}{\sqrt{d}} + \mathrm{lse}(\widehat{z}_B(q)). \quad (1)$$

The same logit residual $e_B(q)$ drives both objectives below.

Retrieval error. Since log-sum-exp is 1-Lipschitz in the infinity norm,

$$|\ell(q) - \widehat{\ell}_B(q)| \le \|e_B(q)\|_\infty \le \|e_B(q)\|_2 .$$

Thus the expected squared logit residual bounds the expected squared page log-mass error.

Output error. The output conditional on attending within the page is

$$\beta(z) = \mathrm{softmax}(z), \qquad f(z) = V^\top \beta(z) = \nu + Y^\top \beta(z).$$

The common mean key logit cancels in this within-page normalization. Redistributing attention among identical values leaves $f(z)$ unchanged; its dependence on the logits is governed by value differences. At uniform attention, $z = 0$ and $\beta(0) = \mathbf{1}/m$. For any logit perturbation $h \in \mathbb{R}^m$,

$$\frac{d}{dt} f(th)\Big|_{t=0} = \frac{1}{m} Y^\top h.$$

Linearizing both $f(z(q))$ and $f(\widehat{z}_B(q))$ at zero therefore gives the linearized output error $Y^\top e_B(q)/m$. We measure it using the seminorm $\|a\|_G = (a^\top Ga)^{1/2}$, with $G \in \mathbb{R}^{d_v \times d_v}$

symmetric positive semidefinite. For example, $G = W_O^\top W_O$ measures the perturbation after applying the current head's output projection block W_O.

Assumption 3.1 (Query second moment)**.** The queries satisfy $\mathbb{E}_q[qq^\top] = I_d$.

Under Assumption 3.1, the two risks are

$$\mathcal{R}_{\text{key}}(B) = \mathbb{E}_q \left\| e_B(q) \right\|_2^2 = \frac{1}{d} \operatorname{tr}\big((I_d - P_B)X^\top X(I_d - P_B)\big), \tag{2}$$

$$\mathcal{R}_{\text{out}}(B) = \mathbb{E}_q \left\| \frac{1}{m} Y^\top e_B(q) \right\|_G^2 = \frac{1}{m^2 d} \operatorname{tr}((I_d - P_B)M_{\text{out}}(I_d - P_B)), \tag{3}$$

where $M_{\text{out}} = X^\top Y G Y^\top X \succeq 0$. The output risk measures the linearized conditional error at uniform attention. Both risks measure the energy of a positive semidefinite matrix outside the retained subspace, so both can be minimized by the following spectral rule.

Lemma 3.2 (Optimal subspace for a quadratic risk)**.** *Let $M \in \mathbb{R}^{d\times d}$ be symmetric positive semidefinite. The minimizers of $\operatorname{tr}((I_d - P_B)M(I_d - P_B))$ over $B \in \operatorname{St}(d, r)$ are exactly the orthonormal bases of leading r-dimensional invariant subspaces of M.*

Applied to $X^\top X$, Lemma 3.2 gives page local key PCA (Hwang, 2026). Applied to M_{out}, it gives the leading right singular subspaces of $G^{1/2}Y^\top X$, where $G^{1/2}$ is the symmetric positive semidefinite square root. The latter weights key directions by their coupling to value differences, rather than by key energy alone.

3.2 Hybrid objective

To combine the risks, scale each component by its total energy, with a stabilizer $\varepsilon > 0$:

$$c_{\text{key}} = (\operatorname{tr}(X^\top X) + \varepsilon)^{-1}, \qquad c_{\text{out}} = (\operatorname{tr}(M_{\text{out}}) + \varepsilon)^{-1}.$$

For $\eta \in [0, 1]$, define

$$\mathcal{R}_\eta(B) = (1 - \eta)c_{\text{key}}\mathcal{R}_{\text{key}}(B) + \eta m^2 c_{\text{out}}\mathcal{R}_{\text{out}}(B).$$

The factor m^2 cancels the $1/m^2$ in the output risk. Substituting (2) and (3) gives $\mathcal{R}_\eta(B) = d^{-1} \operatorname{tr}((I_d - P_B)M_\eta(I_d - P_B))$, where

$$M_\eta = (1 - \eta)c_{\text{key}}X^\top X + \eta c_{\text{out}}M_{\text{out}}. \tag{4}$$

Corollary 3.3 (Output aware local page basis)**.** *The minimizers of $\mathcal{R}_\eta(B)$ over $B \in \operatorname{St}(d, r)$ are exactly the orthonormal bases of leading r-dimensional invariant subspaces of M_η.*

This follows from Lemma 3.2 because $M_\eta \succeq 0$. The endpoint $\eta = 0$ recovers key PCA, while $\eta = 1$ minimizes the conditional output surrogate; intermediate values combine the two scaled risks. At uniform within-page attention, the linearized full-cache response to $e_B(q)$ is $p_0(q)$ times the conditional response, where $p_0(q)$ is the page's attention probability at that reference. We omit the resulting squared-error weight $p_0(q)^2$, which depends on the rest of the cache, so the basis can be constructed from (K, V) alone.

3.3 Computation and summary

For short pages ($m < d$), we reduce the eigenproblem from $d \times d$ to $m \times m$. The normalizers also use only $m \times m$ matrices, since $\operatorname{tr}(X^\top X) = \operatorname{tr}(XX^\top)$ and $\operatorname{tr}(M_{\text{out}}) = \operatorname{tr}(YGY^\top XX^\top)$. Define

$$A_\eta = (1 - \eta)c_{\text{key}}I_m + \eta c_{\text{out}}YGY^\top \succeq 0, \qquad C = A_\eta^{1/2}X,$$

using the symmetric positive semidefinite square root. Since $M_\eta = C^\top C$, its positive eigendirections can be recovered from the $m \times m$ matrix $CC^\top$. Let $s = \min\{r, \operatorname{rank}(C)\}$ and, for $s > 0$, let $U_s \in \mathbb{R}^{m\times s}$ hold orthonormal eigenvectors of $CC^\top$ for its s largest positive eigenvalues, collected in Λ_s. Then

$$CC^\top U_s = U_s\Lambda_s, \qquad B_+ = C^\top U_s\Lambda_s^{-1/2}. \tag{5}$$

Algorithm 1 OVAL: output aware page summary

Require: $K \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{m\times d_v}$, rank $r < d$, $\eta \in [0,1]$, $G \succeq 0$, $\varepsilon > 0$
$\mu \leftarrow \frac{1}{m}K^\top \mathbf{1}$, $\nu \leftarrow \frac{1}{m}V^\top \mathbf{1}$, $X \leftarrow K - \mathbf{1}\mu^\top$, $Y \leftarrow V - \mathbf{1}\nu^\top$
$T \leftarrow XX^\top$, $W \leftarrow YGY^\top$, $c_k \leftarrow (\mathrm{tr}(T)+\varepsilon)^{-1}$, $c_o \leftarrow (\mathrm{tr}(WT)+\varepsilon)^{-1}$
$A_\eta \leftarrow (1-\eta)c_k I_m + \eta c_o W$, $C \leftarrow A_\eta^{1/2}X$, $s \leftarrow \min\{r, \mathrm{rank}(C)\}$
$B_\eta \leftarrow [\,]$
if $s > 0$ **then**
  compute the s leading positive eigenpairs (U_s, Λ_s) of $CC^\top$
  $B_\eta \leftarrow C^\top U_s \Lambda_s^{-1/2}$
end if
if $s < r$ **then**
  $X_\perp \leftarrow X - (XB_\eta)B_\eta^\top$; append up to $r-s$ leading right singular directions of $X_\perp$
  fill remaining columns with orthonormal directions in $\mathrm{span}(B_\eta)^\perp$
end if
$R_\eta \leftarrow XB_\eta$
Ensure: $\mathrm{summary}(K,V) = (\mu, B_\eta, R_\eta)$

Thus $B_+^\top B_+ = I_s$ and $M_\eta B_+ = B_+\Lambda_s$. Algorithm 1 constructs B_η from B_+, handling $s = 0$ and completing the basis when $s < r$. It uses the remaining directions to minimize key reconstruction error without changing the minimum hybrid risk. The stored summary is

$$\mathrm{summary}(K,V) = (\mu, B_\eta, R_\eta), \qquad R_\eta = XB_\eta \in \mathbb{R}^{m\times r}.$$

Scoring uses (1) with $B = B_\eta$ and evaluates XP_Bq as $R_\eta(B_\eta^\top q)$. At decode time, scoring one page costs $O(d + dr + mr + m)$ operations. Values are used only during summary construction and do not enter page scoring.

4 Experiments

We evaluate four questions. First, we test whether the proposed output aware page representation improves sparse KV retrieval across long reasoning, long context understanding, and long generation tasks (Sections 4.1 and 4.2). Unless stated otherwise, we use rank $r = 8$, set $G = W_O^\top W_O$, and report $\eta^\star$, the best value in the evaluated grid at a fixed sparse KV budget. We then study selecting η from a held out calibration set (Section 4.3), sensitivity to η (Section 4.4), and end to end efficiency (Section 4.5). Implementation and reproducibility details are given in Appendix D, with additional rank and page budget ablations in Appendix B.

4.1 Long reasoning Tasks

Setup. We evaluate MATH500(Hendrycks et al., 2021), AIME24(Zhang et al., 2024a;b), and GPQA(Rein et al., 2024) on DeepSeek R1 Llama 8B, DeepSeek R1 Qwen 7B, and DeepSeek R1 Qwen 14B (Guo et al., 2025; Qwen Team, 2024; Qwen et al., 2025; Yang et al., 2024; Grattafiori et al., 2024), using $k = 8$ generations per problem. We report pass@k and avg@k, using avg@k as the primary metric because it captures correctness across all sampled reasoning trajectories.

