ATTENTION UNDER BOUNDED KEY–QUERY SIMILARITY: SPACE COMPLEXITY AND TRANSFORMER ANISOTROPY

Anonymous authors
Paper under double-blind review

ABSTRACT

A growing body of work on KV-cache compression aims to reduce the memory requirements of autoregressive generation. These methods compress the stored context while approximately preserving attention computation, enabling generation with reduced memory consumption. Almost tight theoretical results on the worst case space complexity of attention have recently been obtained, but they are achieved by instances where queries can be perfectly aligned with keys. This directly contradicts the observed anisotropy of the key/query geometry in real transformers. In particular, it has been shown empirically that key/query cosine similarities are small even for most correlated pairs, i.e., near-perfect alignment of queries with keys does not happen in practice. In this paper, we study the space complexity of attention approximation under the assumption that key/query cosine similarities belong to a small interval around zero, i.e., a query cannot align too much with *any* key. We give a nearly space optimal algorithm and a corresponding space lower bound. Surprisingly, the distribution of key/query cosine similarities in our lower bound exhibits the *anisotropy* property empirically established and studied in the literature. Additionally, we show that the simple *uniform sampling* algorithm is nearly optimal for a typical range of parameters, supporting its surprising effectiveness on real datasets.

1 INTRODUCTION

The high memory consumption of modern large language models (LLMs) is a major challenge in their deployment, as the decoding stage of transformers requires storage of all tokens seen so far. The growing body of work on Key-Value (KV) cache compression has been proposed to address this challenge. These algorithms rely on innovative ways of approximating the attention function of the context using small space, either by *quantizing* the vectors, or *evicting* some of them to retain the most informative ones. In both cases, the goal is to approximate the value of the attention function:

$$\mathrm{Attn}(K, V, q) = \frac{\sum_{i=1}^{n} e^{\langle k_i, q\rangle/\sqrt{d}} v_i}{\sum_{i=1}^{n} e^{\langle k_i, q\rangle/\sqrt{d}}}. \tag{1}$$

where $K = (k_1, \ldots, k_n) \subset \mathbb{R}^d$ is the set of all *keys*, $V = (v_1, \ldots, v_n) \subset \mathbb{R}^d$ is the set of corresponding *values* and query $q \in \mathbb{R}^d$ is the *query*.

We consider randomized algorithms for token eviction that are *query-oblivious*. That is, before seeing the specific query, the algorithm constructs a weighted summary by selecting a random subset of token indices $I \subseteq [n]$ together with positive weights w_i (and evicts the other tokens). Given a query q, the algorithm returns

$$\widehat{\mathrm{Attn}}(K, V, q) = \frac{\sum_{i\in I} w_i e^{\langle k_i, q\rangle/\sqrt{d}} v_i}{\sum_{i\in I} w_i e^{\langle k_i, q\rangle/\sqrt{d}}}. \tag{2}$$

For every fixed query q, we require

$$\Pr\Big[\big\|\widehat{\mathrm{Attn}}(K, V, q) - \mathrm{Attn}(K, V, q)\big\| \le \varepsilon\Big] \ge \frac{2}{3}. \tag{3}$$

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.

The probability is over the randomness used to construct the summary. The query is fixed independently of this randomness; we do not require simultaneous success for all queries. Assume that $\|k_i\| \le r_k$ for every i and $\|q\| \le r_q$. We will assume $\|v_i\| \le 1$.

Query-oblivious weighted token eviction (a.k.a. coresets) has been a subject of multiple recent theoretical investigations, leading to (almost) tight bounds on the *worst case* space complexity Haris & Onak (2025); Liberty et al. (2026); Chen et al. (2026), see Table 2 for details. These results show that an $e^{r_k r_q/\sqrt{d}}$ dependence in the space bound is sufficient and necessary. While these results are close to optimal *in the worst case*, some of their implications significantly deviate from what one sees in practical instances:

1. The space lower bound of $e^{r_k r_q/\sqrt{d}}$ can be much larger than the size of the input itself, given that the values of r_k, r_q are often on the order of $15 - 20$ (see, for example, Table 3 in Kochetkova et al. (2025) and our Figure 9) and $d = 128$; in such cases, the upper bound becomes vacuous.
2. The hard distribution of the keys used to prove the lower bounds is isotropic, whereas in practice, keys and queries are known to be *anisotropic*, concentrating in a relatively narrow cone of the high-dimensional space Godey et al. (2024).
3. The worst case hard instances have the queries align perfectly (i.e., have a high dot product) with one of the keys; in contrast, in practice, the key-query cosine similarity is small in absolute value once the attention sink is excluded and stored separately (see e.g., Mazaré et al. (2025) and our Figure 1).

1.1 Our results

Reflecting this discrepancy, we study the space complexity of attention approximation under the following assumption:

(⋆) Every key/query cosine similarity lies in some range $[a, b]$, where the magnitudes $|a|, |b| \ll 1$, i.e., no query can align significantly with *any* key.

Before we state our main theorem, we need to introduce some notation. Fix radius parameters r_k, r_q. Suppose $\|k_i\| \le r_k$ for every i, and fix a pair of parameters $a, b \in [-1, 1]$. We take the admissible query set to be[1]

$$Q := \left\{ q \in \mathbb{R}^d \colon \|q\| \le r_q,\ a \le \frac{\langle k_i, q\rangle}{\|k_i\|\|q\|} \le b \quad \forall i \in [n] \right\}. \tag{4}$$

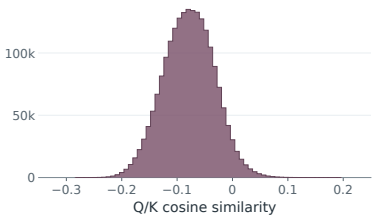


Figure 1: Cosine similarities between the final 256 queries and preceding context keys in layer 8 of Qwen3-8B, for an 8K-token LongBench v2 input, excluding the first key (attention sink).

Let $\mathbf{OPT}_{\mathsf{Attn}}$ denote the optimal worst-case number of retained tokens subject to

$$\Pr\left[\left\|\widehat{\mathrm{Attn}}(K, V, q) - \mathrm{Attn}(K, V, q)\right\| \le \varepsilon\right] \ge \frac{2}{3}$$

for every fixed admissible q. The algorithm knows keys, values, and parameters r_k, r_q, a, b, but not the particular query in Q. All norms are Euclidean, and all logarithms are natural. We use $\widetilde{O}$ and $\widetilde{\Omega}$ to suppress factors polylogarithmic in n.

Our main result shows that importance sampling is a nearly optimal token eviction policy for attention approximation. Surprisingly, for a wide range of parameters (see Figure 2) all importance weights can be set to 1, i.e. the most basic uniform sampling is optimal. Formally, we define two regions in the space of parameters (a, b), see also Figure 2 for an illustration:

Definition 1 (Uniform & importance sampling regimes)**.** *We say that fixed parameters $-1 \le a \le b \le 1$ lie in the* uniform sampling *regime if $a < 0 < b$ and $b - a \le \frac{1+b}{2}$.*

We say that fixed parameters $-1 \le a \le b \le 1$ lie in the importance sampling *regime in any other case.*

Our main result is as follows.

[1] Throughout this paper, we assume that $\|k_i\|, \|q\|$ are non-zero.

Theorem 2. *Fix any constant $\varepsilon \in (0, 1/2]$ independently of n, and assume $d = \omega(\log n)$. Fix parameters $r_k, r_q > 0$ and $-1 \le a < b \le 1$. Suppose $\|k\| \le r_k$ for every key, and the admissible query set Q is described as Equation (4). Let $R^2 = r_k r_q / \sqrt{d}$, then*

1. *In the* uniform sampling *regime, uniform sampling achieves asymptotically tight space bounds:*

$$\boldsymbol{\Omega}\Big(\min\Big\{n,\, \boldsymbol{e^{(b-a)R^2}}\Big\} \cdot e^{-o(R^2)}\Big) \;\le\; \mathbf{OPT}_{\mathsf{Attn}} \;\le\; \boldsymbol{O}\Big(\min\Big\{n,\, \boldsymbol{e^{(b-a)R^2}}\Big\}\Big).$$

2. *In the* importance sampling *regime, importance sampling achieves asymptotically tight space bounds:*

$$\boldsymbol{\Omega}\Big(\min\Big\{n,\, \boldsymbol{e^{\frac{1+b}{2}R^2}}, \boldsymbol{e^{(b-a)R^2}}\Big\} \cdot e^{-o(R^2)}\Big) \;\le\; \mathbf{OPT}_{\mathsf{Attn}} \;\le\; \boldsymbol{O}\Big(\min\Big\{n,\, \boldsymbol{e^{\frac{1+b}{2}R^2}}, \boldsymbol{e^{(b-a)R^2}}\Big\}\Big).$$

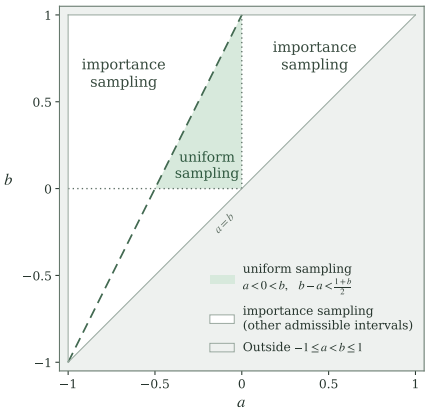


Figure 2: Uniform sampling and importance sampling regions in parameter space.

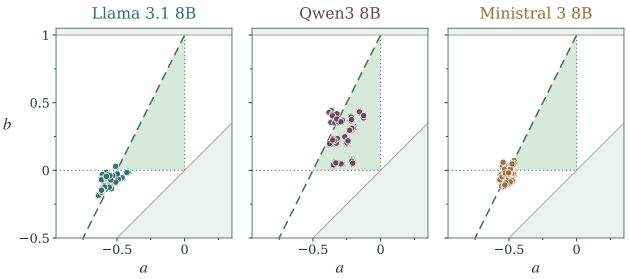


Figure 3: Empirical bounds a and b for 8B models Llama 3.1, Qwen3, and Ministral 3. Each point represents one context (from LongBench v2/InfiniteBench/SCBench)–context-length (8/32/128/256 k)–layer(4/8/16/24) configuration, with endpoints a and b computed as 0.001 and 0.999 quantiles of cosine similarities distribution. Initial-token (attention sink) is excluded. See Section F.1 for more plots.

Theorem 2 is achieved by uniform sampling when the key/query cosine similarities are small ($a < 0 < b$ and $b - a \le \frac{1+b}{2}$), and by an importance sampling scheme for larger cosine similarities. Interestingly, even though uniform sampling is by no means the best token eviction policy, it is quite competitive with some earlier works on token eviction for KV cache compression, especially query-independent ones. See, for example, Table 1 in Kochetkova et al. (2025) for a comparison of SnapKV and BalanceKV with uniform sampling. The recent work of Wang et al. (2026) provides more evidence for this phenomenon. Figure 2 helps explain this from the perspective of theory: the empirical (a, b) values for widely used models lie in or close to the uniform sampling regime.

In addition to the theorem statement, the *proof* of the theorem provides insights into another empirical phenomenon. Specifically, the anisotropy of the key/query space that has been empirically observed in the literature Godey et al. (2024) appears naturally in our lower bound instance. In our hard input distribution on key and query vectors, the keys and queries are similarly distributed except that the keys are positively shifted along a specific direction e_0 while the queries are negatively shifted along e_0. This construction, which is critical to establishing a matching space lower bound, is indeed anisotropic: the keys have a large positive projection on a common direction, and the queries have a large negative projection on the same direction, similarly to distributions observed in practice (Table 1, Figure 5). See Section 2.2 for more details.

