# Positional LSH: Binary Block Matrix Approximation for Attention with Linear Biases

**Daniel Wolfson**
Tel Aviv University
wolfson2@mail.tau.ac.il

**Tal Wagner**
Tel Aviv University
talwag@tauex.tau.ac.il

## Abstract

Positional encoding in transformers is commonly implemented through positional embeddings, attention masks, or bias terms, but formal connections between these mechanisms remain limited. We study attention with positional bias through the lens of locality-sensitive hashing (LSH), focusing on Attention with Linear Biases (ALiBi). We show that the ALiBi bias matrix is the expectation of contiguous block-diagonal binary masks induced by a "positional LSH" scheme. The empirical mean of masks sampled from this scheme yields spectral norm and max-norm approximation guarantees with bounded block sizes with high probability. This structural theorem implies an efficient approximation algorithm for ALiBi-biased attention: with high probability over the sampled masks, the approximate attention output is accurate simultaneously for all query-key-value inputs and can be computed in near-linear time in the context length, reducing long-context ALiBi to a collection of randomized short-context regular (positionally unbiased) attention operations. Conceptually, this connects positional bias, masks, and positional embeddings in a single formal framework and suggests an approach to efficient ALiBi-biased attention. Experiments on large language models validate our theoretical findings.

## 1 Introduction

Although the Transformer architecture has remained remarkably intact since its introduction in [41], even through widespread adoption in large language models, encoding token positions stands out as a still evolving aspect. The challenge is to encode how the structural positions of tokens in the input inform the way they attend to each other. A closely intertwined challenge is long-context attention: attending to all tokens in a long context becomes computationally difficult or infeasible, whereas limiting attention to structurally local windows often degrades long-range reasoning performance.

Three leading paradigms in positional encoding, often used in conjunction, are as follows:

- *Positional embedding*, where each position in the input is endowed with an embedding vector designed to encode its location – e.g., sinusoidal [41] or rotary position embeddings (RoPE) [38];
- *Attention with masks*, which allows tokens to attend to some tokens but not to others through a fixed binary mask – e.g., sliding-window attention [5, 24];
- *Attention with bias*, which modifies attention weights a posteriori according to token positions – e.g., Attention with Linear Biases (ALiBi) [32].

In this work, we undertake a theoretical study of attention with positional bias, focusing on ALiBi, a prominent method that has emerged from empirical research [32] and has been widely implemented and adopted in popular models [12, 43]. We study it through the lens of *locality sensitive hashing* (LSH), a central paradigm in the theory of efficient algorithms. This leads us to uncover fundamental connections between ALiBi and LSH with implications to efficient long-context attention.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Our main structural result, Theorem 2.1, shows that the ALiBi bias matrix can be approximated arbitrarily well by a linear combination of binary contiguous block-diagonal matrices. Conceptually, this draws a formal connection between the three main paradigms in positional encoding: bias, masks and embeddings. Namely, it suggests that attention with bias can be approximated by attention with masks, where the masks are induced by locality-sensitive hash collisions of positional embeddings. In the case of ALiBi, this connection holds in a strong formal sense.

Our main algorithmic result, Theorem 2.2, is an efficient approximation algorithm for ALiBi-biased attention over long contexts, which is based on our structural result. It suggests an approach to long-context ALiBi through a reduction to a collection of randomized short-context regular (positionally unbiased) attention operations, which correspond to the contiguous binary blocks in Theorem 2.1. This yields a near-linear time algorithm for ALiBi with a provable approximation guarantee, addressing the computational bottleneck in long-context ALiBi, a challenge called out in [32].

Conceptually, our contribution goes beyond applying classical LSH to positions in attention: it is in reinterpreting positional bias through hashing on positions, drawing a formal connection between positional embeddings, binary attention masks, and bias terms. Technically, while kernel approximation via LSH is standard, we introduce novel technical tools in order to prove matrix convergence in the spectral norm and high-probability block size control. These results are specialized to attention with ALiBi and do not follow from generic LSH machinery alone. Our proofs draw on results from matrix concentration theory, Fourier analysis of Toeplitz matrices, and sub-gamma tail bounds.

We validate our theoretical findings through experiments on publicly available large language models.

## 2 Our Results in Detail

We start by fixing notation to be used throughout. We use "$\odot$" for element-wise matrix product. We denote the following norms for a matrix $M$: $\|M\|$ for the spectral norm; $\|M\|_{\max} = \max_{ij} |M_{ij}|$; and $\|M\|_{2,\infty} = \max_i \sqrt{\sum_j M_{ij}^2}$. We let $D^{[M]}$ denote the diagonal row-sum matrix $D^{[M]}_{ii} = \sum_{j=1}^n M_{ij}$. Finally, we will use $J \in \mathbb{R}^{n\times n}$ for the fixed lower-triangular all-1 matrix, $J_{ij} = \mathbf{1}\{j \le i\}$.

**Attention.** Let $n$ be the attention context length. Let $Q, K \in \mathbb{R}^{n\times d}$ be the key and query matrices with rows $\{q_i\}, \{k_j\}$ respectively. The unnormalized dot-product attention matrix $A \in \mathbb{R}^{n\times n}$ has entries $A_{ij} = \exp(q_i^T k_j/\sqrt{d})$. The row-wise normalized non-causal attention matrix is $P = (D^{[A]})^{-1}A$. In the causal attention case we instead have $P = (D^{[A\odot J]})^{-1}(A \odot J)$. Given a value matrix $V \in \mathbb{R}^{n\times d'}$, the output of the attention operation without positional bias is $T = PV$.

**ALiBi.** The ALiBi matrix $L^\star$ is defined in [32] as $L^\star_{ij} = e^{-|i-j|/\sigma}$, where $\sigma > 0$ is a bandwidth hyperparameter which is fixed per head and varies between heads. The unnormalized ALiBi-biased attention weights are $A^\star = A \odot L^\star$ in the non-causal case and $A^\star = A \odot J \odot L^\star$ in the causal case. In either case, the normalized ALiBi-biased attention matrix is $P^\star = (D^{[A^\star]})^{-1}A^\star$, and the goal is to compute the output $T^\star = P^\star V$.

The following theorem is our main structural result: an approximation of the ALiBi matrix by random contiguous block-diagonal binary matrices.

**Theorem 2.1.** *There exists a distribution $\mathcal{M}$ over contiguous block-diagonal binary $n \times n$ matrices, where each diagonal block is a square all-1 matrix, such that $\mathbb{E}_{M\sim\mathcal{M}}[M] = L^\star$. Let $M_1, \ldots, M_s \sim \mathcal{M}$ be i.i.d. samples and let $\widetilde{M} = \frac{1}{s}\sum_{i=1}^s M_i$ be their empirical mean. Then we have,*

- *Spectral norm expectation:*

$$\mathbb{E}\|L^\star - \widetilde{M}\| \le C\Psi_\sigma\left(\sqrt{\frac{\log n}{s}} + \frac{\log n}{s}\right) \tag{1}$$

- *Spectral norm concentration:*

$$\forall \varepsilon > 0,\ \ \Pr[\|L^\star - \widetilde{M}\| \ge \varepsilon] \le n \cdot \exp\left(-Cs\min\left\{\frac{\varepsilon^2}{\Psi_\sigma^2}, \frac{\varepsilon}{\Psi_\sigma}\right\}\right) \tag{2}$$

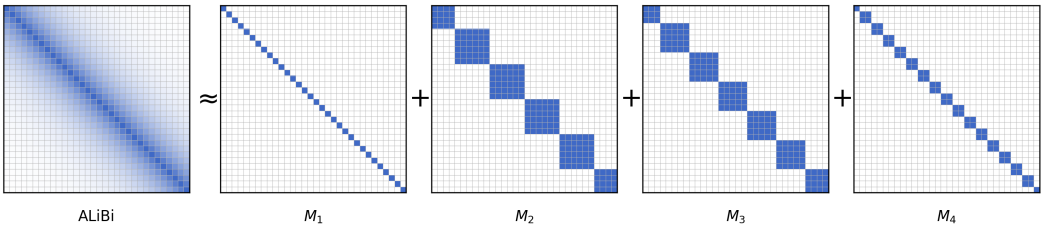


Figure 1: ALiBi approximation with binary block-diagonal matrices by positional LSH.