Results. Table 1 shows the largest gains on avg@k. OVAL is the best compressed method in all nine model and benchmark settings, tying FreeKV only on Qwen 14B GPQA. Averaged across tasks, avg@k improves over FreeKV from 51.83 to 55.69 on Llama 8B, 54.14 to 57.83 on Qwen 7B, and 62.56 to 66.42 on Qwen 14B. The largest gains occur on AIME24, reaching 5.8 to 7.1 points across models. Pass@k is more mixed, with OVAL slightly below FreeKV on Llama 8B and Qwen 7B but above it on Qwen 14B. Overall, the gains primarily reflect higher accuracy across sampled trajectories rather than only a higher chance of obtaining one correct solution.

4.2 Long context understanding and generation

Setup. We next evaluate the representation under long input and long generation workloads. Following the baseline protocol, LongBench v2(Bai et al., 2025) inputs are truncated

Table 1: Long reasoning performance with $k = 8$ generations per problem. $\eta^\star$ denotes the best evaluated mixing parameter for each setting. Each Average row averages MATH500, AIME24, and GPQA for that model. Gold and silver denote the best and second best compressed methods; FullKV is excluded from ranking.

Benchmark	**Metric**	Full	Razor	RaaS	Quest	ArkVale	ShadowKV	InfiniGen	FreeKV	$\eta^\star$
DeepSeek R1 Llama 8B										
MATH500	pass@k	78.00	72.00	74.00	72.00	72.00	76.00	74.00	78.00	**80.00**
	avg@k	67.25	60.50	62.50	62.00	62.50	60.25	66.50	66.75	**69.75**
AIME24	pass@k	80.00	46.67	66.67	73.33	**80.00**	63.33	70.00	76.67	76.67
	avg@k	47.08	30.00	36.25	44.17	46.67	36.50	45.83	47.50	**53.33**
GPQA	pass@k	82.00	60.00	64.00	76.00	72.00	78.00	46.00	**86.00**	74.00
	avg@k	39.75	34.25	33.50	37.25	39.75	36.25	27.50	41.25	**44.00**
Average	pass@k	80.00	59.56	68.22	73.78	74.67	72.44	63.33	**80.22**	76.89
	avg@k	51.36	41.58	44.08	47.81	49.64	44.33	46.61	51.83	**55.69**
DeepSeek R1 Qwen 7B										
MATH500	pass@k	78.00	72.00	74.00	76.00	76.00	74.00	74.00	78.00	**80.00**
	avg@k	71.75	66.75	67.00	68.00	68.25	64.75	70.00	70.00	**73.75**
AIME24	pass@k	83.33	65.33	73.33	76.67	73.33	73.33	63.33	**83.33**	80.00
	avg@k	56.66	35.42	42.92	47.50	47.92	43.75	43.34	52.92	**60.00**
GPQA	pass@k	72.00	60.00	58.00	72.00	72.00	70.00	58.00	**74.00**	72.00
	avg@k	35.75	32.50	33.25	38.75	34.25	33.50	35.50	39.50	**39.75**
Average	pass@k	77.78	65.78	68.44	74.89	73.78	72.44	65.11	**78.44**	77.33
	avg@k	54.72	44.89	47.72	51.42	50.14	47.33	49.61	54.14	**57.83**
DeepSeek R1 Qwen 14B										
MATH500	pass@k	74.00	70.00	68.00	76.00	72.00	76.00	72.00	78.00	**80.00**
	avg@k	70.25	59.75	64.75	67.25	66.25	65.00	67.00	67.50	**72.00**
AIME24	pass@k	86.67	46.67	73.33	83.33	76.67	83.33	76.67	83.33	**86.67**
	avg@k	66.25	32.50	48.75	58.33	61.25	57.25	60.00	64.17	**71.25**
GPQA	pass@k	82.00	68.00	80.00	80.00	**86.00**	**86.00**	58.00	**86.00**	**86.00**
	avg@k	53.25	38.50	44.25	51.25	53.75	51.75	38.00	**56.00**	**56.00**
Average	pass@k	80.89	61.56	73.78	79.78	78.22	81.78	68.89	82.44	**84.22**
	avg@k	63.25	43.58	52.58	58.94	60.42	58.00	55.00	62.56	**66.42**

to 64K tokens and evaluated using Overall accuracy together with Short, Medium, and Long context buckets. LongGenBench(Wu et al., 2025) stresses the cache during 16K token generations and reports completion rate (CR) and CR×Acc. For both benchmarks, we evaluate Llama 3.1 8B Instruct and Qwen 2.5 7B and 14B Instruct (Grattafiori et al., 2024; Qwen Team, 2024; Qwen et al., 2025) with a KV budget of 2048.

Results. On LongBench v2, Table 2 shows that OVAL achieves the highest Overall score among compressed methods on Llama 8B and Qwen 7B, reaching 29.82 and 28.03, respectively, and exceeding FullKV on both. On Qwen 14B, it reaches 33.00, below the strongest compressed baseline at 34.79. Across the context length buckets, OVAL remains competitive at longer inputs, obtaining the best or second best Long score for all three models. LongGenBench presents a less uniform picture. OVAL obtains the highest CR on Qwen 7B and remains close to the strongest methods on Qwen 14B, while Llama 8B favors other retrieval methods. CR×Acc follows a similarly model dependent pattern. Thus the strongest transfer beyond reasoning appears on long input understanding, while long generation is more sensitive to the model and workload. Figure 2(b) and Appendix C additionally report InfiniteBench LongQA accuracy at contexts exceeding 100K tokens.

4.3 Selecting η in practice

The results above report $\eta^\star$, the best value in the evaluated grid for each task. Because this value is selected using evaluation scores, these results characterize the attainable performance of the representation rather than a practical parameter selection procedure.

Table 2: LongBench v2 accuracy by context length and LongGenBench generation metrics. LongBench v2 Overall is the task aggregate for each model.

Benchmark	**Metric**	Full	Razor	RaaS	Quest	ArkVale	ShadowKV	InfiniGen	FreeKV	$\eta^\star$
Llama 3.1 8B Instruct										
LongBench v2	Overall	29.22	27.44	28.23	28.43	28.63	25.45	28.56	29.22	**29.82**
	Short	34.44	33.89	33.89	33.33	33.89	32.78	32.28	**35.00**	**35.00**
	Medium	27.91	25.12	26.51	27.44	26.98	22.79	26.05	27.44	**28.84**
	Long	23.15	21.30	22.02	22.22	23.15	18.52	24.07	23.15	**25.93**
LongGenBench	CR	80.03	35.90	76.63	78.03	39.36	**79.28**	76.68	78.03	74.71
	CR×Acc	26.82	12.20	26.00	27.71	10.36	**30.66**	26.21	27.62	28.50
Qwen 2.5 7B Instruct										
LongBench v2	Overall	27.44	25.25	26.24	27.63	26.84	25.84	26.44	26.84	**28.03**
	Short	36.11	32.78	35.56	**36.67**	36.11	32.22	32.22	34.44	36.11
	Medium	23.72	21.86	21.86	22.79	22.33	20.00	**23.26**	22.33	22.79
	Long	20.37	19.44	19.44	22.22	20.37	**26.85**	23.15	23.15	25.93
LongGenBench	CR	79.56	42.13	77.65	62.89	75.91	35.49	72.96	76.93	**78.54**
	CR×Acc	31.09	21.48	**34.40**	25.96	31.79	11.43	27.67	32.81	32.38
Qwen 2.5 14B Instruct										
LongBench v2	Overall	33.40	34.19	32.60	33.80	34.19	**34.79**	32.31	34.19	33.00
	Short	41.11	**43.33**	40.56	40.00	41.11	40.56	40.56	41.11	41.11
	Medium	31.16	30.70	32.09	33.49	33.49	**34.88**	29.89	33.49	30.23
	Long	25.00	**25.93**	20.37	24.07	24.07	25.00	23.95	24.07	**25.93**
LongGenBench	CR	65.84	26.48	62.29	45.49	45.31	21.25	63.02	**65.46**	63.92
	CR×Acc	29.35	13.89	**29.79**	19.76	19.65	8.25	26.08	29.39	28.64

Table 3: Calibration based selection of η for long reasoning. $\hat{\eta}$ is selected per model on a disjoint calibration pool and frozen across tasks. All methods are scored on the same held out examples. AIME24 contains no calibration problems and is therefore identical to the evaluation set in Table 1. Gap is Sweep best minus Frozen.

Model	Task	$\hat{\eta}$	FreeKV	Frozen	Sweep best	Gap
Llama 8B	MATH500	0.50	63.25	66.00	69.74	3.74
	AIME24		47.50	53.33	53.33	0.00
	GPQA		40.50	41.75	44.17	2.42
	Avg.		**50.42**	**53.69**	**55.75**	**2.05**
Qwen 7B	MATH500	0.25	72.50	72.00	73.75	1.75
	AIME24		52.92	53.33	60.00	6.67
	GPQA		38.75	37.50	39.58	2.08
	Avg.		**54.72**	**54.28**	**57.78**	**3.50**

We therefore select one $\hat{\eta}$ per model using a disjoint calibration pool of 40 problems, with 20 from MATH500 and 20 from GPQA. We evaluate $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$ on this pool, select the value with the highest mean calibration performance, and freeze it across MATH500, AIME24, and GPQA. Calibration problems are excluded from all scores in Table 3; AIME24 contains no calibration problems, so its evaluation set is unchanged. On these matched held out examples, frozen OVAL averages 53.69 versus 50.42 for FreeKV on Llama 8B, improving on all three tasks, including AIME24, which was absent from the calibration pool. On Qwen 7B, calibration selects $\hat{\eta} = 0.25$ and frozen OVAL averages 54.28 versus 54.72 for FreeKV: it improves AIME24 slightly but trails on MATH500 and GPQA. Task specific sweeping reaches 55.75 and 57.78, leaving gaps of 2.05 and 3.50 points. Calibration therefore transfers well on Llama 8B, while on Qwen 7B the 40-problem pool does not identify the best setting; more robust selection of η is an important direction for future work.