Table 1: Cosine similarity between mean query and key vectors. Values are means ± standard deviations across example/context-length configurations common to all models and layers.

Model	Layer 4	Layer 8	Layer 16	Layer 24
Llama 3.1 8B	-0.584 ± 0.148	-0.725 ± 0.098	-0.681 ± 0.076	-0.688 ± 0.105
Qwen3 8B	0.068 ± 0.037	-0.146 ± 0.034	0.024 ± 0.044	-0.143 ± 0.080
Ministral 3 8B	-0.724 ± 0.047	-0.628 ± 0.086	-0.613 ± 0.078	-0.658 ± 0.045

1.2 Related work

Related work on worst-case space complexity of attention approximation. Prior work (Haris & Onak, 2025; Liberty et al., 2026; Chen et al., 2026) has established nearly tight asymptotic bounds on the space complexity of approximating attention in the worst case. The algorithmic landscape is diverse, with approaches drawing on sketching, random sampling, discrepancy theory, and kernel-density estimation. Provable attention approximation has also led to implemented KV cache compression methods. SubGen (Zandieh et al., 2024) combines online key clustering with ℓ_2 sampling of values to obtain approximation guarantees that exploit clustering structure, and evaluates its implementation on long-context question-answering tasks. BalanceKV (Kochetkova et al., 2025) develops a streaming compression algorithm based on geometric discrepancy and vector balancing, with both theoretical guarantees and empirical evaluations of attention approximation and end-to-end generation on long-context benchmarks. A comparison is given in Table 2. We note that the worst case setting is crucially different from our setup in that queries can be arbitrarily well aligned with key.

Related work on KV cache compression. We mention the lines of the most relevant work.

KV subset selection. A prominent approach to KV-cache compression retains a subset of the original key-value pairs. H_2O combines accumulated-attention heavy hitters with recent tokens (Zhang et al., 2023), while Scissorhands uses token importance that persists across decoding steps (Liu et al., 2023). StreamingLLM keeps initial attention sinks and a sliding window of recent tokens (Xiao et al., 2024). SnapKV selects KV positions per head before generation using a window at the prompt's end (Li et al., 2024). PyramidKV splits a fixed cache budget across layers (Cai et al., 2025), while Ada-KV splits it across heads using an attention-output error bound (Feng et al., 2025). These approaches exploit historical attention, prompt observations, or layer/head structure, rather than establishing optimal retained-token complexity under an explicit restriction on future key-query similarities.

Compressed KV representations. Another line of work reduces storage per cached vector. KIVI quantizes keys per channel and values per token (Liu et al., 2024). KVQuant combines distribution-aware quantization, pre-RoPE key quantization, and separate outlier handling (Hooper et al., 2024). TurboQuant uses random rotations for data-oblivious vector quantization, with guarantees on mean-squared and inner-product distortion (Zandieh et al., 2026). Palu caches low-dimensional intermediate representations from low-rank factorizations of key/value projections (Chang et al., 2025). Cartridges learns context representations through self-study (Eyuboglu et al., 2025). These representations fall outside our weighted-subset model, so our retained-token lower bounds do not apply to arbitrary quantized or low-rank encodings.

Random KV eviction. Wang et al. (2026) propose a method that reserves cache space for the input prompt and the most recent tokens, then fills the remaining slots by uniform random sampling, independently across KV heads at each eviction event. Their method is competitive with importance-based policies on reasoning benchmarks, providing further evidence that random sampling can be effective in practice.

Table 2: Worst-case space complexity of attention. $\widetilde{O}, \widetilde{\Omega}$ suppress $\mathrm{poly}(\varepsilon^{-1} \log n)$ factors.

Reference	Model / regime	Upper bound	Lower bound
Kochetkova et al. (2025)	Coreset size (tokens)	$\widetilde{O}\left(\sqrt{d}e^{2R^2}\right)$	$\widetilde{\Omega}(1)$
Haris & Onak (2025)[a]	Space complexity (bits)	$\widetilde{O}(nd)$	$\Omega(nd)$
Liberty et al. (2026)[b]	Coreset size (tokens)	$\widetilde{O}\left(\sqrt{d}e^{R^2+\log R+o(\log\log R)}\right)$	$\widetilde{\Omega}\left(\sqrt{d}e^{R^2}\right)$
Chen et al. (2026)[c]	Space complexity (bits)	$\widetilde{O}\left(e^{(1+o(1))R^2}\right)$	$\widetilde{\Omega}\left(e^{(1-o(1))R^2}\right)$
This work[d], $b-a \le \frac{1+b}{2}$	Coreset size (tokens)	$\widetilde{O}\left(e^{(b-a)R^2}\right)$	$\widetilde{\Omega}\left(e^{(b-a-o(1))R^2}\right)$
This work[d], $b-a > \frac{1+b}{2}$	Coreset size (tokens)	$\widetilde{O}\left(e^{\frac{1+b}{2}R^2}\right)$	$\widetilde{\Omega}\left(e^{(\frac{1+b}{2}-o(1))R^2}\right)$

[a] For entry-wise approximation guarantee.
[b] The coreset guarantee holds simultaneously for all queries in the prescribed norm ball.
[c] Assuming $d = \mathrm{polylog}(n)$ and $R^2 = \Omega(\log n)$.
[d] Figure 3 and Figure 6 suggest that our bounds are $\tilde{O}(e^{\approx 0.6R^2})$ for most practical configurations of a and b.

2 UNIFORM SAMPLING REGIME

2.1 UPPER BOUND

The upper bound in the small cosine similarities regime is shown by a simple approach: sampling tokens uniformly.

Consider the small cosine similarities regime of Definition 1, where $-1 \le a < 0 < b \le 1$ and $b - a \le \frac{1+b}{2}$. Our assumptions are $\|k_i\| \le r_k$, $\|q\| \le r_q$, and

$$a \le \frac{\langle k_i, q\rangle}{\|k_i\|\|q\|} \le b.$$

We write $R^2 = r_k r_q/\sqrt{d}$. Set $m = \left\lceil \frac{96e^{(b-a)R^2}}{\varepsilon^2} \right\rceil$. If $m \ge n$, retain the full dataset. Otherwise, sample m tokens independently and uniformly with replacement. Writing $J_1, \dots, J_m$ for their indices, return the estimate

$$\widehat{\mathrm{Attn}}(K, V, q) = \frac{\sum_{\ell=1}^{m} e^{\langle k_{J_\ell}, q\rangle/\sqrt{d}} v_{J_\ell}}{\sum_{\ell=1}^{m} e^{\langle k_{J_\ell}, q\rangle/\sqrt{d}}}.$$

The intuition behind uniform sampling is that no individual token can account for too large a fraction of the attention mass. Indeed, for every i and every admissible query q,

$$e^{\langle k_i, q\rangle/\sqrt{d}} \in [e^{aR^2}, e^{bR^2}]. \tag{5}$$

Thus, the attention denominator is at least ne^{aR^2}, whereas the contribution of any individual token is at most e^{bR^2}. Consequently, each token accounts for at most an $e^{(b-a)R^2}/n$ fraction of the attention mass.

Lemma 1 (Upper bound, Item 1)**.** *The uniform sampling procedure described above retains at most*

$$O_\varepsilon\left(\min\left\{n,\, e^{(b-a)R^2}\right\}\right)$$

tokens and gives an additive-ε approximation to attention with probability at least 2/3 for each fixed admissible query.

We defer the full proof of Lemma 1 to Section B. We use Chebyshev's inequality to bound the attention denominator, and use Markov's inequality to bound the numerator of the error.

2.2 LOWER BOUND

Recall that in our model (2), the token-eviction algorithm (randomly) selects tokens and assigns them nonnegative weights before seeing the query. We will prove lower bounds against such algorithms.

Theorem 3 (Lower bound, Item 1)**.** *Consider $r_k, r_q > 0$, $\varepsilon \in (0, 1/2]$, and $-1 \le a < b \le 1$ in the uniform sampling regime (Definition 1). Consider n and $d = \omega(\log n)$, and let $\xi = \sqrt{\frac{6\log(2n)}{d-3}} = o(1)$. Pick any integer m satisfying*

$$2 \le m \le n, \qquad m \le \boldsymbol{e^{(b-a)R^2}} \cdot e^{-2\xi R^2}/32. \tag{6}$$

There is a fixed dataset (K, V) of n key-value pairs, with keys of norm r_k and unit-norm values in $\mathbb{R}^d$, together with queries $q_1, \dots, q_m$ of norm r_q, such that every key $k \in K$ satisfies

$$a \le \frac{\langle k, q_i\rangle}{\|k\|\|q\|} \le b \qquad \text{for all } i \in [m].$$

For this dataset, satisfying Equation (3) at each of these queries requires $\mathbf{OPT}_{\mathsf{Attn}} \ge 2m/3$.

Our lower bound constructs m groups of identical key-value pairs, each paired with a different query. A visualization of the construction is given in Figure 4. For each group, we build the key and corresponding query from a set of nearly orthogonal vectors $u_1, \dots, u_m$ (Lemma 2):

Lemma 2 (Nearly orthogonal vectors)**.** *Consider integer $M \ge 2$ and $\xi \in (0, 1)$. Let D be an integer such that $D \ge 6\log(2M)/\xi^2$. There exist M unit vectors $u_1, \dots, u_M \in \mathbb{R}^D$ such that $|\langle u_i, u_j\rangle| \le \xi$ for every $i \ne j$.*

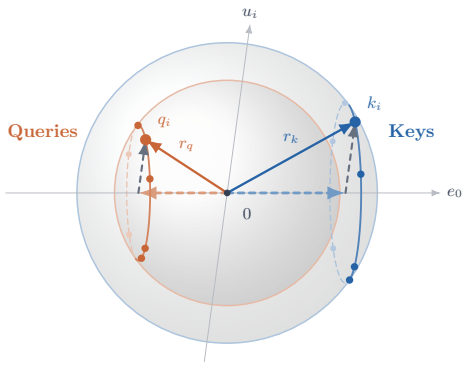


Figure 4: A visualization of the hard instance in our lower bound construction. The keys and queries are shifted away from each other along direction e_0. Orthogonal to e_0, each key-query pair (k_i, q_i) follows its own common direction u_i.

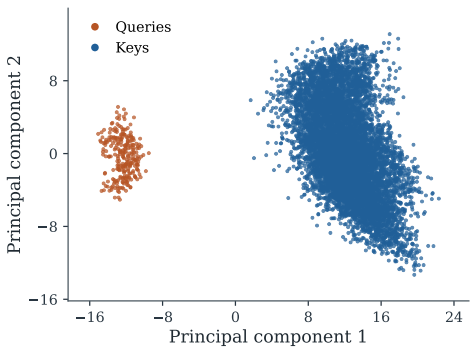


Figure 5: Query-key geometry in Llama 3.1 8B on LongBench v2 example (8K context, layer 8, head 0). A joint PCA is fitted with equal total weight for queries and keys. Initial-token (attention sink) is excluded. See Section F.1 for further plots.

To see this, choose $u_1, \ldots, u_M$ independently and uniformly from $\{-D^{-1/2}, D^{-1/2}\}^D$, so that each vector has unit norm. Hoeffding's inequality and a union bound show that the probability of any pair satisfying $|\langle u_i, u_j \rangle| > \xi$ is at most $M^2 e^{-D\xi^2/2} < 1$ under the assumed bound on D. We give the full proof in Section C.