- *Max-norm concentration:*

$$\forall \varepsilon > 0, \quad \Pr[\|L^\star - \widetilde{M}\|_{\max} \geq \varepsilon] \leq 2n^2 e^{-2s\varepsilon^2} \tag{3}$$

*where* $\Psi_\sigma = 1 + \sigma + \frac{e^{1/\sigma}+1}{e^{1/\sigma}-1}$ *and* $C > 0$ *is a universal constant. Furthermore, let* $b_{\max}$ *be the maximum size of a diagonal block in any of the samples* $M_1, \ldots, M_s$*, then, for all* $\delta \in (0, 1/e)$*,*

$$\Pr[b_{\max} > 1 + \sigma(3\ln(s/\delta) + 2)] < \delta. \tag{4}$$

Structurally, the theorem states that the ALiBi matrix is a convex combination of contiguous non-overlapping local-window attention patterns, each positionally unbiased. Approximation-wise, it implies that with high probability, $s \approx \varepsilon^{-2}\sigma^2 \log n$ samples from $\mathcal{M}$ suffice to approximate ALiBi using diagonal blocks of size at most $b_{\max} \approx \sigma \log\log n$ (suppressing logarithmic factors in $\varepsilon^{-1}, \delta^{-1}, \sigma$ for simplicity). Regarding the values of $\sigma$ and $\Psi_\sigma$, we recall that Press et al. [32] set $\sigma$ as a geometric series from $2^{8/n_h}$ to $2^8$ across $n_h$ heads. Under this setting, $\Psi_\sigma$ satisfies $2 < \Psi_\sigma < 3\sigma + 2$.

Algorithmically, Theorem 2.1 suggests that ALiBi-biased attention can be approximated efficiently over long contexts by a reduction to a collection of small local attention windows. Our next theorem formalizes this intuition and yields an approximation guarantee for the output attention weights.

**Theorem 2.2** (efficient uniform approximation for ALiBi-biased attention)**.** *Let* $\delta \in (0, 1/e)$ *and* $s \geq \log(n/\delta)$*. Let* $\widetilde{M} = \frac{1}{s}\sum_{i=1}^{s} M_i$ *be the empirical mean of s samples from* $\mathcal{M}$ *as in Theorem 2.1. Given an attention instance* $(K, Q, V)$*, let* $A, A^\star, P$ *be the corresponding attention weights as above (in either the causal or non-causal variant), and denote:*

$$\Delta_P = C \min\{\Psi_\sigma \|P\|_{2,\infty}, 1\}\sqrt{\frac{\log(n/\delta)}{s}} \quad \text{and} \quad \beta_P^\star = \min_i \sum_{j=1}^{n} P_{ij} L^\star_{ij}$$

*where* $C > 0$ *is a universal constant. Then, with probability* $1 - \delta$ *over the draw of* $\widetilde{M}$*, the following hold simultaneously for all input attention instances:*

1. *The approximate ALiBi-biased attention output* $\widetilde{T}^\star$*, obtained by replacing* $L^\star$ *with* $\widetilde{M}$*, can be computed in time* $O(nd(\sigma s \log(s/\delta) + 1))$*.*
2. *Whenever* $\Delta_P < \beta_P^\star$*, it holds that* $\|T^\star - \widetilde{T}^\star\|_{\max} \leq \frac{2\Delta_P}{\beta_P^\star - \Delta_P}\|V\|_{\max}$*.*

To unpack the statement, we note that the critical quantity $\beta_P^\star$ is the ratio between the minimum row denominator in $A^\star$ and $A$, that is, in the attention matrix before and after the ALiBi bias. It governs how sensitive the instance is to perturbations in the bias matrix. The term $\Delta_P$ which governs the error decays like $1/\sqrt{s}$ with the number of samples $s$, and may further improve with smaller $\|P\|_{2,\infty}$, which occurs in instances where attention weights are more balanced. Importantly, the convergence in matrix norms in Theorem 2.1 (eqs. (2) and (3)) is what allows the approximation in Theorem 2.2 to hold uniformly — that is, simultaneously for all attention inputs with high probability.

## 3 Positional LSH

In this section we outline our construction and analysis. Omitted proofs appear in Appendix A.

### 3.1 Attention with Bias via Locality Sensitive Hashing

Locality sensitive hashing (LSH) was originally introduced in [23] for metric spaces, with the application of fast nearest neighbor search. A different but closely related form of LSH can be defined for kernel spaces. It originates in [6, 9, 2] and is useful for efficiently estimating similarities for purposes beyond search. This is the notion of LSH we use in this paper.

Let $U$ be a space of objects and $\ker : U \times U \to [0, 1]$ a similarity map (i.e., a kernel) over $U$. An LSH scheme for ker is a distribution $\mathcal{H}$ over hash functions operating on $U$ such that for all $u, u' \in U$

$$\ker(u, u') = \Pr_{h\sim\mathcal{H}}[h(u) = h(u')]. \tag{5}$$

This notion has been widely influential, including recent applications to efficient attention computation over long contexts (see Section 7). While prior work has applied LSH to the token embeddings of the keys and queries in the attention operation, here we will apply it exclusively to their positions in the sequence. Hence, we call this *positional LSH*.

Let $L$ be a positional bias matrix that biases the original attention weights $A$ as $A \odot L$. Let $\{u_i\}_{i=1}^n$ be positional embeddings for the token positions in a context of length $n$. That is, $u_i$ is an object that encodes the location $i$ within the context. Suppose that $L$ has the form $L_{ij} = \ker(u_i, u_j)$. Furthermore, suppose that ker admits an LSH scheme $\mathcal{H}$. We may define a distribution $\mathcal{M}$ over $n \times n$ matrices where a sampled hash function $h \sim \mathcal{H}$ induces the random binary matrix $M^{(h)}$ given by

$$M^{(h)}_{ij} = \begin{cases} 1 & \text{if } h(u_i) = h(u_j) \\ 0 & \text{if } h(u_i) \neq h(u_j). \end{cases}$$

The LSH property (5) immediately implies $L = \mathbb{E}_{h\sim\mathcal{H}}[M^{(h)}]$. Furthermore, $M^{(h)}$ has a distinct structure: its 1-entries are organized in non-overlapping principal combinatorial rectangles. Specifically, for each hash bin $\ell$ in the range of $h$, denote by $I_\ell = \{i : h(u_i) = \ell\}$ the indices that fall into that bin. $M^{(h)}$ has 1's in its $I_\ell \times I_\ell$ principal submatrix for every $\ell$, and 0's elsewhere. Equivalently, the rows and columns of $A$ can be permuted to make $M^{(h)}$ a block-diagonal matrix in which all blocks are all-1 submatrices. Note that without a permutation, the blocks $I_\ell$ need not necessarily be contiguous; it could be, for example, that $h(u_1) = h(u_3) \neq h(u_2)$, in which case indices 1 and 3 are hashed into the same bin while 2 is hashed into a different bin.

The identity $L = \mathbb{E}_{h\sim\mathcal{H}}[M^{(h)}]$ naturally suggests that $L$ can be approximated by the empirical average of i.i.d. samples of $M^{(h)}$. See Figure 1 for illustration. This raises the question of the convergence rate to the mean. Entry-wise scalar convergence follows immediately from standard Chernoff-Hoeffding concentration, and this already yields the max-norm bound eq. (3) in Theorem 2.1. Matrix convergence in the spectral norm (eqs. (1) and (2) in Theorem 2.1) is considerably more intricate and may not hold in general (see Section 5). We will prove it for ALiBi with a specific positional LSH scheme.

### 3.2 Positional LSH for ALiBi via Random Binning Features

ALiBi biases attention weights according to the similarity measure $L^\star_{ij} = \ker(i, j) = e^{-|i-j|/\sigma}$ between token positions $i, j$. This is the familiar one-dimensional Laplacian kernel. Note that the positional embedding object of each index $i$ is simply its integer value, $u_i = i$.