4.4 Sensitivity to η

We next examine sensitivity to η, with $\eta = 0$ recovering page local key PCA and larger values increasing output sensitivity. Table 4 shows no universally optimal setting: long reasoning generally benefits from nonzero η, LongBench v2 (Bai et al., 2025) is relatively stable across interior values, and LongGenBench (Wu et al., 2025) is more model dependent.

Table 4: Effect of η across long reasoning, LongBench v2 (LBv2), and LongGenBench (LGB). Long reasoning results are avg@k.

	Long reasoning			LongBench v2				LongGenBench	
η	MATH500	AIME24	GPQA	Overall	Short	Medium	Long	CR	CR×Acc
Llama 8B									
0	68.00	42.08	39.00	29.42	34.44	27.44	25.00	73.12	27.24
0.25	**69.75**	47.08	39.00	29.42	**35.00**	26.98	25.00	72.64	27.26
0.5	62.25	**53.33**	40.25	**29.82**	**35.00**	27.44	**25.93**	71.82	27.23
0.75	65.75	33.75	**44.00**	**29.82**	33.89	**28.84**	25.00	72.61	27.28
1	63.75	33.75	39.00	**29.82**	**35.00**	27.91	25.00	**74.71**	**28.50**
Qwen 7B									
0	70.00	53.33	30.00	27.63	35.56	21.86	**25.93**	**78.54**	**32.38**
0.25	**73.75**	53.33	**39.75**	**28.03**	35.56	**22.79**	**25.93**	77.64	31.91
0.5	70.25	**60.00**	36.25	27.63	35.56	21.86	**25.93**	77.93	31.86
0.75	71.75	50.42	37.75	**28.03**	**36.11**	22.33	**25.93**	78.29	31.57
1	72.25	**60.00**	36.75	**28.03**	**36.11**	22.33	**25.93**	78.00	32.28
Qwen 14B									
0	71.75	63.75	51.75	32.41	40.56	29.30	25.00	63.78	28.61
0.25	68.00	**71.25**	48.75	32.21	40.56	28.84	25.00	63.74	27.68
0.5	**72.00**	66.25	**56.00**	32.60	**41.11**	28.84	**25.93**	61.08	26.94
0.75	71.50	60.42	50.25	**33.00**	**41.11**	**30.23**	25.00	**63.92**	**28.64**
1	71.00	66.25	48.25	32.60	**41.11**	29.77	24.07	61.34	27.30

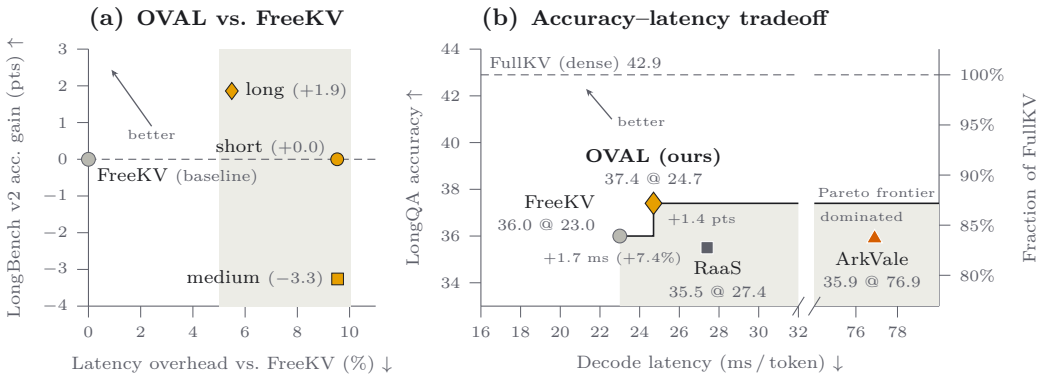


Figure 2: Efficiency of sparse retrieval. (a) Accuracy gain versus decode overhead relative to FreeKV on Qwen2.5-14B LongBench v2. (b) Broader accuracy runtime comparison on InfiniteBench LongQA with Llama-3.1-8B, including RaaS and ArkVale.

4.5 Efficiency

Accuracy Runtime Tradeoff. Figure 2(a) shows that OVAL incurs about 10% decode overhead over FreeKV on LongBench v2 while improving accuracy most on long inputs. Panel (b) extends the comparison to RaaS and ArkVale on InfiniteBench LongQA, where OVAL achieves higher accuracy at modest latency and is substantially faster than ArkVale.

5 Conclusion

We introduced OVAL, an output aware page representation for sparse KV retrieval that combines key fidelity with value dependent output sensitivity. It preserves the summary format and decode time scoring complexity of key only spectral retrieval while better aligning retrieval with the attention output. Across long reasoning, long context understanding, and long generation, OVAL improves over the key only endpoint and performs strongly against recent sparse KV baselines with modest decoding overhead. We further show that η can be selected from held out calibration data, which transfers well on Llama 8B and leaves substantial headroom on Qwen 7B. Overall, the results support incorporating output sensitivity into compact retrieval representations and designing sparse attention summaries around downstream approximation error rather than key geometry alone.

AI use statement

Generative AI tools were used for limited assistance with writing, editing, and code debugging. All technical ideas, methodological development, experiments, analyses, and conclusions were produced and verified by the authors.

Ethics statement

This work studies efficient long context inference for large language models and does not involve human subjects, private user data, or deployment in high risk settings. The proposed method may reduce attention access and computational costs of long context inference, but inherits the broader risks and limitations of the underlying language models and datasets on which it is used.

Reproducibility statement

We provide implementation details, model checkpoints, benchmark protocols, hyperparameters, evaluation settings, memory and latency configurations, and corrections to the reference harness in Appendix D. Code, patches, and per experiment scripts are released through the anonymized repository linked there. All reported comparisons use matched effective KV budgets, and the appendix documents deviations from upstream defaults and implementation details required to reproduce the reported results.

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 Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/.

Yash Akhauri, Ahmed F AbouElhamayed, Yifei Gao, Chi-Chih Chang, Sameh Gobriel, Nilesh Jain, and Mohamed S. Abdelfattah. Tokenbutler: Token importance is predictable, 2026. URL https://arxiv.org/abs/2503.07518.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Priyansh Bhatnagar, Ashkan Moradifirouzabadi, Se-Hyun Yang, SeungJae Lee, Jungwook Choi, and Mingu Kang. STAR-KV: Low-rank KV cache compression via soft thresholding for adaptive rank control. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=lJjH1q6RwY.

Renze Chen, Zhuofeng Wang, Beiquan Cao, Tong Wu, Size Zheng, Xiuhong Li, Xuechao Wei, Shengen Yan, Meng Li, and Yun Liang. Arkvale: Efficient generative LLM inference with recallable key-value eviction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4oAt5L4lYe.

Zhuoming Chen, Ranajoy Sadhukhan, Zihao Ye, Yang Zhou, Jianyu Zhang, Niklas Nolte, Yuandong Tian, Matthijs Douze, Leon Bottou, Zhihao Jia, and Beidi Chen. MagicPIG: LSH sampling for efficient LLM generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ALzTQUgW8a.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

Keqi Deng, Shaoshi Ling, Ruchao Fan, and Jinyu Li. Unique: Universal top-k sparse attention for training-free inference and sparsity-aware training, 2026. URL https://arxiv.org/abs/2605.27740.

Yuan Feng, Junlin Lv, Haoyu Guo, Yukun Cao, S Kevin Zhou, and Xike Xie. CriticalKV: Optimizing KV cache eviction from an output perturbation perspective. In *Forty-third International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=vMwlcLyI4b.

Yao Fu. Challenges in deploying long-context transformers: A theoretical peak performance analysis, 2024. URL https://arxiv.org/abs/2405.08944.

Yizhao Gao, Zhichen Zeng, Dayou Du, Shijie Cao, Peiyuan Zhou, Jiaxing Qi, Junjie Lai, Hayden Kwok-Hay So, Ting Cao, Fan Yang, and Mao Yang. Seerattention: Learning intrinsic sparse attention in your LLMs, 2025. URL https://arxiv.org/abs/2410.13276.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur

Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta,

Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, Sep 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

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, 2024.

Junsung Hwang. Locks: Page-local compact key summaries for efficient long-context decoding, 2026. URL https://arxiv.org/abs/2607.24555.

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

Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. InfiniGen: Efficient generative inference of large language models with dynamic KV cache management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 155–172, Santa Clara, CA, July 2024. USENIX Association. ISBN 978-1-939133-40-3. URL https://www.usenix.org/conference/osdi24/presentation/lee.

Wenhao Li, Jinhao Dong, Hailin Zhang, Wenhang Shi, Wei Lu, and Xiaoyong Du. RabitQCache: Rotated binary quantization for KVCache in long context LLM inference. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=zc8ZDMJmKB.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 22947–22970. Curran Associates, Inc., 2024. doi: 10.52202/079017-0722. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/28ab418242603e0f7323e54185d19bde-Paper-Conference.pdf.

Guangda Liu, Chengwei Li, Jieru Zhao, Chenqi Zhang, and Minyi Guo. ClusterKV: Manipulating LLM KV cache in semantic space for recallable compression. In *Proceedings of the 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7, San Francisco, CA, USA, 2025. IEEE.