In order to keep all key-query cosine similarities in $[a, b]$, queries and keys are anisotropically shifted away from each other along a fixed direction e_0. A pair of keys and queries in the same group k_i, q_i are (a weighted version of) $u_i \pm e_0$ where the sign on e_0 differs for the key and query (along with a normalization term to set their respective norms to r_k, r_q).

Each query q_i has similarity b with its own group's key k_i, while the near orthogonality caps its similarity close to a with the other groups' keys k_j. The exponential in attention amplifies this gap, giving the query's own group almost all the attention mass. We also choose nearly orthogonal value vectors, so a convex combination of other groups' value vectors cannot well-approximate attention. The algorithm must therefore retain a key-value pair from each group with constant probability, otherwise it would fail with probability more than $1/3$ on some group's query.

The resulting construction holds an intriguing connection to patterns in real data. In Figure 5, we see that queries and keys are separated along a common direction and otherwise hold a somewhat similar distribution though with different scales for queries and keys. These qualities also appear in our hard instance (Figure 4).

2.2.1 Lower bound construction

Goal. We construct keys and queries whose scaled inner products are bR^2 for matching pairs (k_i, q_i) and $a(1 - o(1))R^2$ for nonmatching pairs (k_j, q_i), where $R^2 = r_k r_q / \sqrt{d}$. In Section C.1, we use this gap to make each query concentrate its attention on copies of its matching key, forcing an accurate summary to retain many key groups.

Construction. Let e_0, e_K, e_Q be the first three standard basis vectors. Choose m unit vectors satisfying Lemma 2 in dimension $D = d - 3$, with pairwise inner products bounded in magnitude by $\xi = \sqrt{\frac{6 \log(2n)}{D}}$. Embed them into the last $d - 3$ coordinates of $\mathbb{R}^d$, and denote the resulting vectors by $u_1, \ldots, u_m$. Define

$$\begin{aligned} k_i &= r_k\Big(\underbrace{\sqrt{\alpha}\, u_i}_{\text{Shared group direction}} + \underbrace{\sqrt{\beta}\, e_0}_{\text{Anisotropic shift}} + \underbrace{\sqrt{1-\alpha-\beta}\, e_K}_{\text{Norm padding}} \Big), \\ q_i &= r_q\Big(\underbrace{\sqrt{\alpha}\, u_i}_{\text{Shared group direction}} - \underbrace{\sqrt{\beta}\, e_0}_{\text{Anisotropic shift}} + \underbrace{\sqrt{1-\alpha-\beta}\, e_Q}_{\text{Norm padding}} \Big), \end{aligned} \tag{7}$$

where $\alpha \approx b - a$ and $\beta \approx -a$. Specifically, we set

$$\alpha = \frac{b-a}{1+\xi}, \qquad \beta = \frac{-a-\xi b}{1+\xi}. \tag{8}$$

We have $\alpha \geq 0$ since $b \geq a$. Since $a < 0$ and b are fixed and $\xi = o(1)$, for sufficiently large n we have $\xi b < -a$, so $\beta > 0$. Moreover, $\alpha + \beta = 2\alpha - b \leq b - 2a \leq 1$. Therefore, $1 - \alpha - \beta \geq 0$ and the keys and queries are well defined with norms exactly equal to $\|k_i\| = r_k$ and $\|q_i\| = r_q$. The components along u_i and e_0 serve the purpose of achieving the desired inner products while the padding along e_K and e_Q increases the norms of the keys/queries to fix the denominator of cosine similarity.

Verifying the inner products. For every $i, j \in [m]$, $\frac{\langle k_j, q_i \rangle}{\sqrt{d}} = \big(\alpha \langle u_j, u_i \rangle - \beta\big) R^2$. For matching pairs, $\alpha - \beta = b$ gives

$$\frac{\langle k_i, q_i \rangle}{\sqrt{d}} = bR^2. \tag{9}$$

For nonmatching pairs, $|\langle u_j, u_i \rangle| \leq \xi$ gives

$$aR^2 \leq \frac{\langle k_j, q_i \rangle}{\sqrt{d}} \leq (a + 2\alpha\xi)R^2 = aR^2 + \frac{2\xi}{1+\xi}(b-a)R^2. \tag{10}$$

Constructing the value vectors. Apply Lemma 2 a second time to obtain m unit value vectors $v_1, \ldots, v_m \in \mathbb{R}^d$ satisfying

$$|\langle v_i, v_j \rangle| \leq \frac{1}{8} \qquad \text{for all } i, j \in [m] \text{ with } i \neq j. \tag{11}$$

Since $m \leq n$, the assumption $d = \omega(\log n)$ ensures that this is possible.

The final proof of the coreset size lower bound using these key, query and value vectors is given in Section C.

3 Importance Sampling Regime

3.1 Upper Bound

When the interval $[a, b]$ is wide, the exponential in attention can make a few tokens much more influential than the rest. To account for this, uniform sampling would need to retain many tokens. We thus use *importance sampling* instead. Unlike uniform sampling, which treats all tokens equally, importance sampling gives higher sampling probabilities to tokens that can contribute more to attention. This is a standard idea in randomized linear algebra through leverage-score sampling Drineas et al. (2012) and in coreset construction through sensitivity sampling Braverman et al. (2021).

Since we sample before seeing the query, we want to sample a token with probability proportional to the largest possible attention it could receive. This quantity is its *sensitivity*: for a nonempty query set Q containing all admissible queries, define

$$\sigma_i = \sup_{q \in Q} \frac{e^{\langle k_i, q \rangle / \sqrt{d}}}{\sum_{j=1}^{n} e^{\langle k_j, q \rangle / \sqrt{d}}}. \tag{12}$$

Unfortunately, these sensitivities may be difficult to compute exactly. Instead, it suffices to use sampling scores $s_i \geq \sigma_i$, i.e., upper bounds of each sensitivities. In the sampling procedure, we sample each token i with probability $\pi_i = s_i/S$, where $S = \sum_{i=1}^{n} s_i$. We set the number of samples proportional to S, so smaller sampling scores yield a smaller sample size. This leaves two tasks: compute valid sampling scores $s_i \geq \sigma_i$ and show that their sum S is small.

We formalize the sampling procedure in Algorithm 1. For each sampled token i, we retain the corresponding key-value pair (k_i, v_i) together with a positive weight. If we sample the same token more than once, we keep one copy and add the weights together.

Algorithm 1 Attention compression by score-based sampling and reweighting

Require: key-value pairs $(k_1, v_1), \dots, (k_n, v_n)$, accuracy $0 < \varepsilon < 1$, and sampling scores $s_1, \dots, s_n > 0$

$S \leftarrow \sum_{i=1}^{n} s_i, \quad m \leftarrow \lceil 96S/\varepsilon^2 \rceil$
if $m \geq n$ **then**
  return $\{(k_i, v_i, 1) : i \in [n]\}$
end if
$\pi_i \leftarrow s_i/S$ for every $i \in [n]$
Subsample m tokens using probabilities π_i ▷ with replacement
$c_i \leftarrow$ number of times token i was sampled
return $\{(k_i, v_i, c_i/(m\pi_i)) : c_i > 0\}$ ▷ one copy per token

We refer to the set returned by Algorithm 1 as the *summary*. Given a query q, the summary produces an estimator

$$\widehat{\mathrm{Attn}}(K, V, q) = \frac{\sum_{i \in I} w_i e^{\langle k_i, q \rangle / \sqrt{d}} v_i}{\sum_{i \in I} w_i e^{\langle k_i, q \rangle / \sqrt{d}}}, \tag{13}$$

where $I := \{i : c_i > 0\}$ denotes the set of retained token indices. The denominator is positive, since the summary is nonempty and all retained weights are positive.

We first state the approximation guarantee of Algorithm 1. Using a number of tokens proportional to S, the estimator from Equation (13) approximates the attention with high probability.

Lemma 3 (Approximation guarantee)**.** *Suppose $\|v_i\| \leq 1$ and $s_i \geq \sigma_i$ for every $i \in [n]$. Then Algorithm 1 returns a weighted attention summary of size at most*

$$\min\left\{n, \left\lceil \frac{96S}{\varepsilon^2} \right\rceil\right\}.$$

For every fixed query $q \in Q$, its estimate satisfies

$$\Pr\left[\left\|\widehat{\mathrm{Attn}}(K, V, q) - \mathrm{Attn}(K, V, q)\right\| \leq \varepsilon\right] \geq \frac{2}{3},$$

where the probability is over the sampling in Algorithm 1.

We defer the full proof of Lemma 3 to Section D. Similar to the proof of Lemma 1, we use Chebyshev's inequality to bound the attention denominator, and use Markov's inequality to bound the numerator of the error.

Lemma 3 reduces the problem to constructing sampling scores s_i whose sum S is small. To this end, we first bound the total sensitivity $\sum_{i=1}^{n} \sigma_i$ of any query set for which all key-query inner products lie in a fixed interval. This condition is satisfied by our query set Q. However, Q need not be convex, making it difficult to compute sampling scores. We therefore partition the keys and possible query norms into small ranges and, for each pair of ranges, replace the cosine-similarity constraints by fixed inner-product bounds. This partitions the problem into a collection of smaller convex relaxations on which we can approximate the sensitivities. Applying the total-sensitivity bound to these relaxations allows us to control the sum S. These two steps are formalized in Lemmas 4 and 5.

Lemma 4 (Total sensitivity bound)**.** *Suppose $r_k, r_q > 0$ and $\|k_i\| \leq r_k$ for every i, and let $\mathcal{Q}$ be a nonempty compact query set such that*

$$\mathcal{Q} := \left\{q \in \mathbb{R}^d : \|q\| \leq r_q,\ \ell \leq \langle k_i, q \rangle / \sqrt{d} \leq h \quad \forall i \in [n]\right\}.$$

For sensitivities σ_i defined in Equation (12) with the supremum taken over $q \in \mathcal{Q}$, we have

$$\sum_{i=1}^{n} \sigma_i \leq \exp\left(\min\left\{h - \ell, \frac{R^2 + h}{2}\right\}\right). \tag{14}$$

At a high level, the two terms in Equation (14) come from two different ways of controlling the total sensitivity. The first term $h - \ell$ follows from the fact that $\sigma_i \leq e^h/(ne^\ell)$. For the second term, we pair each key k_i with a query q_i attaining its sensitivity and assign weights to these pairs

proportionally to their sensitivities. Writing $\bar{k}, \bar{q}$ for the weighted mean key and query, using Jensen's inequality, the proof gives

$$\log \sum_i \sigma_i \le h - \frac{\langle \bar{k}, \bar{q} \rangle}{\sqrt{d}}, \qquad \log \sum_i \sigma_i \le R^2 + \frac{\langle \bar{k}, \bar{q} \rangle}{\sqrt{d}}.$$

Averaging these two bounds yields $\log \sum_i \sigma_i \le (R^2 + h)/2$. The full proof is deferred to Section D.

It remains to construct the sampling scores.

Lemma 5 (Computing sampling scores). *Let $-1 \le a < b \le 1$ and $r_k, r_q > 0$, and write $R^2 = \frac{r_k r_q}{\sqrt{d}}, \kappa = \min\left\{b - a, \frac{1+b}{2}\right\}$. Suppose $\|k_i\| \le r_k$ for every $i \in [n]$, and let*

$$\mathcal{Q} := \left\{q \in \mathbb{R}^d : \|q\| \le r_q, \quad a\|k_i\|\|q\| \le \langle k_i, q \rangle \le b\|k_i\|\|q\| \quad \forall i \in [n]\right\}.$$

For sensitivities σ_i defined in Equation (12) with the supremum taken over $q \in \mathcal{Q}$, we can compute sampling scores s_i satisfying $s_i \ge \sigma_i$ for every $i \in [n]$, and

$$\sum_{i=1}^{n} s_i \le \frac{128 e^2}{\kappa^2} e^{\kappa R^2}.$$

Moreover, such sampling scores can be obtained by solving $n \cdot \max\{1, \lceil R^2 \rceil\}$ convex optimization problems, using only the keys and the parameters r_k, r_q, a, b.