LSH schemes for the Laplacian kernel are well-known [35, 1, 4]. We will use the one due to Rahimi and Recht [35], which they termed Random Binning Features.

**Definition 3.1** ([35])**.** *The Random Binning Features (RBFs) LSH scheme over $\mathbb{R}$ is defined as follows. To sample $h \sim \mathcal{H}$, first draw $b \sim \Gamma(2, \sigma)$ from the Gamma distribution with shape 2 and scale $\sigma$. It is supported on $[0, \infty)$ with density $p(b) = \sigma^{-2}be^{-b/\sigma}$. Then sample $c$ uniformly from $[0, b]$. The hash function $h = h_{b,c}$ partitions $\mathbb{R}$ into length-$b$ segments $\{[b\ell + c, b\ell + b + c) : \ell \in \mathbb{Z}\}$ and hashes each number into the segment that contains it. Formally, $h(u) = \lfloor \frac{u-c}{b} \rfloor$ for every $u \in \mathbb{R}$.*

**Lemma 3.2** ([35])**.** *RBFs satisfy eq. (5) for the Laplacian kernel $\ker(u, u') = e^{-|u-u'|/\sigma}$ over $\mathbb{R}$.*

Therefore, RBFs form a positional LSH scheme for ALiBi. Although RBFs for the Laplacian kernel are classical, our use of them will not be a black-box application: the nontrivial step will be to analyze their interaction particularly with the space of token indices and achieve matrix approximation for ALiBi-biased attention. To this end, we will require specific properties of their underlying Gamma distribution, summarized in the following fact.

**Fact 3.3.** *A Gamma random variable $Z \sim \Gamma(2, \sigma)$ has moments $\mathbb{E}[Z^k] = \sigma^k(k+1)!$ for all $k \in \mathbb{N}$ and moment generating function $\mathbb{E}[e^{tZ}] = (1-\sigma t)^{-2}$ for all $t < 1/\sigma$.*

We observe some additional useful properties of RBFs as a positional LSH scheme over sequence indices $\{1, \ldots, n\}$. First, the hash bins it forms are contiguous – the indices in each bin are consecutive in the context. Therefore, the diagonal all-1 blocks are contiguous in $M^{(h)}$ as defined in the previous section. Second, the hash bins have nearly uniform sizes: each bin contains exactly $\lceil b \rceil$ consecutive indices, except for the first and last bins, which may contain fewer indices due to being on the boundary of the context.

### 3.3 Approximation in Matrix Norms

We turn to proving eqs. (1) to (3) in Theorem 2.1. As mentioned earlier, eq. (3) follows from standard scalar concentration for Bernoullis (see Lemma A.1). For the spectral norm guarantees, eqs. (1) and (2), we will use the following subexponential matrix Bernstein inequality with unbounded moments.

**Theorem 3.4** ([40], Theorem 6.2 and Remark 6.5)**.** *Let $X_1, \ldots, X_s$ be i.i.d. samples from a distribution $X$ over symmetric $n \times n$ matrices with $\mathbb{E}[X] = L$. Suppose there are $r, \alpha > 0$ such that*

$$\forall\, k \in \mathbb{N}, \quad \|\mathbb{E}[(X-L)^k]\| \le \frac{k!}{2} \cdot r^{k-2} \cdot \alpha^2. \tag{6}$$

*Then, the empirical mean $\widetilde{X} = \frac{1}{s}\sum_{i=1}^{s} X_i$ satisfies $\mathbb{E}[\|L - \widetilde{X}\|] \le \frac{O(1)}{s} \cdot \max\{\alpha\sqrt{s \log n}, r \log n\}$, and $\Pr[\|L - \widetilde{X}\| \ge \varepsilon] \le n \cdot \exp\left(-\frac{s \cdot \varepsilon^2}{2\alpha^2 + 2r\varepsilon}\right)$ for all $\varepsilon > 0$.*

We thus need to establish eq. (6) for $M \sim \mathcal{M}$ with appropriate $r, \alpha$. Fixing $k \in \mathbb{N}$, we have by the sub-additivity and sub-multiplicativity of the spectral norm, and a binomial expansion,

$$\|\mathbb{E}[(M-L^\star)^k]\| \le \mathbb{E}\left[\left(\|M\| + \|L^\star\|\right)^k\right] = \sum_{j=0}^{k} \binom{k}{j} \mathbb{E}[(\|M\|-1)^j]\left(\|L^\star\|+1\right)^{k-j}. \tag{7}$$

Thus we need to bound the moments $\mathbb{E}[(\|M\|-1)^j]$ and the powers $\left(\|L^\star\|+1\right)^j$.

**Lemma 3.5.** $\mathbb{E}[(\|M\|-1)^k] \le \sigma^k(k+1)!$ *for all $k \in \mathbb{N}$.*

*Proof.* $M$ is a block diagonal matrix with all-1 blocks. Let $b_M$ denote the size of its largest block. It is straightforward to see that the largest eigenvalue of $M$ is $b_M$, thus $\|M\| = b_M$. Next, recall that $b_M \le \lceil b \rceil \le b+1$ where $b \sim \Gamma(2, \sigma)$. Hence, $(\|M\|-1)^k \le b^k$. Taking expectation of both sides, the lemma follows from the formula for Gamma moments in Fact 3.3. □

**Lemma 3.6.** $\|L^\star\| \le (e^{1/\sigma}+1)/(e^{1/\sigma}-1)$.

*Proof.* We show two ways to prove the lemma. The first is a more generic approach based on Fourier analysis of Toeplitz matrices, which has the potential advantage of possibly generalizing better to other positional bias matrices beyond ALiBi. Recall that a matrix $T$ is Toeplitz if there is a function $g : \mathbb{Z} \to \mathbb{R}$ such that $T_{ij} = g(i-j)$ for all $i, j$. The ALiBi matrix $L^\star$ is Toeplitz with $g(z) = e^{-|z|/\sigma}$. The following lemma offers a way to bound the spectral norm of a Toeplitz matrix.

**Lemma 3.7** ([19], Lemma 4.1)**.** *Let $T$ be a Toeplitz matrix with $T_{ij} = g(i-j)$. Suppose $g$ satisfies $\sum_{i=-\infty}^{\infty} |g(i)| < \infty$. Then $\|T\| \le \sup f$ where $f$ is the discrete-time Fourier transform (DTFT) of $g$.*

Thus, to bound the spectral norm of $L^\star$ with Lemma 3.7 we need the DTFT of $g(z) = e^{-|z|/\sigma}$. A standard fact, which can be verified by direct derivation, is that the DTFT of $g_\phi(z) = \phi^{|z|}$ for $\phi \in (-1, 1)$ is $f(\omega) = (1-\phi^2)/(1-2\phi\cos(\omega)+\phi^2)$ and its supremum is $(1+\phi)/(1-\phi)$ (attained at $\cos(\omega) = 1$). For $L^\star$ we have $\phi = e^{-1/\sigma}$, and thus Lemma 3.6 follows from Lemma 3.7.