Guangda Liu, Chengwei Li, Zhenyu Ning, Jing Lin, Yiwu Yao, Danning Ke, Minyi Guo, and Jieru Zhao. Freekv: Boosting kv cache retrieval for efficient llm inference, 2026. URL https://arxiv.org/abs/2505.13109.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, Zhiqi Huang, Huan Yuan, Suting Xu, Xinran Xu, Guokun Lai, Yanru Chen, Huabin Zheng, Junjie Yan, Jianlin Su, Yuxin Wu, Yutao Zhang, Zhilin Yang, Xinyu Zhou, Mingxing Zhang, and Jiezhong Qiu. Moba: Mixture of block attention for long-context llms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 17790–17815. Curran Associates, Inc., 2025. doi: 10.52202/085713-0602. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19eae75beed66321d62272e794a9c2ac-Paper-Conference.pdf.

Yanlin Qi, Xinhang Chen, Huiqiang Jiang, Qitong Wang, Botao Peng, and Themis Palpanas. ParisKV: Fast and drift-robust KV-cache retrieval for long-context LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=wxD4wTYQXt.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Luka Ribar, Ivan Chelombiev, Luke Hudlass-Galley, Charlie Blake, Carlo Luschi, and Douglas Orr. SparQ attention: Bandwidth-efficient LLM inference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42558–42583. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ribar24a.html.

Prajwal Singhania, Siddharth Singh, Shwai He, Soheil Feizi, and Abhinav Bhatele. Loki: Low-rank keys for efficient sparse attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=raABeiV71j.

Hanshi Sun, Li-Wen Chang, Wenlei Bao, Size Zheng, Ningxin Zheng, Xin Liu, Harry Dong, Yuejie Chi, and Beidi Chen. ShadowKV: KV cache in shadows for high-throughput long-context LLM inference. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57355–57373. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/sun25b.html.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Danning Ke, Shikuan Hong, Yiwu Yao, and Gongyi Wang. Razorattention: Efficient KV cache compression through retrieval heads. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tkiZQlL04w.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-aware sparsity for efficient long-context LLM inference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47901–47911. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tang24l.html.

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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Qiang Wan, Zilong Huang, Jiachen Lu, Gang Yu, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-qg8MQNrxZw.

Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. Longgenbench: Benchmarking long-form generation in long context llms. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 6851–6872, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/141304a37d59ec7f116f3535f1b74bde-Paper-Conference.pdf.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10819.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Qingyue Yang, Jie Wang, Xing Li, Zhihai Wang, Chen Chen, Lei Chen, Xianzhi Yu, Wulong Liu, Jianye Hao, Mingxuan Yuan, and Bin Li. Attentionpredictor: Temporal patterns matter for KV cache compression. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23097, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1126. URL https://aclanthology.org/2025.acl-long.1126/.

Di Zhang, Jiatong Li, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. Accessing GPT-4 level mathematical olympiad solutions via monte carlo tree self-refine with LLaMa-3 8B. *arXiv preprint arXiv:2406.07394*, 2024a.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. LLaMA-Berry: Pairwise optimization for O1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*, 2024b.

Hailin Zhang, Xiaodong Ji, Yilin Chen, Fangcheng Fu, Xupeng Miao, Xiaonan Nie, Weipeng Chen, and Bin Cui. PQCache: Product quantization-based KVCache for long context LLM inference. *Proc. ACM Manag. Data*, 3(3), June 2025. doi: 10.1145/3725338. URL https://doi.org/10.1145/3725338.

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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710. Curran Associates, Inc., 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6ceefa7b15572587b78ecfcebb2827f8-Paper-Conference.pdf.

A Proofs and extensions

Appendix A.1 collects the softmax and spectral facts used throughout. Appendix A.2 proves the output sensitivity results in Section 3.1. Appendix A.3 proves the first order output expansion and gives additional results separating key fidelity from output sensitivity. Appendix A.4 derives the page mass surrogate and proves Corollary 3.3. Appendix A.5 gives extensions beyond the isotropic query and uniform attention assumptions.

A.1 Preliminaries

Throughout, $\beta(z) = \mathrm{softmax}(z)$ and $f(z) = V^\top \beta(z)$. For a probability vector $b \in \mathbb{R}^m$, define $S(b) = \mathrm{Diag}(b) - bb^\top$, where $\mathrm{Diag}(b)$ places b on the diagonal. Let $H = I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$, so $X = HK$ and $Y = HV$. We write $Df(z)[h] = \frac{d}{dt} f(z+th)\big|_{t=0}$ for the directional derivative. All functions below are smooth on $\mathbb{R}^m$.

Lemma A.1 (Softmax calculus)**.** *For every $z, h \in \mathbb{R}^m$:*

1. *$D\beta(z)[h] = S(\beta(z))h$;*
2. *$S(\beta(z))\mathbf{1} = 0$ and $S(\beta(z)) \succeq 0$;*
3. *$D \operatorname{lse}(z) = \beta(z)^\top$ and $D^2 \operatorname{lse}(z) = S(\beta(z))$.*

In particular, $D \operatorname{lse}(0) = \mathbf{1}^\top/m$ and $D^2 \operatorname{lse}(0) = H/m$.

Proof. For $\beta_i(z) = e^{z_i}/\sum_j e^{z_j}$,

$$\frac{\partial \beta_i}{\partial z_j} = \beta_i(\delta_{ij} - \beta_j),$$

which gives the first claim.

For the second,

$$S(\beta)\mathbf{1} = \beta - \beta(\mathbf{1}^\top \beta) = 0.$$

Moreover, for any $x \in \mathbb{R}^m$,

$$x^\top S(\beta) x = \sum_i \beta_i x_i^2 - \left(\sum_i \beta_i x_i\right)^2 = \mathrm{Var}_{i\sim\beta}(x_i) \geq 0.$$

Finally, $\nabla \operatorname{lse}(z) = \beta(z)$, so $D^2 \operatorname{lse}(z) = S(\beta(z))$. At $z = 0$, $\beta(0) = \mathbf{1}/m$, hence $S(\mathbf{1}/m) = H/m$. □

Proof of Lemma 3.2. Let $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ be the eigenvalues of M. Since $I_d - P_B$ is an orthogonal projector,

$$\mathrm{tr}((I_d - P_B)M(I_d - P_B)) = \mathrm{tr}(M) - \mathrm{tr}(B^\top M B).$$

Thus minimizing the residual is equivalent to maximizing $\mathrm{tr}(B^\top M B)$.

Write $M = \sum_j \lambda_j u_j u_j^\top$ and define $c_j = \left\|B^\top u_j\right\|_2^2$. Then $0 \leq c_j \leq 1$ and $\sum_j c_j = r$, so

$$\mathrm{tr}(B^\top M B) = \sum_j \lambda_j c_j \leq \sum_{j \leq r} \lambda_j.$$

Equality is attained by $B = [u_1, \ldots, u_r]$, giving minimum residual $\sum_{j>r} \lambda_j$.

Equality holds exactly when $c_j = 1$ for $\lambda_j > \lambda_r$ and $c_j = 0$ for $\lambda_j < \lambda_r$. Thus the retained subspace contains all eigendirections above λ_r, with its remaining dimensions in the λ_r eigenspace. These are precisely the leading invariant subspaces; the subspace is unique when $\lambda_r > \lambda_{r+1}$.

Finally, if $M = C^\top C$, its positive eigenvectors are the right singular vectors of C. □

A.2 Output sensitivity

Lemma A.2 (Conditional output sensitivity)**.** *For any logits $z \in \mathbb{R}^m$ and perturbation $h \in \mathbb{R}^m$,*

$$Df(z)[h] = V^\top S(\beta(z))h = \sum_{i=1}^{m} \beta_i(z)\big(v_i - f(z)\big)h_i. \tag{6}$$

Moreover, $V^\top S(\beta) = Y^\top S(\beta)$, since $S(\beta)$ is symmetric and $S(\beta)\mathbf{1} = 0$.

Proof of Lemma A.2. By Lemma A.1,

$$Df(z)[h] = V^\top D\beta(z)[h] = V^\top S(\beta(z))h.$$

Writing $\beta = \beta(z)$,

$$\begin{aligned} V^\top S(\beta)h &= V^\top \operatorname{Diag}(\beta)h - V^\top \beta\, \beta^\top h \\ &= \sum_{i=1}^{m} \beta_i v_i h_i - f(z) \sum_{i=1}^{m} \beta_i h_i \\ &= \sum_{i=1}^{m} \beta_i\big(v_i - f(z)\big)h_i. \end{aligned}$$

Since $S(\beta)\mathbf{1} = 0$ and $S(\beta)$ is symmetric, $\mathbf{1}^\top S(\beta) = 0$. Using $V = \mathbf{1}\nu^\top + Y$,

$$V^\top S(\beta) = \nu\, \mathbf{1}^\top S(\beta) + Y^\top S(\beta) = Y^\top S(\beta).$$

□

Proposition A.3 (Exact path identity)**.** *Let $z, \widehat{z} \in \mathbb{R}^m$, $e = z - \widehat{z}$, $z_t = \widehat{z} + te$, and $\beta_t = \operatorname{softmax}(z_t)$. Then*

$$f(z) - f(\widehat{z}) = \left(\int_0^1 V^\top S(\beta_t)\, dt\right) e, \tag{7}$$

and

$$\|f(z) - f(\widehat{z})\|_G^2 \le \int_0^1 e^\top S(\beta_t) V G V^\top S(\beta_t) e\, dt.$$

Proof. Let $g(t) = f(z_t)$. By Lemma A.2,

$$g'(t) = Df(z_t)[e] = V^\top S(\beta_t)e.$$

The fundamental theorem of calculus gives (7).