The full proof is deferred to Section D.

We now combine the approximation guarantee (Lemma 3) with the sampling score bound (Lemmas 4 and 5) to prove the upper bound in the importance sampling regime.

Corollary 1 (Upper bound, Item 2). *Under the assumptions of Lemma 5, fix a, b independently of n, suppose $\|v_i\| \le 1$ for every $i \in [n]$, and let $0 < \varepsilon < 1$. Then Algorithm 1, with the scores constructed there, retains at most*

$$O_\varepsilon\left(\min\left\{n,\, e^{(b-a)R^2},\, e^{\frac{1+b}{2}R^2}\right\}\right)$$

tokens and gives an additive-ε approximation to attention with probability at least $2/3$ for each fixed admissible query.

Proof. By Lemma 5, the sampling scores satisfy $s_i \ge \sigma_i$ and $S \le 128 e^2 \kappa^{-2} e^{\kappa R^2}$, where $\kappa = \min\{b - a, (1 + b)/2\}$. Since a, b are fixed, κ^{-2} is a constant. Substituting into Lemma 3 and absorbing κ^{-2} into the implicit constant proves the upper bound. □

3.2 Lower bound

The lower bound in the importance sampling regime uses the same construction and analysis as in the uniform sampling regime (Section 2.2). The only change in the theorem statement is that m must satisfy

$$2 \le m \le n, \qquad m \le \min\left\{e^{\frac{1+b}{2}R^2}, e^{(b-a)R^2}\right\} \cdot e^{-2\xi R^2}/32.$$

The full statement (Theorem 4) and proof are given in Section E.

4 Open Problems

The following questions naturally arise from our work:

- Can we provably explain why transformers learn this geometry? In particular, under what assumptions on the data and training dynamics can we prove that small key–query cosine similarities and negative alignment between mean keys and queries emerge?
- Do real KV caches have other structure that our model misses? Our hard instances share some geometric features with trained models, but real caches may have additional properties that make them easier to compress.

AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims in the importance sampling regime, assist in the writing of proofs, propose or refine hypotheses and implement methods.

We have not used generative AI tools to help formulate mathematical claims, develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, interpret results, support qualitative and thematic data analysis. Generating synthetic data sets, cleaning and reformatting datasets and assisting with translation are not applicable to this work.

Additionally, we used generative AI tools for creating scientific figures, suggesting experimental parameters, creating software code, drafting parts of the paper, sourcing/searching for information, editing to improve readability, proposing a title for the paper.

We have reviewed all AI-assisted work. We checked LLM-generated parts of the proof in importance sampling regime and rewrote it ourselves. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vladimir Braverman, Dan Feldman, Harry Lang, Adiel Statman, and Samson Zhou. Efficient coreset constructions via sensitivity sampling. In Vineeth N. Balasubramanian and Ivor Tsang (eds.), *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pp. 948–963. PMLR, 17–19 Nov 2021. URL https://proceedings.mlr.press/v157/braverman21a.html.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=ayi7qezU87.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed Abdelfattah, and Kai-Chiang Wu. Palu: KV-cache compression with low-rank projection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7da6e0e00702c60607a6ae05c802ef85-Abstract-Conference.html.

Justin Y. Chen, Ying Feng, Piotr Indyk, Michael Kapralov, Ekaterina Kochetkova, and Boris Prokhorov. Towards tight bounds for streaming attention, 2026. URL https://arxiv.org/abs/2606.07205.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13 (111):3475–3506, 2012. URL https://www.jmlr.org/papers/v13/drineas12a.html.

Sabri Eyuboglu, Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Will Tennien, Atri Rudra, James Zou, Azalia Mirhoseini, and Christopher Re. Cartridges: Lightweight and general-purpose long context representations via self-study, 2025. URL https://arxiv.org/abs/2506.06266.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=tcisuhGsQZ.

Nathan Godey, Éric Villemonte de la Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pp. 35–48. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EACL-LONG.3. URL https://doi.org/10.18653/v1/2024.eacl-long.3.

Themistoklis Haris and Krzysztof Onak. Compression barriers in autoregressive transformers. In Nika Haghtalab and Ankur Moitra (eds.), *The Thirty Eighth Annual Conference on Learning Theory, June 30 - July 4, 2025, Lyon, France*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2757–2785. PMLR, 2025. URL https://proceedings.mlr.press/v291/haris25a.html.

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. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/028fcbcf85435d39a40c4d61b42c99a4-Abstract-Conference.html.

Ekaterina Kochetkova, Kshiteej Jitesh Sheth, Insu Han, Amir Zandieh, and Michael Kapralov. Streaming attention approximation via discrepancy theory. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/1de510838faddd72f79a7a0e9c564f0c-Abstract-Conference.html.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

Edo Liberty, Alexandr Andoni, and Eldar Kleiner. Nearly optimal attention coresets. *CoRR*, abs/2605.05602, 2026. doi: 10.48550/ARXIV.2605.05602. URL https://doi.org/10.48550/arXiv.2605.05602.

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

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

Pierre-Emmanuel Mazaré, Gergely Szilvasy, Maria Lomeli, Francisco Massa, Naila Murray, Hervé Jégou, and Matthijs Douze. Inference-time sparse attention with asymmetric indexing. arXiv preprint arXiv:2502.08246, 2025. URL https://arxiv.org/abs/2502.08246.

Lieven Vandenberghe and Stephen Boyd. *Convex optimization*, volume 1. Cambridge university press Cambridge, 2004.

Heng Wang, Jielin Qiu, Wenting Zhao, Cheng Qian, Liangwei Yang, Jiawei Han, Heng Ji, Silvio Savarese, Shelby Heinecke, and Huan Wang. Random attention: Rethinking KV cache eviction for efficient reasoning. arXiv preprint arXiv:2609.03430, 2026. URL https://arxiv.org/abs/2609.03430.

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

Amir Zandieh, Insu Han, Vahab Mirrokni, and Amin Karbasi. SubGen: Token generation in sublinear time and memory. arXiv preprint arXiv:2402.06082, 2024. URL https://arxiv.org/abs/2402.06082.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. TurboQuant: Online vector quantization with near-optimal distortion rate. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5c802ef38ab6e366c2ea06eee554c088-Abstract-Conference.html.

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

A PRELIMINARIES

A.1 STANDARD INEQUALITIES AND IDENTITIES

We record several standard inequalities and identities used in the proofs.

Lemma 6 (Markov's inequality)**.** *Let X be a nonnegative random variable. Then, for every $t > 0$,*

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

In particular, for any random vector X with finite second moment,

$$\Pr[\|X\| \geq t] \leq \frac{\mathbb{E}\|X\|^2}{t^2}.$$

Lemma 7 (Chebyshev's inequality)**.** *Let X be a real-valued random variable with finite variance. Then, for every $t > 0$,*

$$\Pr[|X - \mathbb{E}X| \geq t] \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Lemma 8 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [a, b]$ for every i. Define*

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, for every $\varepsilon > 0$,

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right).$$

Lemma 9 (Jensen's inequality)**.** *Let $\lambda_1, \ldots, \lambda_n > 0$ satisfy $\sum_{i=1}^n \lambda_i = 1$, and let $x_1, \ldots, x_n > 0$. Then*

$$\log\left(\sum_{i=1}^{n} x_i\right) = \log\left(\sum_{i=1}^{n} \lambda_i \frac{x_i}{\lambda_i}\right) \geq \sum_{i=1}^{n} \lambda_i \log\left(\frac{x_i}{\lambda_i}\right).$$

Lemma 10 (Cauchy-Schwarz inequality)**.** *Let $\lambda_1, \ldots, \lambda_n \geq 0$, and let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be vectors in a Euclidean space. Then*

$$\left|\sum_{i=1}^{n} \lambda_i \langle x_i, y_i\rangle\right| \leq \left(\sum_{i=1}^{n} \lambda_i \|x_i\|^2\right)^{1/2} \left(\sum_{i=1}^{n} \lambda_i \|y_i\|^2\right)^{1/2}.$$

Lemma 11 (Arithmetic-geometric mean inequality)**.** *For every $x, y \geq 0$,*

$$\sqrt{xy} \leq \frac{x+y}{2}.$$

Lemma 12 (Weighted centering)**.** *Let $w_1, \ldots, w_n \geq 0$ and let $x_1, \ldots, x_n$ be vectors in a Euclidean space. Let*

$$W = \sum_{i=1}^{n} w_i > 0, \qquad \bar{x} = \frac{1}{W}\sum_{i=1}^{n} w_i x_i.$$

Then

$$\sum_{i=1}^{n} w_i \|x_i - \bar{x}\|^2 = \sum_{i=1}^{n} w_i \|x_i\|^2 - W\|\bar{x}\|^2.$$

In particular, if $\sum_i w_i = 1$, then

$$\sum_{i=1}^{n} w_i \|x_i - \bar{x}\|^2 = \sum_{i=1}^{n} w_i \|x_i\|^2 - \|\bar{x}\|^2.$$

A.2 Convex Optimization

We record the convex optimization facts used to compute the sampling scores. The optimization guarantee below is stated in the oracle model.

Lemma 13. *Let $u_1, \dots, u_N \in \mathbb{R}^d$ and $c_1, \dots, c_N \in \mathbb{R}$. The function*

$$f(x) = \log\left(\sum_{j=1}^{N} e^{\langle u_j, x\rangle + c_j}\right)$$

is convex and continuously differentiable, with

$$\nabla f(x) = \frac{\sum_{j=1}^{N} e^{\langle u_j, x\rangle + c_j} u_j}{\sum_{j=1}^{N} e^{\langle u_j, x\rangle + c_j}}.$$

The convexity statement follows from convexity of log-sum-exp and composition with an affine map; see Vandenberghe & Boyd (2004, Sections 3.1.5 and 3.2.2). The gradient formula follows by differentiation.

Lemma 14. *A set defined by finitely many affine inequalities and constraints of the form*

$$\|Ax + c\| \le \langle u, x\rangle + \beta$$

is closed and convex. If it is bounded, it is compact.

Proof. The function $x \mapsto \|Ax + c\| - \langle u, x\rangle - \beta$ is continuous and convex. Thus its non-positive sublevel set is closed and convex, and these properties are preserved by intersection. Compactness follows from closedness and boundedness. □

Lemma 15 (Additive-accuracy convex minimization)**.** *Let $C \subseteq \mathbb{R}^d$ be a nonempty compact convex set, and let $f : C \to \mathbb{R}$ be continuous and convex. Then, for every $\eta > 0$, the convex optimization problem*

$$\min_{x \in C} f(x)$$

admits an additive-η solution; that is, there exists $\widehat{x} \in C$ satisfying

$$\min_{x \in C} f(x) \le f(\widehat{x}) \le \min_{x \in C} f(x) + \eta.$$

Proof. Since C is compact and f is continuous, f attains its minimum on C. An exact minimizer is, in particular, an additive-η solution. □

B Proofs from Section 2.1

Lemma 1 (Upper bound, Item 1)**.** *The uniform sampling procedure described above retains at most*

$$O_\varepsilon\left(\min\left\{n,\, e^{(b-a)R^2}\right\}\right)$$

tokens and gives an additive-ε approximation to attention with probability at least 2/3 for each fixed admissible query.