The second proof for Lemma 3.6 is more direct, albeit more specialized to ALiBi. Recall the matrix norm definitions $\|M\|_1 = \max_j \sum_i |M_{ij}|$ and $\|M\|_\infty = \max_i \sum_j |M_{ij}|$. By a standard Hölder

inequality for matrix norms, $\|L^\star\| \le \sqrt{\|L^\star\|_1 \|L^\star\|_\infty}$. Since $L^\star$ is symmetric, $\|L^\star\|_1 = \|L^\star\|_\infty$, hence $\|L^\star\| \le \|L^\star\|_\infty$. Thus, it suffices to bound the row sums in $L^\star$. The sum of row $i$ is

$$\sum_{j=1}^{n} L^\star_{ij} = 1 + \sum_{t=1}^{i-1} e^{-t/\sigma} + \sum_{t=1}^{n-i} e^{-t/\sigma} \le 1 + 2\sum_{t=1}^{\infty} e^{-t/\sigma} = \frac{e^{1/\sigma}+1}{e^{1/\sigma}-1},$$

where the final equality is established by calculating the sum of the geometric series. □

We plug Lemmas 3.5 and 3.6 into eq. (7) and get

$$\|\mathbb{E}[(M-L^\star)^k]\| \le \sum_{j=0}^{k} \binom{k}{j} \cdot \sigma^j (j+1)! \cdot \left(\frac{e^{1/\sigma}+1}{e^{1/\sigma}-1} + 1\right)^{k-j} \le 2^k k! \sum_{j=0}^{k} \binom{k}{j} \cdot \sigma^j \cdot \left(\frac{e^{1/\sigma}+1}{e^{1/\sigma}-1} + 1\right)^{k-j}$$

$$= 2^k k! \cdot \left(1 + \sigma + \frac{e^{1/\sigma}+1}{e^{1/\sigma}-1}\right)^k = (2\Psi_\sigma)^k k! = \frac{k!}{2} \cdot (2\Psi_\sigma)^{k-2} \cdot (2\sqrt{2}\Psi_\sigma)^2,$$

where we have used that $(j+1)! \le (k+1)! \le 2^k k!$ for all $k \ge j \ge 0$. Thus, eq. (6) holds with $r = 2\Psi_\sigma$ and $\alpha = 2\sqrt{2}\Psi_\sigma$, and thus eqs. (1) and (2) in Theorem 2.1 now follow from Theorem 3.4.

### 3.4 Maximal Block Size

We proceed to proving eq. (4) in Theorem 2.1. By the previous sections, this amounts to bounding the maximal bin size in RBFs. To this end, we recall subgamma concentration (see, e.g., [42]).

**Definition 3.8.** *A real-valued random variable $Z$ is called* $(\nu, \kappa)$-right-subgamma *if for all $t \in (0, 1/\kappa)$ it holds that* $\mathbb{E}\left[e^{t(Z-\mathbb{E}[Z])}\right] \le \exp\left(\nu t^2/(2-2\kappa t)\right)$.

**Fact 3.9.** *if $Z$ is $(\nu,\kappa)$-right-subgamma, then for all $t > 0$,* $\Pr[Z \ge \mathbb{E}[Z] + \sqrt{2\nu t} + \kappa t] \le e^{-t}$.

**Lemma 3.10** (see Lemma A.2 in the appendix)**.** *$Z \sim \Gamma(2,\sigma)$ is $(2\sigma^2, \sigma)$-right-subgamma.*

In Theorem 2.1 we draw i.i.d. samples $M_1, \ldots, M_s \sim \mathcal{M}$ and denote by $b_{\max}$ their maximal block size. Let $b_1, \ldots, b_s \sim \Gamma(2,\sigma)$ be the corresponding samples of their RBFs. By the previous sections, each $M_j$ is block-diagonal with maximal bin size $\lceil b_j \rceil \le b_j + 1$. Therefore, $b_{\max} \le 1 + \max_j b_j$. We use Fact 3.9 with $t = \ln(s/\delta)$, $\mathbb{E}[b] = 2\sigma$ from Fact 3.3, and $\nu, \kappa$ from Lemma 3.10. We get

$$\forall j, \quad \Pr[b_j > 1 + 2\sigma + 2\sigma\sqrt{\ln(s/\delta)} + \sigma \ln(s/\delta)] < \delta/s.$$

Since $s \ge 1$ and $\delta < 1/e$, eq. (4) follows by a union bound over $j = 1, \ldots, s$.

### 3.5 Near-Linear time Approximation Algorithm

In this section, we turn to Theorem 2.2. The algorithm in the theorem replaces the exact ALiBi matrix $L^\star$ with the approximation $\widetilde{M}$. Given an attention instance $(K, Q, V)$, in the notation from Section 2, we have $A^\star = A \odot L^\star$ (non-causal case) or $A^\star = A \odot J \odot L^\star$ (causal case), $P^\star = (D^{[A^\star]})^{-1} A^\star$, and the exact output is $T^\star = P^\star V$. The algorithm in Theorem 2.2 computes $\widetilde{A}^\star = A \odot \widetilde{M}$ (non-causal case) or $\widetilde{A}^\star = A \odot J \odot \widetilde{M}$ (causal case), then $\widetilde{P}^\star = (D^{[\widetilde{A}^\star]})^{-1} \widetilde{A}^\star$, and returns $\widetilde{T}^\star = \widetilde{P}^\star V$.

**Running time analysis.** Let $M$ be a sample from $\mathcal{M}$ and let $I_1, \ldots, I_t \subset \{1, \ldots, n\}$ be its contiguous blocks of indices. Let $\widetilde{A} = A \odot M$. Since $M$ and thus $\widetilde{A}$ is block-diagonal, we can compute $\widetilde{A}V$ block by block as $\widetilde{A}V = \sum_{j=1}^{t} \widetilde{A}_{[I_j,I_j]} V_{[I_j,:]}$. Since the blocks in $M$ are all-1 matrices, $\widetilde{A}_{[I_j,I_j]} = (A \odot M)_{[I_j,I_j]} = A_{[I_j,I_j]}$, and hence $\widetilde{A}V = \sum_{j=1}^{t} A_{[I_j,I_j]} V_{[I_j,:]}$. Similarly, $D^{[\widetilde{A}]}$ can be computed as $D^{[\widetilde{A}]} = \sum_{j=1}^{t} D^{[A_{[I_j,I_j]}]}$. The upshot is that $\widetilde{A}$ and $D^{[\widetilde{A}]}$ can be computed by computing only the restriction of $A$ to the blocks of $M$. Note that $A$ is the regular (positionally unbiased) attention matrix, and therefore, computing $\widetilde{A}$ and $D^{[\widetilde{A}]}$ is reduced to a collection of smaller regular attention computations on the subsets of keys and queries that correspond to the blocks $\{I_j\}$. The time complexity is $d\sum_{j=1}^{t} |I_j|^2$. Observe that $\sum_j |I_j| = n$ since the blocks form a partition of $\{1, \ldots, n\}$, and therefore $d\sum_{j=1}^{t} |I_j|^2 \le d \max_j |I_j| \sum_j |I_j| = nd \max_j |I_j|$. With the empirical mean $\widetilde{M} =$

$\frac{1}{s}\sum_{i=1}^{s} M_i$ of $s$ samples with block sizes $\{I_{ij}\}$, the overall time complexity for computing $\widetilde{T}^*$ is thus $\sum_{i=1}^{s} nd \max_j |I_{ij}| \le nd s b_{\max}$. The probabilistic bound on $b_{\max}$ in Theorem 2.1 (eq. (4)) now establishes the running time in Theorem 2.2 in the non-causal case. The same arguments hold in the causal case with $A \odot J$ instead of $A$.

**Approximation guarantee.** Here we only outline how matrix norm approximation yields attention approximation, and leave the full details to the appendix. For all $i, j$ one can verify that $T^\star_{ij} = (\sum_k P_{ik} L^\star_{ik} V_{kj})/(\sum_k P_{ik} L^\star_{ik})$ and $\widetilde{T}^\star_{ij} = (\sum_k P_{ik} \widetilde{M}_{ik} V_{kj})/(\sum_k P_{ik} \widetilde{M}_{ik})$, where recall that $P$ is the row-stochastic matrix $P = (D^{[A]})^{-1} A$. The error in the numerators can be bounded either with the max-norm of $L^\star - \widetilde{M}$ as

$$\left|\sum_k P_{ik}(L^\star - \widetilde{M})_{ik} V_{kj}\right| \le \|L^\star - \widetilde{M}\|_{\max} \|V\|_{\max} \left|\sum_k P_{ik}\right| = \|L^\star - \widetilde{M}\|_{\max} \|V\|_{\max}, \tag{8}$$

or with the spectral norm of $L^\star - \widetilde{M}$ as

$$\left|\sum_k P_{ik}(L^\star - \widetilde{M})_{ik} V_{kj}\right| = \left|(L^\star - \widetilde{M})^\top_{i,*}(P_{i,*} \odot V_{*,j})\right| \le \|L^\star - \widetilde{M}\| \cdot \|V\|_{\max} \|P\|_{2,\infty}. \tag{9}$$

A similar argument applies to the denominators with an all-1 matrix instead of $V$. From Theorem 2.1 we now get that the numerator error is bounded by $\Delta_P \|V\|_{\max}$ and the denominator error is bounded by $\Delta_P$. The bound on $\|T^\star - \widetilde{T}^\star\|_{\max}$ now follows by manipulation. See Appendix A for full details.