For the second claim, define $a_t = G^{1/2}V^\top S(\beta_t)e$. Jensen's inequality yields

$$\left\| \int_0^1 a_t\, dt \right\|_2^2 \le \int_0^1 \|a_t\|_2^2\, dt,$$

and

$$\|a_t\|_2^2 = e^\top S(\beta_t) V G V^\top S(\beta_t) e.$$

□

Proposition A.4 (Deterministic value diameter bound)**.** *Define*

$$\operatorname{diam}_G(V) = \max_{i,j} \|v_i - v_j\|_G .$$

Then for all $z, \widehat{z} \in \mathbb{R}^m$,

$$\|f(z) - f(\widehat{z})\|_G \le \operatorname{diam}_G(V)\, \|z - \widehat{z}\|_\infty . \tag{8}$$

Proof. For any probability vector b, $V^\top b$ lies in the convex hull of $\{v_1, \dots, v_m\}$. Hence

$$\left\|v_i - V^\top b\right\|_G \le \operatorname{diam}_G(V).$$

Using Lemma A.2,

$$\begin{aligned}\|Df(z)[h]\|_G &\le \sum_{i=1}^m \beta_i(z)\, \|v_i - f(z)\|_G\, |h_i| \\ &\le \operatorname{diam}_G(V)\, \|h\|_\infty .\end{aligned}$$

Integrating along the segment between $\widehat{z}$ and z proves (8). □

Proposition A.5 (Global sensitivity of centered page perturbations)**.** *Fix the query, the page values, and the background cache. Let the background contribute denominator $Z_0 > 0$ and value numerator $n_0 \in \mathbb{R}^{d_v}$. Hold the mean key factor $c = \exp(\mu^\top q/\sqrt{d}) > 0$ fixed, and define the full output and page attention probability by*

$$o_{\mathrm{glob}}(z) = \frac{n_0 + cV^\top \exp(z)}{Z_0 + c\mathbf{1}^\top \exp(z)}, \qquad p(z) = \frac{c\mathbf{1}^\top \exp(z)}{Z_0 + c\mathbf{1}^\top \exp(z)},$$

where exp *acts componentwise. For any $z, h \in \mathbb{R}^m$,*

$$Do_{\mathrm{glob}}(z)[h] = p(z)Df(z)[h] + p(z)\big(f(z) - o_{\mathrm{glob}}(z)\big)\beta(z)^\top h,$$

and

$$Dp(z)[h] = p(z)(1 - p(z))\beta(z)^\top h.$$

Let $T_0 = Z_0 + cm$ and $p_0 = p(0) = cm/T_0$. For every centered perturbation h with $\mathbf{1}^\top h = 0$, $Dp(0)[h] = 0$ and

$$Do_{\mathrm{glob}}(0)[h] = p_0 Df(0)[h] = \frac{p_0}{m} Y^\top h = \frac{c}{T_0} Y^\top h. \tag{9}$$

Proof. Write $o_{\mathrm{glob}}(z) = N(z)/T(z)$ with $N(z) = n_0 + cV^\top \exp(z)$ and $T(z) = Z_0 + c\mathbf{1}^\top \exp(z)$. Their derivatives are

$$DN(z)[h] = cV^\top \operatorname{Diag}(\exp(z))h, \qquad DT(z)[h] = c\exp(z)^\top h.$$

Since $c\exp(z)/T(z) = p(z)\beta(z)$, the quotient rule gives

$$\begin{aligned}Do_{\mathrm{glob}}(z)[h] &= p(z)\left(V^\top \operatorname{Diag}(\beta(z))h - o_{\mathrm{glob}}(z)\beta(z)^\top h\right) \\ &= p(z)Df(z)[h] + p(z)\big(f(z) - o_{\mathrm{glob}}(z)\big)\beta(z)^\top h.\end{aligned}$$

The second equality uses Lemma A.2 and $f(z) = V^\top \beta(z)$. Applying the quotient rule to $p(z) = (T(z) - Z_0)/T(z)$ also yields

$$Dp(z)[h] = \frac{Z_0}{T(z)} \frac{DT(z)[h]}{T(z)} = (1 - p(z))p(z)\beta(z)^\top h.$$

At uniform attention, $\beta(0) = \mathbf{1}/m$, so a centered h satisfies $\beta(0)^\top h = 0$. Hence the page mass derivative and the additional term in $Do_{\mathrm{glob}}(0)[h]$ both vanish. Finally, $S(\beta(0)) = H/m$ and $V^\top H = Y^\top$, so $Df(0)[h] = Y^\top h/m$. Substituting $p_0 = cm/T_0$ proves (9). □

A.3 Output aware page basis

Proposition A.6 (First order conditional output error)**.** *For any $q \in \mathbb{R}^d$ and $B \in \mathrm{St}(d, r)$,*

$$f(z(q)) - f(\widehat{z}_B(q)) = \frac{1}{m\sqrt{d}} Y^\top X(I_d - P_B)q + r_B(q). \tag{10}$$

Let $\rho = \max\{\|z(q)\|_2, \|\widehat{z}_B(q)\|_2\}$ and define

$$L_{V,G}(\rho) = \sup_{\substack{u,a,b \in \mathbb{R}^m \\ \|u\|_2 \le \rho,\ \|a\|_2 = \|b\|_2 = 1}} \left\|D^2 f(u)[a,b]\right\|_G,$$

where $D^2 f(u)[a,b]$ is the second derivative applied to directions a, b. This quantity is finite, and the remainder satisfies

$$\|r_B(q)\|_G \le L_{V,G}(\rho)\rho\, \|e_B(q)\|_2 .$$

Proof of Proposition A.6. Fix q and B, and write $z = z(q)$, $\widehat{z} = \widehat{z}_B(q)$, $e = e_B(q) = z - \widehat{z}$, and $\rho = \max\{\|z\|_2, \|\widehat{z}\|_2\}$. The quantity $L_{V,G}(\rho)$ is finite because $D^2 f$ is continuous on the compact product of the closed ball and the two unit spheres in its definition. At uniform attention, $S(\beta(0)) = H/m$, so

$$Df(0)[e] = \frac{1}{m} V^\top H e = \frac{1}{m} Y^\top e = \frac{1}{m\sqrt{d}} Y^\top X(I_d - P_B)q.$$

Let $z_t = \widehat{z} + te$ for $0 \le t \le 1$. The exact path identity gives

$$f(z) - f(\widehat{z}) = Df(0)[e] + r_B(q), \qquad r_B(q) = \int_0^1 \big(Df(z_t) - Df(0)\big)[e]\, dt.$$

This defines the remainder as a vector in $\mathbb{R}^{d_v}$ and establishes (10) independently of whether G is invertible. Since $\|z_t\|_2 \le \rho$,

$$\big(Df(z_t) - Df(0)\big)[e] = \int_0^1 D^2 f(sz_t)[z_t, e]\, ds.$$

Taking the G-seminorm, applying the triangle inequality, and using the definition of $L_{V,G}(\rho)$ yields

$$\|r_B(q)\|_G \le L_{V,G}(\rho)\rho \, \|e_B(q)\|_2 . \tag{11}$$

The triangle inequality remains valid for singular G, since $\|a\|_G = \left\|G^{1/2}a\right\|_2$.

For completeness, Taylor expansion at zero also gives

$$f(a) = f(0) + Df(0)[a] + R(a), \qquad R(a) = \int_0^1 (1-t) D^2 f(ta)[a, a]\, dt.$$

For $\|a\|_2 \le \rho$, $\|R(a)\|_G \le L_{V,G}(\rho) \, \|a\|_2^2 /2$. Since $r_B(q) = R(z) - R(\widehat{z})$, this implies

$$\|r_B(q)\|_G \le \frac{1}{2} L_{V,G}(\rho) \left(\|z(q)\|_2^2 + \|\widehat{z}_B(q)\|_2^2 \right). \tag{12}$$

□

Derivation and local meaning of the output surrogate. The expected squared linearized conditional error is

$$\begin{aligned}
\mathcal{R}_{\text{out}}(B) &= \mathbb{E}_q \left\| \frac{Y^\top X(I_d - P_B)q}{m\sqrt{d}} \right\|_G^2 \\
&= \frac{1}{m^2 d} \mathbb{E}_q \left[q^\top (I_d - P_B) M_{\text{out}} (I_d - P_B) q \right] \\
&= \frac{1}{m^2 d} \operatorname{tr}\big((I_d - P_B) M_{\text{out}} (I_d - P_B)\big),
\end{aligned}$$

where the last equality uses $\mathbb{E}_q[qq^\top] = I_d$ and $M_{\text{out}} = X^\top Y G Y^\top X$. This proves (3). The Gram factorization $M_{\text{out}} = (G^{1/2} Y^\top X)^\top (G^{1/2} Y^\top X)$ shows it is positive semidefinite. Lemma 3.2 therefore gives the stated eigenspace and singular subspace minimizers, including the freedom to choose within tied eigenspaces. If its rank is below r, all positive eigendirections must be retained and the remaining directions can be chosen orthonormally in its nullspace. This includes $M_{\text{out}} = 0$, for which every $B \in \mathrm{St}(d, r)$ has zero surrogate risk.

To relate the surrogate to the exact conditional error, scale queries by $t > 0$ while keeping X, V, G, and B fixed. Differentiability at zero gives

$$\frac{f(z(tq)) - f(\widehat{z}_B(tq))}{t} \longrightarrow \frac{1}{m} Y^\top e_B(q) \qquad \text{as } t \to 0.$$

By Proposition A.4 and $e_B(tq) = te_B(q)$,

$$\left\| \frac{f(z(tq)) - f(\widehat{z}_B(tq))}{t} \right\|_G^2 \le \operatorname{diam}_G(V)^2 \, \|e_B(q)\|_2^2 .$$

The right side is integrable under Assumption 3.1. Dominated convergence therefore proves

$$\mathbb{E}_q \left\| f(z(tq)) - f(\widehat{z}_B(tq)) \right\|_G^2 = t^2 \mathcal{R}_{\text{out}}(B) + o(t^2) \qquad \text{as } t \to 0.$$

Thus the surrogate is the leading coefficient of the expected squared conditional error in this local limit. No higher query moment is needed for this statement, and it does not assert optimality for finite-logit nonlinear error.