Proof. If the algorithm retains all tokens, the estimate is exact. Otherwise, fix an admissible query q, we write the true average and the empirical average of kernels as

$$D = \frac{1}{n}\sum_i \exp(\langle k_i, q\rangle/\sqrt{d}) \qquad \text{and} \qquad \widehat{D} = \frac{1}{m}\sum_{\ell=1}^{m} \exp(\langle k_{J_\ell}, q\rangle/\sqrt{d}).$$

We will show that

$$\Pr\left[\left\|\widehat{\mathrm{Attn}}(K, V, q) - \mathrm{Attn}(K, V, q)\right\| \le \varepsilon\right] \ge \frac{2}{3}, \quad \text{where} \quad \widehat{\mathrm{Attn}}(K, V, q) := \frac{\sum_{\ell=1}^{m} e^{\langle k_{J_\ell}, q\rangle/\sqrt{d}} v_{J_\ell}}{\sum_{\ell=1}^{m} e^{\langle k_{J_\ell}, q\rangle/\sqrt{d}}}$$

for uniform samples $J_1, \cdots, J_m$. We write the attention error vector as

$$\widehat{\mathrm{Attn}}(K,V,q) - \mathrm{Attn}(K,V,q) = \frac{\frac{1}{m}\sum_{\ell=1}^{m} \exp(\langle k_{J_\ell}, q\rangle/\sqrt{d})\big(v_{J_\ell} - \mathrm{Attn}(K,V,q)\big)}{\widehat{D}}.$$

We will use Chebyshev's inequality to bound the denominator, and use Markov's inequality to bound the norm of the numerator. By Equation (5), we have $D \geq e^{aR^2}$. Moreover, since $J_1, \ldots, J_m$ are independent,

$$\mathrm{Var}(\widehat{D}) = \frac{1}{m}\,\mathrm{Var}\Big(e^{\langle k_J, q\rangle/\sqrt{d}}\Big) \leq \frac{1}{m}\mathbb{E}\Big[e^{2\langle k_J, q\rangle/\sqrt{d}}\Big] \leq \frac{e^{bR^2}D}{m},$$

where J is uniform on $[n]$. Thus, by Chebyshev's inequality (Lemma 7),

$$\Pr[\widehat{D} < D/2] \leq 4e^{bR^2}/(mD) \leq 4e^{(b-a)R^2}/m \leq \varepsilon^2/24 \leq 1/96.$$

On the other hand, we can bound the expected squared norm of the numerator as

$$\begin{aligned}
&\mathbb{E}\Big\|\frac{1}{m}\sum_{\ell=1}^{m} \exp(\langle k_{J_\ell}, q\rangle/\sqrt{d})\big(v_{J_\ell} - \mathrm{Attn}(K,V,q)\big)\Big\|^2 \\
&\quad= \frac{1}{mn}\sum_i \exp(2\langle k_i, q\rangle/\sqrt{d})\big\|v_i - \mathrm{Attn}(K,V,q)\big\|^2 \\
&\quad\overset{\text{Equation (5)}}{\leq} \frac{e^{bR^2}}{m}\cdot\frac{1}{n}\sum_i \exp(\langle k_i, q\rangle/\sqrt{d})\big\|v_i - \mathrm{Attn}(K,V,q)\big\|^2 \\
&\quad\overset{\text{Lemma 12}}{=} \frac{e^{bR^2}}{m}\left(\frac{1}{n}\sum_i \exp(\langle k_i, q\rangle/\sqrt{d})\|v_i\|^2 - D\|\,\mathrm{Attn}(K,V,q)\|^2\right) \\
&\quad\leq \frac{e^{bR^2}D}{m}.
\end{aligned}$$

Hence, by Markov's inequality (Lemma 6), the probability that the numerator has norm larger than $\varepsilon D/2$ is at most $4e^{bR^2}/(m\varepsilon^2 D) \leq 1/24$.

Therefore, with probability at least $1 - 1/96 - 1/24 > 2/3$, we have simultaneously $\widehat{D} \geq D/2$ and numerator norm at most $\varepsilon D/2$, which implies $\|\widehat{\mathrm{Attn}}(K,V,q) - \mathrm{Attn}(K,V,q)\| \leq \varepsilon$. □

C Proofs from Section 2.2

Here we prove a standard lemma which is used in the lower bound about packing nearly orthogonal vectors in high dimensions.

Lemma 2 (Nearly orthogonal vectors)**.** *Consider integer $M \geq 2$ and $\xi \in (0,1)$. Let D be an integer such that $D \geq 6\log(2M)/\xi^2$. There exist M unit vectors $u_1, \ldots, u_M \in \mathbb{R}^D$ such that $|\langle u_i, u_j\rangle| \leq \xi$ for every $i \neq j$.*

Proof. We will prove the result by a standard application of the probabilistic method. For all $i \in [M]$, choose each coordinate of each vector u_i independently and uniformly from $\{-D^{-1/2}, D^{-1/2}\}$. Then, u_i is a unit vector. Note that for any $i \neq j$, $\langle u_i, u_j\rangle$ is the sum of D independent random variables with mean zero and bounded in magnitude by $1/D$. Hoeffding's inequality (Lemma 8) gives $\Pr[|\langle u_i, u_j\rangle| > \xi] \leq 2e^{-D\xi^2/2}$. A union bound over the $\binom{M}{2}$ pairs upper bounds the failure probability by

$$M^2 e^{-D\xi^2/2} \leq M^2 e^{-6\log(2M)/2} = M^2(2M)^{-3} = \frac{1}{8M} < 1.$$

Therefore, such a set of vectors must exist. □

C.1 Proof of Theorem 3

Here we present the proof of the lower bound in the uniform sampling regime. Refer to the main body in Section 2.2 for the construction of key, query, and value vectors.

Proof. Each group will contain $t = \lfloor n/m \rfloor \geq 1$ copies of the same key-value pair, with one additional copy in the first s groups, so that $n = tm + s$ and $0 \leq s < m$. For $i \in [m]$, let $N_i = t + 1$ if $i \leq s$ and $N_i = t$ if $i > s$. The attention instance K, V will have N_i copies of the key-value pair (k_i, v_i). By the setting of t, s, N_i, there will be exactly n key-value pairs. As $t \geq 1$, $N_j/N_i \leq 2$ for all $i, j \in [m]$.

At distinct query q_i, let α_i be the attention mass on the copies of (k_i, v_i). Bounding $\langle k_j, q_i \rangle - \langle k_i, q_i \rangle$ by Equation (9) and Equation (10), we get

$$\alpha_i = \frac{N_i e^{\langle k_i, q_i \rangle / \sqrt{d}}}{\sum_{j=1}^{m} N_j e^{\langle k_j, q_i \rangle / \sqrt{d}}} = \frac{1}{1 + \sum_{j \neq i} \frac{N_j}{N_i} e^{(\langle k_j, q_i \rangle - \langle k_i, q_i \rangle)/\sqrt{d}}} \geq \frac{1}{1 + 2(m-1)e^{-(1-2\xi)(b-a)R^2}}.$$

Recall the condition $m \leq e^{(1-2\xi)(b-a)R^2}/32$. Then, $1 - \alpha_i \leq 2me^{-(1-2\xi)(b-a)R^2} \leq \frac{1}{16}$.

The attention output assigns mass α_i to v_i and total mass $1 - \alpha_i$ to the other value vectors. Since all value vectors have unit norm, the triangle inequality gives

$$\| \operatorname{Attn}(K, V, q_i) - v_i \| \leq (1 - \alpha_i) \max_{j \neq i} \|v_j - v_i\| \leq 2(1 - \alpha_i) \leq \frac{1}{8}. \tag{15}$$

Let H_i be the event that the coreset retains at least one copy of (k_i, v_i). On H_i^c, the nonnegative coreset weights imply that

$$\widehat{\operatorname{Attn}}(K, V, q_i) = \sum_{j \neq i} \lambda_j v_j, \qquad \lambda_j \geq 0, \qquad \sum_{j \neq i} \lambda_j = 1.$$

Since $|\langle v_i, v_j \rangle| \leq 1/8$ by Equation (11) for every $j \neq i$, we obtain

$$\langle v_i, \widehat{\operatorname{Attn}}(K, V, q_i) \rangle = \sum_{j \neq i} \lambda_j \langle v_i, v_j \rangle \leq \frac{1}{8} \sum_{j \neq i} \lambda_j = \frac{1}{8}.$$

Since $\|v_i\| = 1$, the Cauchy-Schwarz inequality and Equation (15) give

$$\langle v_i, \operatorname{Attn}(K, V, q_i) \rangle = 1 + \langle v_i, \operatorname{Attn}(K, V, q_i) - v_i \rangle \geq 1 - \| \operatorname{Attn}(K, V, q_i) - v_i \| \geq \frac{7}{8}.$$

On H_i^c, we have already shown that $\langle v_i, \widehat{\operatorname{Attn}}(K, V, q_i) \rangle \leq 1/8$. Since v_i is a unit vector, another application of the Cauchy-Schwarz inequality gives

$$\|\widehat{\operatorname{Attn}}(K, V, q_i) - \operatorname{Attn}(K, V, q_i)\| \geq \langle v_i, \operatorname{Attn}(K, V, q_i) \rangle - \langle v_i, \widehat{\operatorname{Attn}}(K, V, q_i) \rangle \geq \frac{7}{8} - \frac{1}{8} = \frac{3}{4} > \varepsilon.$$

Hence success at the fixed query q_i implies H_i. Since answering q_i accurately requires retaining a copy of (k_i, v_i), the success guarantee Equation (3) implies $\Pr[H_i] \geq 2/3$ for every $i \in [m]$. Since each retained pair represents at most one group,

$$\mathbb{E}|I| \geq \sum_{i=1}^{m} \Pr[H_i] \geq \frac{2m}{3}.$$

By definition, $\mathbf{OPT}_{\mathsf{Attn}}$ is the optimal worst-case number of tokens retained by an algorithm, so, from above, $\mathbf{OPT}_{\mathsf{Attn}} \geq 2m/3$. □

D Proofs from Section 3.1

Lemma 3 (Approximation guarantee)**.** *Suppose $\|v_i\| \leq 1$ and $s_i \geq \sigma_i$ for every $i \in [n]$. Then Algorithm 1 returns a weighted attention summary of size at most*

$$\min \left\{ n, \left\lceil \frac{96S}{\varepsilon^2} \right\rceil \right\}.$$

For every fixed query $q \in Q$, its estimate satisfies

$$\Pr\left[\left\|\widehat{\mathrm{Attn}}(K,V,q) - \mathrm{Attn}(K,V,q)\right\| \le \varepsilon\right] \ge \frac{2}{3},$$

where the probability is over the sampling in Algorithm 1.