## 4 Bounding the Number of HBM Accesses

Theorem 2.2 bounds the time complexity of our algorithm. On modern GPUs, however, the running time of attention is often dominated by the cost of moving data between memory levels, rather than by its time complexity. In this section we analyze our algorithm in a model that captures this cost, following the IO-aware analysis of FlashAttention [15].

A GPU has two main levels of memory. Its main memory, the HBM, is large (tens of gigabytes) and stores the model parameters and the inputs $Q, K, V$. Computation, however, takes place in the on-chip SRAM, which is much smaller but roughly an order of magnitude faster [15]. Every computation therefore requires moving its inputs from HBM to SRAM and its outputs back to HBM. Since the compute throughput of GPUs has grown faster than their memory bandwidth, standard attention is largely memory-bound: its running time is governed mainly by the number of HBM accesses. For this reason, FlashAttention [15] measures the cost of attention by the number of HBM accesses, and reduces it by tiling the computation so that the full $n \times n$ attention matrix is never written to HBM.

Let $M$ denote the SRAM size, which satisfies $d \le M \le nd$. FlashAttention requires $\Theta(n^2 d^2 M^{-1})$ HBM accesses [15]. In our algorithm, each sample $M_r \sim \mathcal{M}$ is a contiguous block-diagonal binary matrix, so the attention computation for $M_r$ reduces to independent local attention operations over its blocks, each of which can be carried out with FlashAttention. This yields the following bound.

**Proposition 4.1.** *Assume $d \le M$. Then the expected number of HBM accesses incurred by the Positional LSH attention algorithm of Theorem 2.2 is $O(snd(1 + d\sigma/M))$.*

Note that for $s \approx \varepsilon^{-2}\sigma^2 \log n$, as implied by Theorem 2.1 for an $\varepsilon$-accurate approximation, the expected number of HBM accesses is $O\big(\varepsilon^{-2}\sigma^2 \log n \cdot nd(1 + d\sigma/M)\big)$, which is near-linear in $n$.

The proof of Proposition 4.1 is given in Appendix A.3.

## 5 Generality: Positional LSH beyond ALiBi

So far we have focused on ALiBi as a concrete case of a popular positional bias that adheres to our positional LSH framework. In this section, we draw a distinction between the results that extend to any positional bias scheme that adheres to Section 3.1 and the results specific to ALiBi.

*What extends to any positional LSH:* Let $L$ be any positional bias matrix that satisfies eq. (5) in Section 3.1. Then the identity $\mathbb{E}_{M\sim\mathcal{M}}[M] = L$ in Theorem 2.1 immediately holds. Furthermore, the

max-norm convergence bound, eq. (3) in Theorem 2.1, also holds, since its proof relies only on scalar Bernoulli concentration (Lemma A.1). Consequently, the uniform attention approximation guarantee of Theorem 2.2 holds in a slightly weaker form, with $\Delta'_P = C\sqrt{\log(n/\delta)/s}$ instead of $\Delta_P$.

*What is specific to ALiBi:* In Theorem 2.1, the spectral norm convergence bounds eqs. (1) and (2) in Theorem 2.1 and the block size bound eq. (4) rely on the specific properties of the RBFs LSH scheme for ALiBi (Definition 3.1), particularly on properties of its underlying Gamma distribution and its interplay with the space of indices in a sequence (Sections 3.3 and 3.4). This also includes the contiguity of the blocks (see Sections 3.1 and 3.2). For other positional bias and LSH schemes, the analogous results would require their own proofs. Note that the bound on the block size is key to the running time of the positionally-biased attention approximation algorithm in Theorem 2.2.

As a concrete positive example, consider the logarithmic KERPLE positional bias [11]. Its additive logit bias for positions $i, j$ is $\text{bias}(i,j) = -r_1 \log(1 + r_2|i-j|)$ with parameters $r_1, r_2 > 0$. In our notation, the corresponding multiplicative bias matrix is $L_{ij}^{\text{KERPLE}} = e^{\text{bias}(i,j)} = (1 + r_2|i-j|)^{-r_1}$. By the standard formula for the Laplace transform of the Gamma distribution, one can check that $L_{ij}^{\text{KERPLE}} = \mathbb{E}_{\lambda\sim\Gamma(r_1,r_2)}[e^{-\lambda|i-j|}]$. Thus, $L_{ij}^{\text{KERPLE}}$ is a Gamma-weighted mixture of Laplace kernels. This gives rise to a positional LSH scheme for KERPLE: sample $\lambda \sim \Gamma(r_1, r_2)$ and then run RBF with bandwidth $1/\lambda$ (i.e., conditioned on $\lambda$, the RBF bin width is sampled from $\Gamma(2, 1/\lambda)$).

On the other hand, one can establish negative examples, where no (exact) positional LSH scheme exists. For example, it is known that the Gaussian kernel $\ker(u,u') = e^{-\|u-u'\|_2^2}$ over $\mathbb{R}^d$ has no LSH scheme even for $d = 1$ [18]. In particular, no LSH scheme exists for $e^{-\frac{1}{4}|i-j|^2}$ over integers $i, j$. To see this, in [9] it is proven that if $\ker(u,u')$ admits an LSH scheme then $1 - \ker(u,u')$ must satisfy the triangle inequality over $U$; however, one can check the triangle inequality $1 - e^{-\frac{1}{4}|i-k|^2} \leq 1 - e^{-\frac{1}{4}|i-j|^2} + 1 - e^{-\frac{1}{4}|j-k|^2}$ fails for any triplet of consecutive integers $i, j, k$. Therefore, an additive quadratic bias term $-\frac{1}{4}|i-j|^2$ cannot have an exact positional LSH scheme.

## 6 Experimental Validation

### 6.1 Convergence in Matrix Norms

First, we validate our theoretical results on convergence in matrix norms in Theorems 2.1 and 2.2 directly on attention operations in pre-trained models. We use two publicly available large language models: Qwen3-14B and Mistral-7B. We use as inputs 30 text segments from the Wikitext-103 dataset [31] with context lengths of 4k-4.5k and record the output of the models. We then retrieve the corresponding $Q$,$K$,$V$ matrices for each model per input, layer and attention head. We compute $\|L^\star - \widetilde{M}\|$, $\|L^\star - \widetilde{M}\|_{\max}$ and $\|T^\star - \widetilde{T}^\star\|_{\max}$, for varying sample sizes.

The results are reported in Figure 2. Results for additional $\sigma$ values is displayed in Appendix B. The results show that for both models, norm errors decay towards zero as the sample size increases.

### 6.2 Model Training Experiments

Next, we evaluate the effects of our approximation method on downstream model performance. To this end, we implemented a prototype of our positional LSH method for ALiBi in FlashAttention-2 [14]. Our code is available online[1]. We note that our implementation is not optimized for performance from an engineering perspective (see also the discussion of our limitations in Section 8). The goal of the experiments in this section is to gain empirical insight into the approximation quality of our method rather than demonstrate efficiency gains.

We train publicly available models on publicaly available data:

- Qwen3-0.6B [39], pre-trained from scratch on Wikitext-103 [31] (103M raw words).
- Qwen3-4B [39], pre-trained from scratch on PG-19 dataset [33] (1.97B raw words).
- Mistral-7B [25], fine-tuned from its publicly available pre-trained checkpoint, on Wikitext-103.