The full cache derivative gives a different weighting when averaged over queries. For each query, fix its background and mean key factor as in Proposition A.5, and let $p_0(q)$ be the page probability at uniform within page attention. Applied to $e_B(q)$, the expected squared first order full cache response is

$$\frac{1}{m^2 d} \mathbb{E}_q \left[p_0(q)^2 q^\top (I_d - P_B) M_{\text{out}} (I_d - P_B) q \right].$$

The condition $\mathbb{E}_q[qq^\top] = I_d$ alone does not remove the additional query-dependent weight $p_0(q)^2$. The objective in (3) deliberately uses the conditional response, allowing construction from the page alone.

Why output sensitivity differs from key fidelity.

Proposition A.7 (Marginal value covariance does not identify output sensitivity)**.** *For every token permutation matrix* Π *and* $Y' = \Pi Y$*,*

$$(Y')^\top Y' = Y^\top Y.$$

However, there exist X, Y, Π*, and* G *such that*

$$X^\top Y' G (Y')^\top X \neq X^\top Y G Y^\top X.$$

Consequently, the marginal value covariance $Y^\top Y$ *does not determine* M_{out}*.*

Proof. Since every permutation matrix is orthogonal,

$$(Y')^\top Y' = Y^\top \Pi^\top \Pi Y = Y^\top Y.$$

For a separation example, take $m = 3$, $d = d_v = 1$, $G = 1$, and

$$X = Y = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

Let Π exchange the first two tokens, so

$$Y' = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}.$$

Then

$$X^\top Y Y^\top X = (X^\top Y)^2 = 4, \qquad X^\top Y' (Y')^\top X = (X^\top Y')^2 = 1.$$

Since X is unchanged, the pages have identical $X^\top X$ and $Y^\top Y$ but different output sensitivity matrices. □

Proposition A.8 (Key PCA versus the output aware basis)**.** *Let* $m \geq 3$*,* $d = 2$*,* $d_v = 1$*, and choose orthonormal* $a, b \in \mathbf{1}^\perp$*. Set*

$$X = [\sigma_1 a \quad \sigma_2 b], \qquad Y = b, \qquad G = 1,$$

with $\sigma_1 > \sigma_2 > 0$*. For rank* $r = 1$*, key PCA uniquely selects the first feature direction, whereas the output aware basis uniquely selects the second. The leading output risk of the latter is zero, while that of key PCA is* $\sigma_2^2/(m^2 d)$*.*

Proof. By orthonormality,

$$X^\top X = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}, \qquad Y^\top X = \begin{pmatrix} 0 & \sigma_2 \end{pmatrix}.$$

Therefore

$$M_{\text{out}} = X^\top Y Y^\top X = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Since $\sigma_1^2 > \sigma_2^2$, key PCA selects the first direction, whereas Lemma 3.2 gives the second direction for the output objective.

Using (3), the second direction discards only the zero eigenvalue and therefore incurs zero leading output risk. The first direction discards σ_2^2, yielding $\sigma_2^2/(m^2 d)$. □

A.4 Page mass fidelity and the hybrid basis

Lemma A.9 (Log sum exponential expansion with explicit remainder)**.** *For every $h \in \mathbf{1}^\perp$,*

$$\left| \operatorname{lse}(h) - \log m - \frac{1}{2m} \|h\|_2^2 \right| \le \frac{1}{3} \|h\|_\infty \|h\|_2^2 . \tag{13}$$

Proof. Let $g(t) = \operatorname{lse}(th)$ and $\beta^{(t)} = \operatorname{softmax}(th)$. By Lemma A.1,

$$g'(t) = \beta^{(t)\top} h, \qquad g''(t) = h^\top S(\beta^{(t)}) h.$$

Writing $m_t = \beta^{(t)\top} h$, differentiating once more gives

$$g'''(t) = \sum_i \beta_i^{(t)} (h_i - m_t)^3.$$

Therefore

$$\begin{aligned} |g'''(t)| &\le \max_i |h_i - m_t| \sum_i \beta_i^{(t)} (h_i - m_t)^2 \\ &\le 2 \|h\|_\infty \sum_i \beta_i^{(t)} h_i^2 \\ &\le 2 \|h\|_\infty \|h\|_2^2 . \end{aligned}$$

Since $h \in \mathbf{1}^\perp$,

$$g(0) = \log m, \qquad g'(0) = 0, \qquad g''(0) = \frac{1}{m} \|h\|_2^2 .$$

Taylor's theorem gives

$$\operatorname{lse}(h) = \log m + \frac{1}{2m} \|h\|_2^2 + \frac{1}{6} g'''(\tau)$$

for some $\tau \in (0, 1)$. Applying the bound above proves (13). □

Derivation of the key fidelity surrogate. Applying Lemma A.9 to the centered vectors $z(q)$ and $\widehat{z}_B(q)$, and canceling the preserved mean key term, gives

$$\ell(q) - \widehat{\ell}_B(q) = \frac{\|z(q)\|_2^2 - \|\widehat{z}_B(q)\|_2^2}{2m} + \Delta_B(q),$$

where

$$|\Delta_B(q)| \le \frac{1}{3} \left(\|z(q)\|_\infty \|z(q)\|_2^2 + \|\widehat{z}_B(q)\|_\infty \|\widehat{z}_B(q)\|_2^2 \right).$$

Since $z(q) = \widehat{z}_B(q) + e_B(q)$,

$$\|z(q)\|_2^2 - \|\widehat{z}_B(q)\|_2^2 = \|e_B(q)\|_2^2 + 2\widehat{z}_B(q)^\top e_B(q).$$

Although $P_B q$ and $(I_d - P_B)q$ are orthogonal, multiplication by X need not preserve their orthogonality. Under $\mathbb{E}_q[qq^\top] = I_d$, however, cyclicity of the trace and $(I_d - P_B)P_B = 0$ give

$$\begin{aligned}\mathbb{E}_q[\widehat{z}_B(q)^\top e_B(q)] &= \frac{1}{d}\operatorname{tr}\big(P_B X^\top X(I_d - P_B)\big) \\ &= \frac{1}{d}\operatorname{tr}\big((I_d - P_B)P_B X^\top X\big) = 0.\end{aligned}$$

Moreover, the same second moment assumption yields the exact identity

$$\mathcal{R}_{\text{key}}(B) = \mathbb{E}_q \left\|e_B(q)\right\|_2^2 = \frac{1}{d}\operatorname{tr}\big((I_d - P_B)X^\top X(I_d - P_B)\big)\,, \tag{14}$$

which is (2) and holds without a small-logit assumption.

There is also a global bound on squared log-mass error. If $a = \left\|e_B(q)\right\|_\infty$, then componentwise $\widehat{z}_B(q) - a\mathbf{1} \le z(q) \le \widehat{z}_B(q) + a\mathbf{1}$. Monotonicity and the identity $\operatorname{lse}(u + c\mathbf{1}) = \operatorname{lse}(u) + c$ imply

$$|\ell(q) - \widehat{\ell}_B(q)| \le \left\|e_B(q)\right\|_\infty \le \left\|e_B(q)\right\|_2.$$

Squaring and taking expectations therefore gives

$$\mathbb{E}_q[(\ell(q) - \widehat{\ell}_B(q))^2] \le \mathcal{R}_{\text{key}}(B).$$

This bound requires only the finite second moment, with no restriction on the logit magnitudes.

To state the local approximation in expectation, additionally assume $\mathbb{E}_q \left\|q\right\|_2^3 < \infty$. For fixed X and B, the remainder bound is integrable, and

$$\mathbb{E}_q[\ell(q) - \widehat{\ell}_B(q)] = \frac{1}{2m}\mathcal{R}_{\text{key}}(B) + \mathbb{E}_q \Delta_B(q).$$

In particular, scaling queries by $t \ge 0$ gives the precise local statement

$$\mathbb{E}_q[\ell(tq) - \widehat{\ell}_B(tq)] = \frac{t^2}{2m}\mathcal{R}_{\text{key}}(B) + O(t^3) \qquad \text{as } t \to 0,$$

because both centered logit vectors scale linearly with t and the remainder bound scales cubically. Here q retains the assumed second moment $\mathbb{E}_q[qq^\top] = I_d$; t only parameterizes the local expansion. Thus the local expansion identifies the leading expected signed log-mass error; the squared-error relationship above is an upper bound. By Lemma 3.2 with $M = X^\top X$, the risk minimizers are exactly the bases of leading rank r eigenspaces of $X^\top X$.

Interpretation of the trace scaling. For either component matrix $M \in \{X^\top X, M_{\text{out}}\}$ and $J = I_d - P_B$, positive semidefiniteness gives

$$0 \le \frac{\operatorname{tr}(JMJ)}{\operatorname{tr}(M) + \varepsilon} \le \frac{\operatorname{tr}(M)}{\operatorname{tr}(M) + \varepsilon} < 1.$$

Without the stabilizer, this ratio is the fraction of the component's energy discarded by the basis whenever $\operatorname{tr}(M) > 0$. With $\varepsilon > 0$, a zero component contributes zero and all denominators remain positive. The scaled component has trace $\operatorname{tr}(M)/(\operatorname{tr}(M) + \varepsilon)$, which approaches one when $\operatorname{tr}(M)$ is large relative to ε. The combined risk is a weighted sum of these stabilized ratios divided by d.