Proof. If the algorithm retains all tokens, the estimate is exact. Otherwise, fix $q \in \mathcal{Q}$ and write

$$Z = \sum_{i=1}^{n} e^{\langle k_i, q\rangle/\sqrt{d}}, \qquad Y = \sum_{i=1}^{n} e^{\langle k_i, q\rangle/\sqrt{d}} v_i.$$

Thus $\mathrm{Attn}(K,V,q) = Y/Z$. The sampled denominator and numerator in (13) are

$$\widehat{Z} = \frac{1}{m}\sum_{\ell=1}^{m} \frac{e^{\langle k_{J_\ell}, q\rangle/\sqrt{d}}}{\pi_{J_\ell}}, \qquad \widehat{Y} = \frac{1}{m}\sum_{\ell=1}^{m} \frac{e^{\langle k_{J_\ell}, q\rangle/\sqrt{d}} v_{J_\ell}}{\pi_{J_\ell}}.$$

Since each index i is sampled with probability π_i, we have $\mathbb{E}\widehat{Z} = Z$ and $\mathbb{E}\widehat{Y} = Y$.
The sensitivity upper bound gives

$$e^{\langle k_i, q\rangle/\sqrt{d}} \le \sigma_i Z \le s_i Z \qquad \text{for every } i \in [n]. \tag{16}$$

Using $\pi_i = s_i/S$,

$$\mathrm{Var}\,\widehat{Z} \le \frac{1}{m}\sum_{i=1}^{n} \frac{e^{2\langle k_i, q\rangle/\sqrt{d}}}{\pi_i} \le \frac{SZ}{m}\sum_{i=1}^{n} e^{\langle k_i, q\rangle/\sqrt{d}} = \frac{S}{m} Z^2.$$

Similarly, using independence and applying Lemma 12 to a centered sample, we obtain

$$\mathbb{E}\|\widehat{Y} - Y\|^2 = \frac{1}{m}\left(\sum_{i=1}^{n} \frac{e^{2\langle k_i, q\rangle/\sqrt{d}}\|v_i\|^2}{\pi_i} - \|Y\|^2\right) \le \frac{1}{m}\sum_{i=1}^{n} \frac{e^{2\langle k_i, q\rangle/\sqrt{d}}}{\pi_i} \le \frac{S}{m} Z^2,$$

where we used $\|v_i\| \le 1$ and (16).

Using Chebyshev's inequality (Lemma 7) for $\widehat{Z}$ and Markov's inequality (Lemma 6) for $\|\widehat{Y} - Y\|^2$, followed by a union bound, we obtain

$$|\widehat{Z} - Z| \le \frac{\varepsilon}{4} Z, \qquad \|\widehat{Y} - Y\| \le \frac{\varepsilon}{4} Z$$

hold together with probability at least $1 - \frac{32S}{m\varepsilon^2} \ge \frac{2}{3}$, since $m = \lceil 96 S/\varepsilon^2 \rceil$.

On this event, $\widehat{Z} \ge (1 - \varepsilon/4)Z$. Also, $\|Y\| \le Z$ because $\|v_i\| \le 1$. Therefore,

$$\left\|\frac{\widehat{Y}}{\widehat{Z}} - \frac{Y}{Z}\right\| \le \frac{\|\widehat{Y} - Y\|}{\widehat{Z}} + \frac{\|Y\|\,|\widehat{Z} - Z|}{Z\widehat{Z}} \le \frac{\varepsilon/2}{1 - \varepsilon/4} \le \varepsilon,$$

where the last inequality uses $0 < \varepsilon < 1$. The algorithm retains at most $\min\{n, m\}$ tokens, which gives the stated size bound. □

Lemma 4 (Total sensitivity bound)**.** *Suppose $r_k, r_q > 0$ and $\|k_i\| \le r_k$ for every i, and let $\mathcal{Q}$ be a nonempty compact query set such that*

$$\mathcal{Q} := \left\{q \in \mathbb{R}^d : \|q\| \le r_q,\ \ell \le \langle k_i, q\rangle/\sqrt{d} \le h \quad \forall i \in [n]\right\}.$$

For sensitivities σ_i defined in Equation (12) with the supremum taken over $q \in \mathcal{Q}$, we have

$$\sum_{i=1}^{n} \sigma_i \le \exp\left(\min\left\{h - \ell, \frac{R^2 + h}{2}\right\}\right). \tag{14}$$

Proof. In this proof, we replace each key k_i by $\sqrt{r_q/r_k}\,k_i$ and each query q by $\sqrt{r_k/r_q}\,q$. This preserves all inner products and hence all sensitivities, while giving both keys and queries the common norm bound $\sqrt{r_k r_q} = R\,d^{1/4}$. We use the same notation for the rescaled keys and query set.

Since $\mathcal{Q}$ is compact and the attention weights are continuous, for each $i \in [n]$ there exists a query $q_i \in \mathcal{Q}$ attaining the sensitivity of token i. We let q_i be a maximizing query for token i; these $q_1, \ldots, q_n$ need not be the same.

Define

$$\lambda_i = \sigma_i / \sum_{j=1}^{n} \sigma_j, \qquad \bar{k} = \sum_{i=1}^{n} \lambda_i k_i, \qquad \bar{q} = \sum_{i=1}^{n} \lambda_i q_i.$$

For each i, applying Jensen's inequality (Lemma 9, with each x_j set to $e^{\langle k_j, q_i\rangle/\sqrt{d}}$) gives

$$\log\left(\sum_{j=1}^{n} e^{\langle k_j, q_i\rangle/\sqrt{d}}\right) = \log\left(\sum_{j=1}^{n} \lambda_j \frac{e^{\langle k_j, q_i\rangle/\sqrt{d}}}{\lambda_j}\right) \geq \sum_{j=1}^{n} \lambda_j \log \frac{e^{\langle k_j, q_i\rangle/\sqrt{d}}}{\lambda_j} = \frac{\langle \bar{k}, q_i\rangle}{\sqrt{d}} - \sum_{j=1}^{n} \lambda_j \log \lambda_j. \tag{17}$$

By the definition of q_i, $\sigma_i = e^{\langle k_i, q_i\rangle/\sqrt{d}} / \sum_{j=1}^{n} e^{\langle k_j, q_i\rangle/\sqrt{d}}$. Combining this with $\sigma_i/\lambda_i = \sum_{j=1}^{n} \sigma_j$, we have

$$\log\left(\sum_{i=1}^{n} \sigma_i\right) = \sum_{i=1}^{n} \lambda_i \left(\frac{\langle k_i, q_i\rangle}{\sqrt{d}} - \log\left(\sum_{j=1}^{n} e^{\langle k_j, q_i\rangle/\sqrt{d}}\right) - \log \lambda_i\right).$$

Applying Equation (17) to the right-hand side,

$$\log\left(\sum_{i=1}^{n} \sigma_i\right) \leq \sum_{i=1}^{n} \lambda_i \frac{\langle k_i, q_i\rangle}{\sqrt{d}} - \frac{\langle \bar{k}, \bar{q}\rangle}{\sqrt{d}} = \sum_{i=1}^{n} \lambda_i \frac{\langle k_i - \bar{k}, q_i - \bar{q}\rangle}{\sqrt{d}}. \tag{18}$$

We bound this expression in two ways. First, $\sum_i \lambda_i \langle k_i, q_i\rangle/\sqrt{d} \leq h$. Also, $\langle \bar{k}, \bar{q}\rangle/\sqrt{d} = \sum_{i,j} \lambda_i \lambda_j \langle k_i, q_j\rangle/\sqrt{d} \geq \ell$, since every q_j belongs to $\mathcal{Q}$. Hence

$$\log\left(\sum_{i=1}^{n} \sigma_i\right) \leq h - \frac{\langle \bar{k}, \bar{q}\rangle}{\sqrt{d}} \leq h - \ell. \tag{19}$$

Second, centering the vectors (Lemma 12) gives $\sum_i \lambda_i \|k_i - \bar{k}\|^2 = \sum_i \lambda_i \|k_i\|^2 - \|\bar{k}\|^2 \leq R^2\sqrt{d} - \|\bar{k}\|^2$, and the same bound holds for the queries. Applying Cauchy-Schwarz (Lemma 10) to (18), we obtain

$$\begin{aligned}
\log\left(\sum_{i=1}^{n} \sigma_i\right) &\leq \frac{1}{\sqrt{d}} \sqrt{\left(\sum_i \lambda_i \|k_i - \bar{k}\|^2\right)\left(\sum_i \lambda_i \|q_i - \bar{q}\|^2\right)} \\
&\leq \frac{1}{\sqrt{d}} \sqrt{(R^2\sqrt{d} - \|\bar{k}\|^2)(R^2\sqrt{d} - \|\bar{q}\|^2)} \leq R^2 - \frac{\|\bar{k}\|^2 + \|\bar{q}\|^2}{2\sqrt{d}} \leq R^2 + \frac{\langle \bar{k}, \bar{q}\rangle}{\sqrt{d}}.
\end{aligned}$$

The last two inequalities follow from the arithmetic-geometric mean inequality (Lemma 11) and Young's inequality for inner products, $2\langle \bar{k}, \bar{q}\rangle \geq -\|\bar{k}\|^2 - \|\bar{q}\|^2$, respectively. Averaging this bound with $\log(\sum_{i=1}^{n} \sigma_i) \leq h - \langle \bar{k}, \bar{q}\rangle/\sqrt{d}$ from (19) gives

$$\log\left(\sum_{i=1}^{n} \sigma_i\right) \leq \frac{R^2 + h}{2}.$$

Together with $\log(\sum_{i=1}^{n} \sigma_i) \leq h - \ell$, this proves the lemma.

□

Lemma 5 (Computing sampling scores). *Let $-1 \leq a < b \leq 1$ and $r_k, r_q > 0$, and write $R^2 = \frac{r_k r_q}{\sqrt{d}}$, $\kappa = \min\left\{b - a, \frac{1+b}{2}\right\}$. Suppose $\|k_i\| \leq r_k$ for every $i \in [n]$, and let*

$$\mathcal{Q} := \left\{q \in \mathbb{R}^d : \|q\| \leq r_q, \quad a\|k_i\|\|q\| \leq \langle k_i, q\rangle \leq b\|k_i\|\|q\| \quad \forall i \in [n]\right\}.$$

For sensitivities σ_i defined in Equation (12) with the supremum taken over $q \in \mathcal{Q}$, we can compute sampling scores s_i satisfying $s_i \geq \sigma_i$ for every $i \in [n]$, and

$$\sum_{i=1}^{n} s_i \leq \frac{128e^2}{\kappa^2} e^{\kappa R^2}.$$

Moreover, such sampling scores can be obtained by solving $n \cdot \max\{1, \lceil R^2 \rceil\}$ convex optimization problems, using only the keys and the parameters r_k, r_q, a, b.

Proof. We group keys and queries by norm, compute sampling scores within each pair of groups by convex optimization, and then bound the sum of these scores using Lemma 4.

Partitioning by norm. Set $M = \max\{1, \lceil R^2 \rceil\}$. Partition the keys into norm bins

$$I_1 = \left\{ i \in [n] : \|k_i\| \leq \frac{r_k}{M} \right\}, \qquad I_u = \left\{ i \in [n] : \frac{(u-1)r_k}{M} < \|k_i\| \leq \frac{ur_k}{M} \right\} \quad (u = 2, \ldots, M).$$

Similarly, partition the possible query norms into M intervals of width r_q/M, assigning each endpoint to the interval on its left and including zero in the first interval.

For $u, v \in [M]$, denote

$$A_{uv} = \frac{R^2(u-1)(v-1)}{M^2}, \quad B_{uv} = \frac{R^2 uv}{M^2}, \quad \ell_{uv} = \min\{aA_{uv}, aB_{uv}\}, \quad h_{uv} = \max\{bA_{uv}, bB_{uv}\}.$$

If $i \in I_u$ and an admissible query q has norm in the v-th interval, then

$$A_{uv} \leq \frac{\|k_i\|\|q\|}{\sqrt{d}} \leq B_{uv}, \qquad \ell_{uv} \leq \frac{\langle k_i, q \rangle}{\sqrt{d}} \leq h_{uv}.$$

Thus, for each nonempty I_u, the compact convex domain

$$\widetilde{\mathcal{Q}}_{uv} = \left\{ z \in \mathbb{R}^d : \|z\| \leq \frac{vr_q}{M}, \quad \ell_{uv} \leq \frac{\langle k_j, z \rangle}{\sqrt{d}} \leq h_{uv} \quad \forall j \in I_u \right\}$$

contains every admissible query whose norm lies in the v-th interval.