[1] https://github.com/DanielWolfsonTAU/positional_lsh

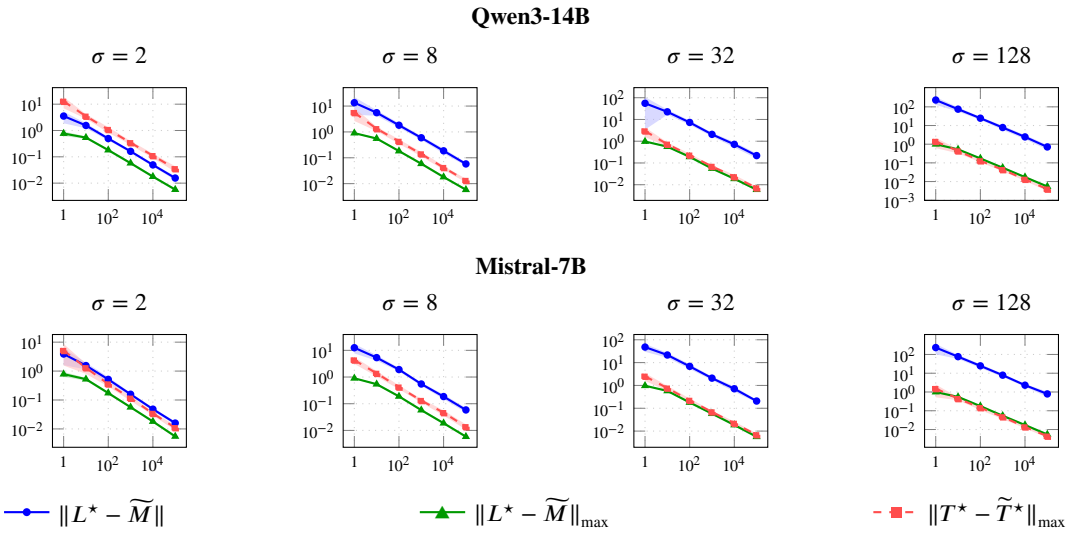


Figure 2: $\|L^\star - \widetilde{M}\|$, $\|L^\star - \widetilde{M}\|_{\max}$ and $\|T^\star - \widetilde{T}^\star\|_{\max}$ as functions of the sample size $s$, for four attention heads (with $\sigma$ values: $\sigma \in \{2, 8, 32, 128\}$), for the models Qwen3-14B and Mistral-7B. Both axes show values on a log scale. Shaded regions are mean±std across inputs.

We evaluate perplexity (PPL) and next-token prediction accuracy (Acc), both on the original context length used in training and under context length extrapolation (where the model is evaluated on longer context lengths than it was trained on). Qwen3-0.6B and Qwen3-4B are trained with an 8k context length, and Mistral-7B is trained with a 4k context length. We evaluate them at their trained context lengths as well as extrapolated context lengths up to 16k or 64k.

We evaluate Positional LSH with different sample sizes $s$. We compare it with the original models without positional bias and with ALiBi. As a baseline for comparison, we also include a "fixed blocks" baseline with a single block-diagonal mask matrix with a fixed block size of 512.

The results are displayed in Tables 1 to 3. For all models, as the number of samples in Positional LSH increases, performance improves, approaching that of ALiBi. Furthermore, Positional LSH outperforms the fixed blocks baseline. Finally, for Qwen3-0.6B, Positional LSH outperforms the original model, even with a small sample size. For Qwen3-4B, this occurs only with $s = 10$ samples and at the highest tested extrapolated context length of 64k. On Mistral-7B, neither positional bias method outperforms the original model in length extrapolation.

Additional experimental details and parameters are provided in Appendix B.

## 7 Related Work

Positional encodings and relative-position-aware attention in transformers are widely studied. Apart from RoPE, a central line of work injects positional information directly into attention through relative-position terms. Shaw et al. [36] introduced relative position representations inside self-attention, and T5 [34] used a simplified scalar relative bias added to attention logits. Fixed and kernelized bias schemes for length extrapolation include ALiBi and KERPLE [11], while FIRE [29] provides a unified functional view. Jelassi et al. [24] also studied hard-ALiBi, a binary sliding-window variant with head-dependent window sizes, while Kazemnejad et al. [27] studied NoPE, i.e., using no positional encoding at all. Our work is complementary to these methods: rather than proposing a new positional bias family, we propose a structural and algorithmic analysis of ALiBi.

From an efficiency perspective, exact attention with bias has recently been optimized at the systems level, including ALiBi-FlashAttention [12] for exact computation and FlashBias [43] via a low-rank decomposition. Finally, LSH has a long history in similarity search [23, 17, 9, 3] and kernel methods [8, 37, 4, 7, 13, 18], and has more recently been used for efficient attention and KV-cache compression [28, 20, 10, 22, 16, 26]. Prior LSH-based attention work hashes token representations or semantic content; in contrast, we use LSH in a different way by hashing positions.

Table 1: Length extrapolation results on Wikitext-103 for Qwen3-0.6B (trained from scratch)

| Context length: | 8k (training) | | 16k (extrapolation) | |
|---|---|---|---|---|
| Method | PPL | Acc | PPL | Acc |
| Original model | 20.213±0.088 | 44.796±0.182 | 21.316±0.058 | 43.652±0.126 |
| ALiBi | 18.920±0.009 | 45.249±0.035 | 18.753±0.021 | 45.351±0.020 |
| Positional LSH, $s = 1$ | 19.217±0.059 | 45.051±0.044 | 19.278±0.027 | 44.985±0.029 |
| Positional LSH, $s = 10$ | 19.073±0.054 | 45.125±0.025 | 19.043±0.070 | 45.179±0.021 |
| Positional LSH, $s = 20$ | 19.050±0.033 | 45.146±0.032 | 18.961±0.033 | 45.192±0.046 |
| Positional LSH, $s = 50$ | 19.028±0.048 | 45.161±0.018 | 18.920±0.020 | 45.206±0.041 |
| Positional LSH, $s = 100$ | 19.017±0.018 | 45.172±0.034 | 18.893±0.035 | 45.219±0.023 |
| Fixed blocks | 19.788±0.054 | 44.725±0.058 | 19.784±0.041 | 44.658±0.067 |

Table 2: Length extrapolation results on PG-19 for Qwen3-4B (trained from scratch)

| Context length: | 8k (training) | | 16k (ext.) | | 32k (ext.) | | 64k (ext.) | |
|---|---|---|---|---|---|---|---|---|
| Method | PPL | Acc | PPL | Acc | PPL | Acc | PPL | Acc |
| Original model | 11.775 | 47.481 | 11.446 | 47.838 | 11.776 | 47.353 | 12.633 | 46.183 |
| ALiBi | 11.751 | 47.568 | 11.445 | 47.930 | 11.450 | 47.919 | 11.402 | 47.973 |
| Positional LSH, $s = 1$ | 12.408 | 46.777 | 13.230 | 45.936 | 13.158 | 45.927 | 13.029 | 46.086 |
| Positional LSH, $s = 10$ | 12.298 | 46.915 | 12.315 | 46.919 | 12.318 | 46.885 | 12.218 | 46.965 |
| Fixed blocks | 12.841 | 45.473 | 13.511 | 45.349 | 13.515 | 45.340 | 13.462 | 45.593 |

Table 3: Length extrapolation results on Wikitext-103 for Mistral-7B (fine-tuned)

| Context length: | 4k (training) | | 8k (extrapolation) | | 16k (extrapolation) | |
|---|---|---|---|---|---|---|
| Method | PPL | Acc | PPL | Acc | PPL | Acc |
| Original model | 6.347±0.001 | 58.371±0.016 | 6.273±0.002 | 58.523±0.016 | 6.270±0.004 | 58.501±0.039 |
| ALiBi | 6.498±0.008 | 57.924±0.142 | 6.455±0.022 | 58.086±0.134 | 6.461±0.021 | 57.960±0.207 |
| Positional LSH, $s = 1$ | 6.957±0.009 | 56.807±0.031 | 6.941±0.015 | 56.926±0.025 | 6.923±0.011 | 56.975±0.027 |
| Positional LSH, $s = 5$ | 6.725±0.013 | 57.186±0.051 | 6.711±0.019 | 57.325±0.038 | 6.721±0.012 | 57.235±0.058 |
| Positional LSH, $s = 10$ | 6.698±0.010 | 57.280±0.026 | 6.683±0.011 | 57.389±0.028 | 6.705±0.034 | 57.284±0.021 |
| Positional LSH, $s = 15$ | 6.672±0.011 | 57.324±0.023 | 6.664±0.017 | 57.412±0.023 | 6.668±0.016 | 57.338±0.032 |
| Fixed blocks | 7.352±0.014 | 54.919±0.022 | 7.327±0.017 | 54.944±0.019 | 7.378±0.015 | 54.930±0.013 |

# 8 Conclusions and Limitations

We introduced a positional LSH framework for attention with bias, establishing a formal connection between positional bias, binary attention masks, and positional embeddings. For ALiBi, this yields both a structural approximation of the bias matrix by randomized contiguous block masks and a uniform near-linear time approximation theorem for ALiBi-biased attention. Our experiments support this approximation perspective: as the number of samples increases, the approximation improves and downstream behavior approaches exact ALiBi. Our framework is more general than ALiBi and may extend other fixed positional bias kernels.