Proof of Corollary 3.3. Let $J = I_d - P_B$. Under $\mathbb{E}_q[qq^\top] = I_d$, substituting (2) and (3) gives

$$\begin{aligned}\mathcal{R}_\eta(B) &= \frac{1-\eta}{d} c_{\text{key}} \operatorname{tr}(JX^\top XJ) + \frac{\eta}{d} c_{\text{out}} \operatorname{tr}(JM_{\text{out}}J) \\ &= \frac{1}{d}\operatorname{tr}(JM_\eta J).\end{aligned}$$

The factors m^2 and $1/m^2$ cancel in the second term, giving exactly the matrix M_η in (4).

Both component matrices are positive semidefinite, and their coefficients are nonnegative, so $M_\eta \succeq 0$. Equivalently, define

$$A_\eta = (1-\eta)c_{\text{key}} I_m + \eta c_{\text{out}} Y G Y^\top \succeq 0.$$

Then

$$M_\eta = X^\top A_\eta X = (A_\eta^{1/2} X)^\top (A_\eta^{1/2} X).$$

Lemma 3.2 now gives the leading r-dimensional invariant subspaces of M_η as the minimizers. If its eigenvalues are $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, the minimum risk is $d^{-1}\sum_{j>r}\lambda_j$. The Gram factorization also identifies the minimizers as leading right singular subspaces of $A_\eta^{1/2} X$, consistent with the construction in Section 3.3. □

Recovery from the token space Gram matrix. The token space matrices $T = XX^\top$ and $W = YGY^\top$ used in Algorithm 1 satisfy $\operatorname{tr}(T) = \operatorname{tr}(X^\top X)$ and $\operatorname{tr}(WT) = \operatorname{tr}(X^\top Y G Y^\top X)$ by cyclicity of the trace. Thus the algorithm computes the same normalization constants without forming $d \times d$ matrices.

For the positive eigenpairs in (5),

$$B_+^\top B_+ = \Lambda_s^{-1/2} U_s^\top C C^\top U_s \Lambda_s^{-1/2} = I_s,$$
$$M_\eta B_+ = C^\top (C C^\top U_s) \Lambda_s^{-1/2} = B_+ \Lambda_s.$$

The positive eigenvalues of $C^\top C$ and $CC^\top$ coincide, so B_+ contains leading orthonormal eigenvectors of M_η. If $s = r$, it is already an optimal rank r basis.

If $s < r$, then $s = \operatorname{rank}(C)$ and B_+ spans $\operatorname{range}(M_\eta)$. Thus every column added by an orthonormal completion lies in $\ker(M_\eta)$ and preserves the minimum hybrid risk. To select among these completions, let $D \in \mathbb{R}^{d\times(r-s)}$ satisfy $D^\top D = I_{r-s}$ and $B_+^\top D = 0$. With $B_\eta = [B_+, D]$, the retained key energy is

$$\operatorname{tr}(B_\eta^\top X^\top X B_\eta) = \|XB_+\|_F^2 + \|XD\|_F^2 = \|XB_+\|_F^2 + \|X_\perp D\|_F^2 .$$

The first term is fixed. The second is maximized by the leading right singular directions of $X_\perp$, restricted to the orthogonal complement of B_+. Every positive right singular direction already lies in this complement because $X_\perp B_+ = 0$. After all positive directions are retained, any further orthonormal columns in the complement contribute zero key energy. This proves that the completion used in Algorithm 1 minimizes $\mathcal{R}_{\text{key}}$ among all completions of B_+ while remaining optimal for $\mathcal{R}_\eta$.

A.5 Extensions

A.5.1 Nonisotropic query distributions

The main construction assumes $\mathbb{E}_q[qq^\top] = I_d$. A general positive definite query second moment can be incorporated by whitening.

Corollary A.10 (Query whitened output aware basis)**.** *Let $\Sigma_q = \mathbb{E}_q[qq^\top] \succ 0$, $\xi = \Sigma_q^{-1/2} q$, and $\widetilde{X} = X\Sigma_q^{1/2}$. Then $\mathbb{E}[\xi\xi^\top] = I_d$.*

Define

$$\widetilde{M}_{\text{out}} = \widetilde{X}^\top Y G Y^\top \widetilde{X},$$

together with

$$\widetilde{c}_{\text{key}} = \frac{1}{\operatorname{tr}(\widetilde{X}^\top \widetilde{X}) + \varepsilon}, \qquad \widetilde{c}_{\text{out}} = \frac{1}{\operatorname{tr}(\widetilde{M}_{\text{out}}) + \varepsilon}.$$

The normalized page matrix in whitened coordinates is

$$\widetilde{M}_\eta = (1-\eta)\widetilde{c}_{\text{key}} \widetilde{X}^\top \widetilde{X} + \eta \widetilde{c}_{\text{out}} \widetilde{M}_{\text{out}}. \tag{15}$$

Its leading rank r eigenspace minimizes the corresponding normalized risk in the whitened query coordinates.

Table 5: Rank and sparse KV budget ablations on GPQA with DeepSeek R1 Qwen 7B. The rank study fixes the budget at 2048 tokens, while the budget study fixes $r = 8$. Higher is better.

Ablation	Setting	GPQA
Rank	$r = 1$	36.67
	$r = 4$	38.75
	$r = 8$	**39.58**
	$r = 16$	39.58
KV budget	512	32.50
	1024	36.25
	2048	39.58
	4096	**41.25**

Proof. Since $q = \Sigma_q^{1/2}\xi$,

$$z(q) = \frac{Xq}{\sqrt{d}} = \frac{\widetilde{X}\xi}{\sqrt{d}},$$

and

$$\mathbb{E}[\xi\xi^\top] = \Sigma_q^{-1/2}\Sigma_q\Sigma_q^{-1/2} = I_d.$$

The derivations of (2) and (3) therefore apply directly with $\widetilde{X}$ in place of X. Applying Corollary 3.3 in the whitened coordinates gives (15). □

At runtime, the corresponding query code is $\widetilde{B}^\top\Sigma_q^{-1/2}q$, with stored coefficients $\widetilde{X}\widetilde{B}$. If Σ_q is singular, whitening can be performed on its support or using $\Sigma_q + \varepsilon I_d$ for numerical stability.

A.5.2 Nonuniform within page attention

The main method evaluates output sensitivity at uniform within page attention. For a general reference point z, define

$$A_{\text{out}}(z) = S(\beta(z))VGV^\top S(\beta(z)) \succeq 0.$$

By Lemma A.2, for a small logit perturbation e,

$$\|f(z) - f(z-e)\|_G^2 = e^\top A_{\text{out}}(z)e + O(\|e\|_2^3).$$

This yields the Gauss Newton risk

$$\mathcal{R}_{\text{GN}}(B) = \frac{1}{d}\mathbb{E}_{q\sim\mathcal{D}}\left[q^\top(I_d - P_B)X^\top A_{\text{out}}(z(q))X(I_d - P_B)q\right]. \tag{16}$$

This gives a hierarchy of increasingly query dependent constructions. The main method evaluates sensitivity at uniform attention and retains a single query free eigensolution. A fixed nonuniform reference can instead define a frozen local metric, while (16) can be estimated using sampled queries at the cost of losing the single closed form eigenproblem.

B Rank and Page Budget Ablations

We further evaluate sensitivity to the representation rank and sparse KV budget on GPQA using DeepSeek R1 Qwen 7B. For the rank ablation, we fix the budget to 2048 tokens and vary $r \in \{1, 4, 8, 16\}$. For the budget ablation, we fix $r = 8$ and vary the retrieved KV budget over $\{512, 1024, 2048, 4096\}$ tokens. All other settings follow the main experiments.

The rank ablation measures the sensitivity of retrieval quality to the capacity of the page representation, while the budget ablation measures the effect of increasing the number of retrieved KV tokens. Together, these experiments test the robustness of the default configuration $r = 8$ with a 2048 token budget used in the main evaluation.

Table 6: InfiniteBench LongQA at contexts exceeding 100K tokens with Llama-3.1-8B. LongQA averages En.MC, En.QA, Zh.QA, En.Sum, and En.Dia. All compressed methods use a KV budget of 2048; OVAL uses $\eta = 0.25$. Overhead is decode latency relative to FreeKV. Gold and silver rank compressed methods by accuracy; FullKV is a reference and is excluded from ranking.

Metric	FullKV	ArkVale	RaaS	FreeKV	OVAL
LongQA accuracy	42.9	35.9	35.5	36.0 🥈	**37.4** 🥇
Fraction of FullKV (%)	100.0	83.7	82.8	83.9	87.2
Decode latency (ms/token)	–	76.9	27.4	23.0	24.7
Overhead vs. FreeKV	–	+234.3%	+19.1%	0.0%	+7.4%

C InfiniteBench Results

We additionally evaluate on InfiniteBench to test a longer context regime than the 64K LongBench v2 setup. Table 6 reports the LongQA results underlying Figure 2(b), which average En.MC, En.QA, Zh.QA, En.Sum, and En.Dia with 50 examples per task.

OVAL reaches 37.4 LongQA accuracy, compared with 36.0 for FreeKV, 35.9 for ArkVale, and 35.5 for RaaS, recovering 87.2% of FullKV accuracy versus 83.9% for FreeKV. This gain costs 1.7 ms per token, a 7.4% decode overhead relative to FreeKV. OVAL is also faster than both RaaS (27.4 ms) and ArkVale (76.9 ms), placing it on the accuracy–latency Pareto frontier among the evaluated retrieval methods. A gap of 5.5 points to FullKV remains at this budget, indicating that contexts beyond 100K tokens remain challenging for all compressed methods.

D Implementation details and reproducibility

Artifact. Code, the FreeKV patch, and per experiment scripts are available at `https://anonymous.4open.science/r/oval-kv-4424`. The implementation patches FreeKV commit `2c8a7d2` rather than vendoring the upstream repository.