Bounding sensitivities locally. For $i \in I_u$, define the local sensitivity

$$\sigma_i^{uv} = \max_{z \in \widetilde{\mathcal{Q}}_{uv}} \frac{e^{\langle k_i, z \rangle / \sqrt{d}}}{\sum_{j \in I_u} e^{\langle k_j, z \rangle / \sqrt{d}}},$$

with $\sigma_i^{uv} = 0$ when $\widetilde{\mathcal{Q}}_{uv}$ is empty. For an admissible query q whose norm lies in the v-th interval,

$$\frac{e^{\langle k_i, q \rangle / \sqrt{d}}}{\sum_{j=1}^{n} e^{\langle k_j, q \rangle / \sqrt{d}}} \leq \frac{e^{\langle k_i, q \rangle / \sqrt{d}}}{\sum_{j \in I_u} e^{\langle k_j, q \rangle / \sqrt{d}}} \leq \sigma_i^{uv}.$$

Consequently, $\sigma_i \leq \sum_{v=1}^{M} \sigma_i^{uv}$.

Computing the sampling scores. For $i \in I_u$, define

$$f_i^u(z) := \log \left(\sum_{j \in I_u} e^{\langle k_j - k_i, z \rangle / \sqrt{d}} \right).$$

By Lemma 13, this objective is convex and continuously differentiable, with an explicit gradient. Whenever $\widetilde{Q}_{uv}$ is nonempty, it is the logarithm of the reciprocal of token i's attention weight within I_u, so

$$\min_{z \in \widetilde{Q}_{uv}} f_i^u(z) = -\log \sigma_i^{uv}.$$

Moreover, for every $z \in \widetilde{Q}_{uv}$, $0 \leq f_i^u(z) \leq \log |I_u| + h_{uv} - \ell_{uv}$. The lower bound follows from the term $j = i$, and the upper bound follows from the defining inner-product bounds of $\widetilde{Q}_{uv}$.

By Lemma 14 and the definition of $\widetilde{Q}_{uv}$, each nonempty domain is compact and convex. Applying Lemma 15 with $\eta = \log 2$, we obtain $z_i^{uv} \in \widetilde{Q}_{uv}$ satisfying

$$-\log \sigma_i^{uv} \le f_i^u(z_i^{uv}) \le -\log \sigma_i^{uv} + \log 2. \tag{20}$$

Set $s_i^{uv} := 2e^{-f_i^u(z_i^{uv})}$. Exponentiating Equation (20) gives $\sigma_i^{uv} \le s_i^{uv} \le 2\sigma_i^{uv}$. For empty domains, set $s_i^{uv} := 0$, so the same inequalities hold under the convention $\sigma_i^{uv} = 0$. We are now ready to describe the sampling scores s_i. For $i \in I_u$, define

$$s_i := \sum_{v=1}^{M} s_i^{uv}.$$

Together with the local sensitivity bound, this gives $\sigma_i \le \sum_{v=1}^{M} \sigma_i^{uv} \le s_i \le 2\sum_{v=1}^{M} \sigma_i^{uv}$. Since $0 \in Q$, we have $\sigma_i \ge 1/n$, so the scores are positive.

Bounding the sum within each pair of bins. Write $D_{uv} = B_{uv} - A_{uv} = \frac{R^2(u+v-1)}{M^2} \le \frac{2R^2}{M} \le 2$. For $x_+ = \max\{x, 0\}$, the endpoint definitions give

$$\ell_{uv} = aB_{uv} - a_+ D_{uv}, \qquad h_{uv} = bB_{uv} + (-b)_+ D_{uv}.$$

Since $a_+ + (-b)_+ \le 1$,

$$h_{uv} - \ell_{uv} \le (b-a)B_{uv} + 2, \qquad \frac{B_{uv} + h_{uv}}{2} \le \frac{1+b}{2} B_{uv} + 1.$$

The keys in I_u have norm at most ur_k/M, and queries in $\widetilde{\mathcal{Q}}_{uv}$ have norm at most vr_q/M. The product of these bounds, divided by $\sqrt{d}$, is B_{uv}. For each nonempty domain, Lemma 4 therefore gives

$$\sum_{i \in I_u} \sigma_i^{uv} \le \exp\left(\min\left\{h_{uv} - \ell_{uv}, \frac{B_{uv} + h_{uv}}{2}\right\}\right) \le e^{\kappa B_{uv} + 2}.$$

The same bound holds trivially for empty domains. Hence

$$\sum_{i=1}^{n} s_i \le 2e^2 \sum_{u=1}^{M} \sum_{v=1}^{M} e^{\kappa B_{uv}}.$$

Summing over the bins. If $R^2 < 1$, then $M = 1$, and the desired bound follows immediately. Otherwise, $M = \lceil R^2 \rceil \le 2R^2$. Since $u, v \le M$,

$$B_{uv} \le \frac{R^2(u+v)}{2M} = R^2 - \frac{R^2}{2M}\big((M-u) + (M-v)\big).$$

Using $R^2/(2M) \ge 1/4$, we obtain

$$\sum_{u=1}^{M} \sum_{v=1}^{M} e^{\kappa B_{uv}} \le e^{\kappa R^2} \left(\sum_{t=0}^{M-1} e^{-\kappa t/4}\right)^2 \le \frac{e^{\kappa R^2}}{(1 - e^{-\kappa/4})^2}.$$

Finally, $0 < \kappa \le 1$ and $1 - e^{-\kappa/4} \ge \kappa/8$, so

$$\sum_{i=1}^{n} s_i \le \frac{128e^2}{\kappa^2} e^{\kappa R^2}.$$

Each key participates in at most M optimization problems, giving at most nM such problems in total. □

E PROOFS FROM SECTION 3.2

In this section, we give the proof of Theorem 4, the lower bound in the importance sampling regime.

Theorem 4 (Lower bound, Item 2). *Consider $r_k, r_q > 0$, $\varepsilon \in (0, 1/2]$, and $-1 \le a < b \le 1$ in the importance sampling regime (Definition 1). Consider n and $d = \omega(\log n)$, and let $\xi = \sqrt{\frac{6\log(2n)}{d-3}} = o(1)$. Pick any integer m satisfying*

$$2 \le m \le n, \qquad m \le \min\left\{e^{\frac{1+b}{2}R^2}, e^{(b-a)R^2}\right\} \cdot e^{-2\xi R^2}/32. \tag{21}$$

There is a fixed dataset (K, V) of n key-value pairs, with keys of norm r_k and unit-norm values in $\mathbb{R}^d$, together with queries $q_1, \ldots, q_m$ of norm r_q, such that every key $k \in K$ satisfies

$$a \le \frac{\langle k, q_i\rangle}{\|k\|\|q_i\|} \le b \qquad \text{for all } i \in [m].$$

For this dataset, satisfying Equation (3) at each of these queries requires $\mathbf{OPT}_{\mathsf{Attn}} \ge 2m/3$.

The construction and analysis are essentially the same as that in the uniform sampling regime Section 2.2 with slightly more complicated parameter settings in order to capture all possible values of a, b.

Goal. We construct keys and queries such that each matching pair (k_i, q_i) has scaled inner product bR^2, exceeding that of every nonmatching pair (k_j, q_i), $j \ne i$, by at least $(1 - o(1)) \min\{b - a, (1 + b)/2\}R^2$, where $R^2 = r_k r_q/\sqrt{d}$. We will then use this gap to make each query concentrate its attention on copies of its matching key, forcing an accurate summary to retain many key groups.

Construction. Let e_0, e_K, e_Q be the first three standard basis vectors. Choose m unit vectors satisfying Lemma 2 in dimension $D = d - 3$, with pairwise inner products bounded in magnitude by $\xi = \sqrt{\frac{6\log(2n)}{D}}$. Embed them into the last $d - 3$ coordinates of $\mathbb{R}^d$, and denote the resulting vectors by $u_1, \ldots, u_m$. Define

$$\begin{aligned} k_i &= r_k\Big(\underbrace{\sqrt{\alpha}\, u_i}_{\text{Shared group direction}} + \underbrace{\sqrt{|\beta|}\, e_0}_{\text{Anisotropic shift}} + \underbrace{\sqrt{1-\alpha-|\beta|}\, e_K}_{\text{Norm padding}} \Big), \\ q_i &= r_q\Big(\underbrace{\sqrt{\alpha}\, u_i}_{\text{Shared group direction}} - \underbrace{\operatorname{sgn}(\beta)\sqrt{|\beta|}\, e_0}_{\text{Anisotropic shift}} + \underbrace{\sqrt{1-\alpha-|\beta|}\, e_Q}_{\text{Norm padding}} \Big), \end{aligned} \tag{22}$$

where we set

$$\alpha = \min\left\{\frac{1+b}{2}, \frac{b-a}{1+\xi}\right\}, \qquad \beta = \alpha - b = \min\left\{\frac{1-b}{2}, \frac{-a-\xi b}{1+\xi}\right\}. \tag{23}$$

Note that $\alpha \ge 0$ as $b \ge -1$ and $b \ge a$. Note that $\alpha + |\beta| \le 1$ as $\alpha + \beta = 2\alpha - b \le 2\left(\frac{1+b}{2}\right) - b = 1$ and $\alpha - \beta = b \le 1$. Therefore, $1 - \alpha - |\beta| \ge 0$ and the keys and queries are well defined with norms exactly equal to $\|k_i\| = r_k$ and $\|q_i\| = r_q$. The components along u_i and e_0 serve the purpose of achieving the desired inner products while the padding along e_K and e_Q increases the norms of the keys/queries to fix the denominator of cosine similarity.

Verifying the inner products. It remains to bound the scaled inner products between queries and keys. For any $i \in [m]$, the inner product between the matching pair is:

$$\frac{\langle k_i, q_i\rangle}{\sqrt{d}} = (\alpha - \beta)\, R^2 = bR^2.$$

For any non-matching pair $i \ne j$,

$$(b - \alpha\,(1 + \xi))\, R^2 \le \frac{\langle k_i, q_j\rangle}{\sqrt{d}} = (\alpha\langle u_i, u_j\rangle - \beta)\, R^2 \le bR^2.$$

The second choice in the minimum defining α ensures $b - \alpha\,(1 + \xi) \ge a$. Thus, every key-query cosine similarity belongs to $[a, b]$, since $R^2\sqrt{d} = r_k r_q = \|k_i\|\|q_j\|$.

For any $i \neq j$, the scaled inner product between key-query pairs at the same index exceeds that of any other key-query pair by

$$\frac{\langle k_i, q_i\rangle - \langle k_j, q_i\rangle}{\sqrt{d}} \geq bR^2 - (\alpha\xi - \beta)\,R^2 = \alpha\,(1-\xi)\,R^2.$$

By the definition of α,

$$\frac{\langle k_i, q_i\rangle - \langle k_j, q_i\rangle}{\sqrt{d}} \geq \min\left\{b-a, \frac{1+b}{2}\right\} R^2 \frac{1-\xi}{1+\xi} \geq (1-2\xi)\min\left\{b-a, \frac{1+b}{2}\right\} R^2. \quad (24)$$

The last inequality uses $(1-x)/(1+x) \geq 1-2x$ for $x \geq 0$.