Our contribution is theoretical and algorithmic rather than a systems result, and our experiments are of limited scale meant only to validate the theory. Although our analysis gives a near-linear asymptotic algorithm, we do not demonstrate wall-clock speedups over hardware-optimized exact ALiBi implementations. Some of this gap may be due to our implementation being under-optimized, but it likely also reflects the fact that current hardware and low-level kernels are highly optimized for large dense operations, so asymptotic gains from decomposing attention into many smaller local computations need not translate into speedups at the context lengths we experimented with. Identifying regimes where the favorable asymptotics of positional LSH lead to practical gains may involve much longer context lengths and is left an intriguing direction for future work.

## AI Use Statement

GPT 5.2 and 5.4 were used for writing assistance and polishing, literature search, and improving the clarity of mathematical presentation. GPT 5.4 also proposed the second (Hölder-based) proof of Lemma 3.7 after reviewing the authors' original (Fourier-based) proof. GPT and Claude were also used for interactive coding. The authors are accountable for the accuracy of all content and artifacts associated with this paper.

## Acknowledgments and Disclosure of Funding

This research was supported by the Israeli Ministry of Innovation, Science & Technology, by Len Blavatnik and the Blavatnik Family foundation, and by an Alon Scholarship. This research was made possible through a GPU compute resource grant funded by the Association of University Heads, the Council for Higher Education and the AI Research Compute Center.

## References

[1] Alex Andoni and Piotr Indyk. Dimension reduction in kernel spaces from locality-sensitive hashing. 2009.

[2] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117–122, 2008.

[3] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. *Advances in neural information processing systems*, 28, 2015.

[4] Arturs Backurs, Piotr Indyk, and Tal Wagner. Space and time efficient kernel density estimation in high dimensions. *Advances in neural information processing systems*, 32, 2019.

[5] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[6] Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE, 1997.

[7] Moses Charikar, Michael Kapralov, Navid Nouri, and Paris Siminelakis. Kernel density estimation through density constrained near neighbor search. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 172–183. IEEE, 2020.

[8] Moses Charikar and Paris Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1032–1043. IEEE, 2017.

[9] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388, 2002.

[10] Zhuoming Chen, Ranajoy Sadhukhan, Zihao Ye, Yang Zhou, Jianyu Zhang, Niklas Nolte, Yuandong Tian, Matthijs Douze, Leon Bottou, Zhihao Jia, et al. Magicpig: Lsh sampling for efficient llm generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399, 2022.

[12] Sanghun Cho, Ofir Press, and Tri Dao. ALiBi-FlashAttention: Speeding up ALiBi by 3–5x with a hardware-efficient implementation. Princeton Language and Intelligence Blog, Jan 2024. Accessed: 2026-04-27.

[13] Benjamin Coleman and Anshumali Shrivastava. Sub-linear race sketches for approximate kernel density estimation on streaming data. In *Proceedings of The Web Conference 2020*, pages 1739–1749, 2020.

[14] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

[16] Aditya Desai, Shuo Yang, Alejandro Cuadron, Matei Zaharia, Joseph E Gonzalez, and Ion Stoica. Hashattention: Semantic sparsity for faster inference. In *Forty-second International Conference on Machine Learning*, 2025.

[17] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, volume 99, pages 518–529, 1999.

[18] Barak Gorodissky and Tal Wagner. Locality sensitive hashing for p-exponential kernels with applications to density estimation. In *Neural information processing systems (NeurIPS)*, 2026.

[19] Robert M Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006.

[20] Insu Han, R Jayaram, A Karbasi, V Mirrokno, D Woodruff, and A Zandieh. Hyperattention: Long-context attention in near-linear time. In *International Conference on Learning Representations*. International Conference on Learning Representations, 2024.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[22] Piotr Indyk, Michael Kapralov, Kshiteej Sheth, and Tal Wagner. Improved algorithms for kernel matrix-vector multiplication under sparsity assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025.

[23] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[24] Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.

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

[26] Sahil Joshi, Agniva Chowdhury, Amar Kanakamedala, Ekam Singh, Evan Tu, and Anshumali Shrivastava. Race attention: A strictly linear-time attention for long-sequence training. In *The Fourteenth International Conference on Learning Representations*, 2026.

[27] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

[28] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

[29] Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[31] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[32] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

[33] Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[35] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[36] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018.

[37] Paris Siminelakis, Kexin Rong, Peter Bailis, Moses Charikar, and Philip Levis. Rehashing kernel evaluation in high dimensions. In *International Conference on Machine Learning*, pages 5789–5798. PMLR, 2019.

[38] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[39] Qwen Team. Qwen3 technical report, 2025.

[40] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Bo Waggoner. Sub-gamma variables and concentration, 2019. Available at https://www.bowaggoner.com/blog/2019/02-03-sub-gamma-concentration/.

[43] Haixu Wu, Minghao Guo, Yuezhou Ma, Yuanxu Sun, Jianmin Wang, Wojciech Matusik, and Mingsheng Long. Flashbias: Fast computation of attention with bias. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[44] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. PyTorch FSDP: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860, 2023.

# A Deferred Proofs

## A.1 Omitted Lemmas for Theorem 2.1

**Lemma A.1** (max-norm bound in Theorem 2.1)**.** *Under the assumptions of Theorem 2.1, the max-norm bound in eq. (3) holds.*

*Proof.* Let $i, j \in [n]$. By Lemma 3.2 and the positional LSH construction in Section 3.1, we have $L^\star_{ij} = \Pr_{M\sim\mathcal{M}}[M_{ij} = 1]$. Therefore, $\widetilde{M}_{ij}$ is the average of $s$ Bernoulli random variables with expectation $L^\star_{ij}$. Therefore, by the Chernoff-Hoeffding inequality,

$$\Pr\left[\left|L^\star_{ij} - \widetilde{M}_{ij}\right| > \varepsilon\right] < 2e^{-2s\varepsilon^2}$$

The lemma follows by a union bound over all $i, j$. □

**Lemma A.2** (Lemma 3.10 restated)**.** $Z \sim \Gamma(2, \sigma)$ *is* $(2\sigma^2, \sigma)$*-right-subgamma.*

*Proof.* Let $t \in (0, 1/\sigma)$. By plugging the expectation and MGF from Fact 3.3,

$$\ln \mathbb{E}[e^{t(Z-\mathbb{E}[Z])}] = \ln\left(\mathbb{E}[e^{tZ}] \cdot e^{-t\mathbb{E}[Z]}\right) = \ln((1-\sigma t)^{-2} \cdot e^{-2\sigma t}) = 2(-\ln(1-\sigma t) - \sigma t).$$

From the Taylor expansion of $-\ln(1-x)$ at 0 we have, for every $x \in (0, 1)$,

$$-\ln(1-x) - x = \sum_{i=1}^{\infty} \frac{x^i}{i} - x = \sum_{i=2}^{\infty} \frac{x^i}{i} \le \frac{x^2}{2} \sum_{i=0}^{\infty} x^i = \frac{x^2}{2} \cdot \frac{1}{1-x}.$$

Plugging this above with $x = \sigma t$,

$$\ln \mathbb{E}[e^{t(Z-\mathbb{E}[Z])}] \le 2 \cdot \frac{\sigma^2 t^2}{2} \cdot \frac{1}{1-\sigma t} = \frac{2\sigma^2 t^2}{2(1-\sigma t)}.$$

Exponentiating both sides proves the claim. □

## A.2 Proof of Theorem 2.2

The running time was proven in Section 3.5. Here we prove the approximation guarantee. Let $A' = A$ in the non-causal case and $A' = A \odot J$ in the causal case. In either case we have $P = (D^{[A']})^{-1}A'$, and $P$ is row-stochastic.