D.1 Software, hardware, and build

Inference	Python 3.10, torch 2.5.1+cu121, transformers 4.45.2, flash-attn 2.7.4.post1	**Judge**	Python 3.11, vLLM 0.9.2, transformers 4.53.2
GPU	RTX A6000, 48 GB, `sm_86`	**Driver**	CUDA 12.2
Build	CMake + Ninja, target `freekv_cpp`	**Arch.**	`TORCH_CUDA_ARCH_LIST=8.6`
Patch	`patches/freekv.patch`	**Size**	989 lines, 19 files

The CUDA extension is built from `source/freekv_cpp`. `CUDA_HOME` points to the active conda environment and its `nvcc`. Sources are discovered through `file(GLOB "src/*.cu" "src/*.cpp")`, so adding a CUDA source file requires no CMake modification.

Environment constraints. `flash_attn` is imported at module load and is therefore a hard dependency. The judge requires transformers < 4.54 because later versions register an `aimv2` configuration that conflicts with vLLM 0.9.2. We support cu121 through cu128; cu130 does not load with the CUDA 12.2 driver.

D.2 Models

Tag	Checkpoint	Used for
`llama-3.1-chat-8b`	`NousResearch/Meta-Llama-3.1-8B-Instruct`	LBv2, LGB, latency
`qwen-2.5-chat-7b`	`Qwen/Qwen2.5-7B-Instruct`	LBv2, LGB, latency, memory
`qwen-2.5-chat-14b`	`Qwen/Qwen2.5-14B-Instruct`	LBv2, LGB
`ds-r1-llama-8b`	`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`	reasoning
`ds-r1-qwen-7b`	`deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`	reasoning
`ds-r1-qwen-14b`	`deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`	reasoning

Shared loading configuration.

Precision	bfloat16	**Attention**	`eager`
Device mapping	`auto`	**CPU loading**	`low_cpu_mem_usage=True`
Tokenizer	slow, `use_fast=False`	**Context limit**	65,536 tokens

Stop ids are taken from each checkpoint's `GenerationConfig`. For Qwen2.5 14B, the default `device_map="auto"` headroom can place layers on meta tensors and cause the attention patch to fail. We therefore increase the per device allowance through `KVC_MAX_MEMORY`, using `45GiB` when required.

D.3 Method configuration

Core method configuration. We use page size 32, rank $r = 8$, and $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$ throughout.

Rank cap	$\min(r, \text{page} - 1, d)$
Record dtype	bfloat16
GQA aggregation	`avgSM`
Spectral retrieval	2 steps
Correction group	`avg`
Correction threshold	$\tau = 0.8$ or 0.9, depending on the benchmark
Dense layers	layer 0 remains dense
Last layer budget	0

Numerical construction. Summaries are constructed in float32 and cast back to the cache dtype. A_η is formed in the page $\times$ page Gram domain. We clamp traces at 10^{-30} and normalize each matrix to unit mean diagonal before eigendecomposition. Eigenvalues are clamped at 0 before the square root and singular values at 10^{-10} before basis construction. Because centering removes one degree of freedom, a page of size 32 supports rank at most 31.

D.4 Benchmark protocols

	Reasoning	LongBench v2	LongGenBench	InfiniteBench
Data	AIME 30, MATH 50, GPQA 50	503: 180/215/108	400, 100/type	11×50
$\mathcal{B}$	2048	2048	2048	2048
$\mathcal{S} = \mathcal{W}$	512	128	512	128
CLI budget	$2048 + 512 + 512$	$1792 + 128 + 128$	$1024 + 512 + 512$	$1792 + 128 + 128$
τ	0.9	0.8	0.9	0.8
Decoding	$T = .6,\ p = .95$	greedy	$T = .95,\ p = .95$	greedy
Max gen.	16,384	128	16,000	task specific
Input	unchanged	middle truncate 65,536	unchanged	middle truncate 65,536
Stop	EOS	EOS	`*** finished`	EOS
Seed	42	42	42	42
Metric	pass@k, avg@k	Overall/S/M/L	CR, CR$\times$Acc	official scorers

Shared sparse retrieval settings.

`--method spec_ret --page_rep oval --oval_rank 8 --eta` η `--GQA_policy avgSM --spec_ret_steps 2 --page_size 32 --sparsity 1 --skip_layer 1 --last_layer_budget 0 --corr_group avg --batch_size 1.`

Baseline arms differ only in `--page_rep` (`quest`, `arkv`) or `--method` (`raas`, `full`).

Budget matching. Every method uses the same effective KV budget, $\mathcal{B} = 2048$. The accuracy and performance harnesses express this budget through different combinations of selectable, sink, and recent regions; differences in the raw CLI fields therefore do not imply different effective budgets.

Released code	Reported experiments
120,000 token truncation	65,536
LongGenBench $p = 1.0$	$p = 0.95$
authors' absolute judge path	local judge path
`page_rep=quest`	`page_rep=oval`
`GQA_policy=avgS`	`GQA_policy=avgSM`

LongBench v2. We use `0shot.txt` without chain of thought or retrieval. Inputs exceeding 65,536 tokens retain the first and last 32,768. Decoding is greedy `argmax`; although the harness exposes a temperature argument, it is unused. A prediction is valid only if `"The correct answer is (X)"` is produced within 128 tokens. Consequently, 6 to 13% of examples remain unparsable because generation is still in the reasoning phase at the limit; this behavior is shared across methods. `KVC_MAX_TOKENS=65536` fixes the summary slab size.

Reasoning. We sample $k = 8$ generations per problem with $T = 0.6$ and $p = 0.95$. Independent generations use separate `--seed` values, encoded by the harness using a `-seed`n suffix for nondefault seeds. MATH500 and GPQA use the 50 problem subsets distributed with the harness.

InfiniteBench. Because InfiniteBench is absent from upstream FreeKV, we add its evaluator and execute both OVAL and all baseline arms through the same harness. Each task uses 50 examples sampled with seed 42 rather than the first 50 records. This avoids bias from the presorted distributed files, including `kv_retrieval`, which is ordered by needle position. Prompt templates, task specific generation limits, and official scorers come from the InfiniteBench reference implementation. We omit `math_calc`, leaving 11 tasks.

LongGenBench judge. The evaluator is Qwen3 32B with tensor parallelism 2, $T = 0.95$, $p = 0.95$, maximum generation 50, seed 42, `max_model_len=20480`, `max_num_seqs=8`, and `gpu_memory_utilization=0.82`. On systems without NVLink, `NCCL_P2P_DISABLE=1` avoids initialization deadlock. Because prediction files append rather than overwrite, the judge evaluates the final 400 records and requires the file length to be a multiple of 400.

D.5 Latency Evaluation

	Long input	**Long generation**
Dataset	`gov_report`	`lgbench`
Input	repeated and truncated to 32,000	620 to 660 tokens
Generated tokens	512	16,384
τ	0.8	0.9
Batch	$1, 2, 4$	$1, 2, 4$

Timing. Measurements use `source/pred.py`. One complete generation is used as warmup and all of its per step measurements are discarded. We report the mean interval between subsequent generated tokens. Every method uses $\mathcal{B} = 2048$.

The performance harness uses `--budget 2048 --sink 512 --recent 512 --page_size 32`. Sink and recent regions are accounted for internally when constructing the selectable budget.

Fixed offload configuration. 6 GiB GPU page pool · 20 GiB host pool · `HND` · 2 recall threads · 2 recall streams · `FREEKV_WORKSPACE_MIB=512`.

The measured arms are ArkVale (`--recall_impl arkvale`), RaaS (`--method raas`, drop only, $\alpha = 0.02$), FreeKV (`--spec_ret --corr` τ), and OVAL with the same FreeKV configuration plus `--page_rep oval --eta` η. `OVAL_QUANT` selects bfloat16 or packed records.

LongBench v2 latency uses `--dataset longbench2`, `--lb2_length {short,medium,long}`, items 148/190/39, `--max_length 65536`, and `--max_gen 128`.

D.6 Decoding released output paths

⟨model⟩-spec_ret-s128-r128-1.00-manual-p32-b1792-avgSM-oval_r8_eta0.25_q0-pQ2-llb0-pQcs0.8-skl1

The fields decode as follows: `s/r` = sink/recent; `1.00` = sparsity; `manual` = attention pattern; `p32/b1792` = page size/budget field; `avgSM` = GQA policy; `oval_r8` = representation and rank; `_eta` = η; `_q` = packed record indicator; `pQ2` = retrieval steps; `llb0` = last layer budget; `pQcs0.8` = τ; and `skl1` = number of initially dense layers. The optional `-t`, `-topp`, and `-Mg` fields appear only when they differ from their defaults.

D.7 Corrections to the reference harness

1. Index range selection does not slice metadata.
Prompts are sliced but the associated metadata is not, so generations from a selected subrange are evaluated against another example's expected structure. We correct the metadata indexing together with the prompt range.

2. Head reordering can duplicate columns.
When full and dynamic attention masks overlap, the reconstructed output projection can exceed $n_{\text{heads}}d$. On Qwen2.5 14B we observe width 5151 rather than 5120. The issue occurs only under multi GPU sharding. We therefore derive the GQA group size from the reordered projection instead of `q_proj`/`k_proj`, validate the resulting width, and fail explicitly on mismatch.

D.8 Scope and reported quantities

What OVAL changes. Only the page representation used for *retrieval*. After top k selection, attention is computed using the original keys of the selected pages, exactly as in the compared retrieval baselines.

In the accuracy harness, the full KV cache remains resident and page summaries are additional state. These experiments therefore quantify *sparse retrieval* rather than compressed key attention.

Best over η. For each reported metric,

$$\eta^\star = \arg \max_{\eta \in \{0, 0.25, 0.5, 0.75, 1\}} \text{metric}(\eta).$$

The maximizing value may differ across metrics for the same model. Accordingly, entries within a complete table row do not necessarily arise from one common η. Section 4.4 reports the corresponding sensitivity.