Constructing the value vectors. Apply Lemma 2 a second time to obtain m unit value vectors $v_1, \ldots, v_m \in \mathbb{R}^d$ satisfying

$$|\langle v_i, v_j\rangle| \leq \frac{1}{8} \qquad \text{for all } i, j \in [m] \text{ with } i \neq j. \quad (25)$$

Since $m \leq n$, the assumption $d = \omega(\log n)$ ensures that this is possible.

E.1 Full construction

Proof of Theorem 4. Choose $t = \lfloor n/m \rfloor \geq 1$ and $0 \leq s < m$ such that $n = tm + s$. For $i \in [m]$, let $N_i = t+1$ if $i \leq s$ and $N_i = t$ if $i > s$. The attention instance K, V will have N_i copies of the key-value pair (k_i, v_i). By the setting of t, s, N_i, there will be exactly n key-value pairs. As $t \geq 1$, $N_j/N_i \leq 2$ for all $i, j \in [m]$.

At distinct query q_i, let α_i be the attention mass on the copies of (k_i, v_i). By Equation (24),

$$\alpha_i = \frac{N_i e^{\langle k_i, q_i\rangle/\sqrt{d}}}{\sum_{j=1}^m N_j e^{\langle k_j, q_i\rangle/\sqrt{d}}} = \frac{1}{1 + \sum_{j\neq i} \frac{N_j}{N_i} e^{(\langle k_j, q_i\rangle - \langle k_i, q_i\rangle)/\sqrt{d}}} \geq \frac{1}{1 + 2(m-1)e^{-(1-2\xi)\min\{b-a,(1+b)/2\}R^2}}.$$

Recall the condition $m \leq e^{(1-2\xi)\min\{b-a,(1+b)/2\}R^2}/32$. Then,

$$1 - \alpha_i \leq 2me^{-(1-2\xi)\min\{b-a,(1+b)/2\}R^2} \leq \frac{1}{16}.$$

So, attention focuses significant mass on v_i:

$$\|\operatorname{Attn}(K, V, q_i) - v_i\| \leq 2(1-\alpha_i) \leq \frac{1}{8}. \quad (26)$$

The first inequality follows because all value vectors have norm one, so their distance from v_i is at most two.

Let H_i be the event that the coreset retains at least one copy of (k_i, v_i). On the complementary event, the approximation at q_i is a convex combination of the value vectors v_j with $j \neq i$. This follows directly from the nonnegative weights in the definition of query-oblivious weighted token-eviction. Therefore

$$\langle v_i, \widehat{\operatorname{Attn}}(K, V, q_i)\rangle \leq \frac{1}{8}$$

as all distinct value vectors are nearly orthogonal.

On the other hand, Equation (26) gives $\langle v_i, \operatorname{Attn}(K, V, q_i)\rangle \geq 1 - 1/8$. Projecting the error onto the unit vector v_i yields, whenever H_i does not hold,

$$\begin{aligned}\|\widehat{\operatorname{Attn}}(K, V, q_i) - \operatorname{Attn}(K, V, q_i)\| &\geq |\langle v_i, \widehat{\operatorname{Attn}}(K, V, q_i)\rangle - \langle v_i, \operatorname{Attn}(K, V, q_i)\rangle| \\ &\geq 1 - \frac{1}{8} - \frac{1}{8} = \frac{3}{4} > \varepsilon.\end{aligned}$$

Hence success at the fixed query q_i implies H_i.

The per-query guarantee of a query-oblivious weighted token-eviction policy now gives $\Pr[H_i] \geq 2/3$ for each i separately. Since each retained pair represents at most one group,

$$\mathbb{E}|I| \geq \sum_{i=1}^m \Pr[H_i] \geq \frac{2m}{3}.$$

By definition, $\mathbf{OPT}_{\mathsf{Attn}}$ is the optimal worst-case number of tokens retained by an algorithm, so, from above, $\mathbf{OPT}_{\mathsf{Attn}} \geq 2m/3$. □

F EXPERIMENTS

F.1 Q/K COSINE SIMILARITY

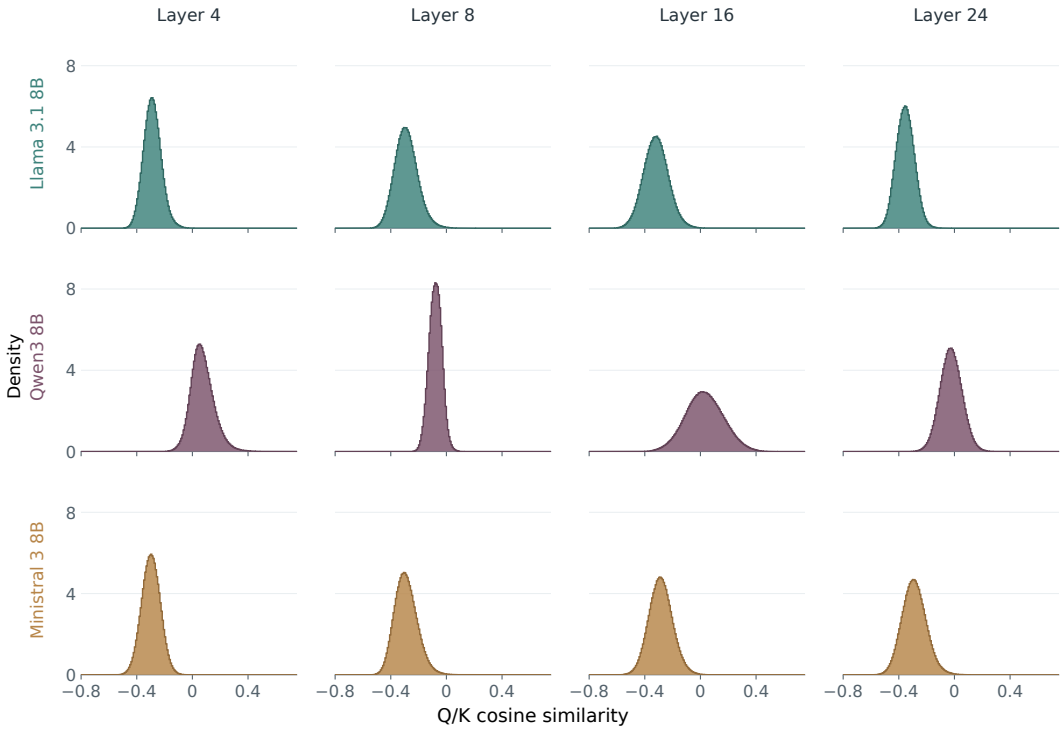


Figure 6: **Query–key cosine similarities across models and attention layers.** Rows show Llama 3.1 8B, Qwen3 8B, Ministral 3 8B; columns show layers 4, 8, 16, 24. All panels use the same prepared LongBench v2 example. Each panel uses an input of 8,192 tokens and compares the final 256 queries from query head 0 against 7,935 earlier-context keys from the corresponding KV head 0. The initial key is excluded. Cosines are computed after model normalization and RoPE.

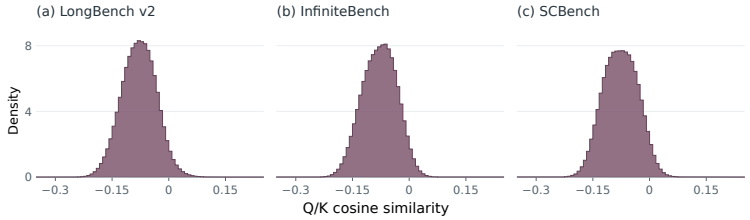


Figure 7: **Query–key cosine similarities across inputs.** The plot uses Qwen3 8B at layer 8 on (a) LongBench v2, (b) InfiniteBench, (c) SCBench. Each panel uses an input of 8,192 tokens and compares the final 256 queries from query head 0 against 7,935 earlier-context keys from the corresponding KV head 0. The initial key (attention sink) is excluded. Cosines are computed after model normalization and RoPE.

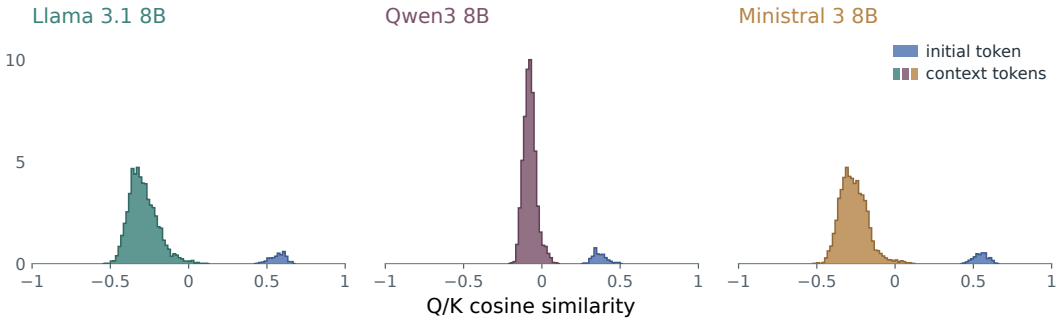


Figure 8: **Query–key cosine similarities with initial token (attention sink)** Qwen3 8B at layer 8 on LongBench v2 example input of 8,192 tokens and compares the final 256 queries from query head 0 against a sample of 16 earlier-context keys from the corresponding KV head 0 and the initial key (attention sink). Cosines are computed after model normalization and RoPE.

F.2 KEY/QUERY NORMS

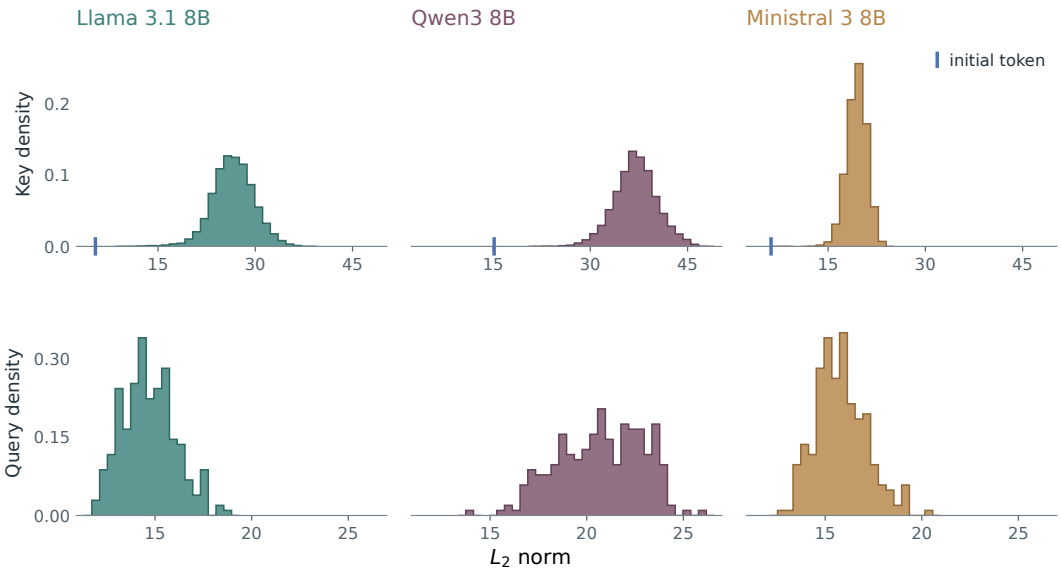


Figure 9: **Key and query magnitudes across models.** Layer 8 on the LongBench v2 input (8,192 tokens). Columns show Llama 3.1 8B, Qwen3 8B, Ministral 3 8B. The top row uses 7,935 keys per model: all eligible earlier-context keys. Blue baseline marks indicate the initial key (attention sink) norm in each model. The bottom row uses the final 256 queries.