Observe that the exact target output $T^\star$ and the returned approximation $\widetilde{T}^\star$ are computed as

$$T^\star_{ij} = \frac{\sum_{k=1}^n A'_{ik} L^\star_{ik} V_{kj}}{\sum_{k=1}^n A'_{ik} L^\star_{ik}} \quad \text{and} \quad \widetilde{T}^\star_{ij} = \frac{\sum_{k=1}^n A'_{ik} \widetilde{M}_{ik} V_{kj}}{\sum_{k=1}^n A'_{ik} \widetilde{M}_{ik}}.$$

We may divide both the numerators and the denominators by $D^{[A']}_{ii}$ and write,

$$T^\star_{ij} = \frac{\sum_{k=1}^n P_{ik} L^\star_{ik} V_{kj}}{\sum_{k=1}^n P_{ik} L^\star_{ik}} \quad \text{and} \quad \widetilde{T}^\star_{ij} = \frac{\sum_{k=1}^n P_{ik} \widetilde{M}_{ik} V_{kj}}{\sum_{k=1}^n P_{ik} \widetilde{M}_{ik}}.$$

For convenience let us denote the numerators and the denominators as

$$N_{ij} = \sum_{k=1}^n P_{ik} L^\star_{ik} V_{kj}, \quad \widetilde{N}_{ij} = \sum_{k=1}^n P_{ik} \widetilde{M}_{ik} V_{kj}, \quad D_i = \sum_{k=1}^n P_{ik} L^\star_{ik}, \quad \widetilde{D}_i = \sum_{k=1}^n P_{ik} \widetilde{M}_{ik}.$$

Equations (8) and (9) in Section 3.5 show that

$$\forall i, j \quad \left|N_{ij} - \widetilde{N}_{ij}\right| \le \min\{\|L^\star - \widetilde{M}\|_{\max}, \|P\|_{2,\infty}\|L^\star - \widetilde{M}\|\} \cdot \|V\|_{\max},$$

and the same arguments with an all-1 matrix instead of $V$ yield

$$\forall i \quad \left|D_i - \widetilde{D}_i\right| \le \min\{\|L^\star - \widetilde{M}\|_{\max}, \|P\|_{2,\infty}\|L^\star - \widetilde{M}\|\}.$$

With probability $1-\delta$ we have by Equations (2) and (3) in Theorem 2.1,

$$\|L^\star - \widetilde{M}\|_{\max} \le O\left(\sqrt{\frac{\log(n/\delta)}{s}}\right) \quad \text{and} \quad \|L^\star - \widetilde{M}\| \le O\left(\Psi_\sigma\sqrt{\frac{\log(n/\delta)}{s}} + \frac{\log(n/\delta)}{s}\right).$$

In the statement of Theorem 2.2 we assume that $s \ge \log(n/\delta)$ and hence $\frac{\log(n/\delta)}{s} \le \sqrt{\frac{\log(n/\delta)}{s}}$. Plugging this above, we get

$$\forall i,j \quad \left|N_{ij} - \widetilde{N}_{ij}\right| \le \Delta_P \|V\|_{\max} \quad \text{and} \quad \forall i \quad \left|D_i - \widetilde{D}_i\right| \le \Delta_P. \tag{10}$$

Furthermore, we have

$$\forall i,j \quad \left|\frac{N_{ij}}{D_i}\right| = \left|\frac{\sum_{k=1}^n P_{ik}L^\star_{ik}V_{kj}}{\sum_{k=1}^n P_{ik}L^\star_{ik}}\right| \le \left|\frac{\sum_{k=1}^n P_{ik}L^\star_{ik}}{\sum_{k=1}^n P_{ik}L^\star_{ik}}\right| \cdot \|V\|_{\max} = \|V\|_{\max}. \tag{11}$$

From Equations (10) and (11),

$$\left|\widetilde{T}^\star_{ij} - T^\star_{ij}\right| = \left|\frac{\widetilde{N}_{ij}}{\widetilde{D}_i} - \frac{N_{ij}}{D_i}\right| = \left|\frac{\widetilde{N}_{ij} - N_{ij}}{\widetilde{D}_i} + \frac{N_{ij}}{D_i}\cdot\frac{D_i - \widetilde{D}_i}{\widetilde{D}_i}\right| \le \frac{2\Delta_P\|V\|_{\max}}{\widetilde{D}_i}. \tag{12}$$

Finally, observe that $\beta^\star_P = \min_i D_i$ in the notation of Theorem 2.2. Therefore, if the condition of item 2 in the theorem holds, then by Equation (10) we have for all $i$ that $\widetilde{D}_i \ge D_i - \left|\widetilde{D}_i - D_i\right| \ge \beta^\star_P - \Delta_P$. Plugging this in Equation (12) yields the theorem. □

### A.3 Proof of Proposition 4.1

We use the notation of Section 4. Recall that the algorithm of Theorem 2.2 draws $s$ i.i.d. samples $M_1, \dots, M_s \sim \mathcal{M}$. For each sample $M_r$, if its blocks have sizes $|I_1|, \dots, |I_t|$ (with $\sum_j |I_j| = n$), the attention computation restricted to $M_r$ reduces to $t$ independent local attention operations over sequences of lengths $|I_1|, \dots, |I_t|$. We first explain how the outputs of these local operations are combined, and then bound the number of HBM accesses.

**Combining the local attention outputs.** Since FlashAttention returns normalized attention outputs, whereas the algorithm of Theorem 2.2 averages the sampled masks before normalization, the local outputs must be combined with appropriate weights. Let $A' = A$ in the non-causal case and $A' = A \odot J$ in the causal case. For a row $i$ and a sample $r \in \{1, \dots, s\}$, let

$$z_{ir} = \sum_{k=1}^n A'_{ik}(M_r)_{ik} \quad \text{and} \quad o_{ir} = \frac{1}{z_{ir}}\sum_{k=1}^n A'_{ik}(M_r)_{ik}V_{k,*}.$$

Since the blocks of $M_r$ are all ones and partition $\{1, \dots, n\}$, $z_{ir}$ and $o_{ir}$ are the softmax normalizer and the output of row $i$ in the local attention operation over the block of $M_r$ that contains $i$. Since $\widetilde{M} = \frac{1}{s}\sum_{r=1}^s M_r$, the output of the algorithm is

$$\widetilde{T}^\star_{i,*} = \frac{\sum_{k=1}^n A'_{ik}\widetilde{M}_{ik}V_{k,*}}{\sum_{k=1}^n A'_{ik}\widetilde{M}_{ik}} = \frac{\sum_{r=1}^s z_{ir}\, o_{ir}}{\sum_{r=1}^s z_{ir}}, \tag{13}$$

where the factors $\frac{1}{s}$ cancel. FlashAttention maintains for each row the maximum $m$ of the attention scores and the sum $\ell$ of the exponentiated scores shifted by $m$, and writes both to HBM (see Algorithm 1 in [15]); the normalizer is then $z = e^m\ell$. Hence each local operation provides both $o_{ir}$ and $z_{ir}$, and the outputs can be combined by eq. (13) (in practice, in a numerically stable way using $\log z_{ir} = m + \log \ell$).

*Proof of Proposition 4.1.* Fix a single sample $M_r \sim \mathcal{M}$, with contiguous blocks $I_1, \dots, I_t$ partitioning $\{1, \dots, n\}$, induced by a draw $b \sim \Gamma(2, \sigma)$ from the RBF scheme of Definition 3.1. By that construction, every block has size $|I_j| \le b + 1$.

**Local attention operations.** We apply Algorithm 1 of [15] as a black box to each block independently. On a sequence of length $N$, it uses column blocks of size $B_c = \left\lceil \frac{M}{4d} \right\rceil$, and by the proof of Theorem 2

in [15], it incurs $\Theta(N d T_c)$ HBM accesses, where $T_c = \left\lceil \frac{N}{B_c} \right\rceil$ is the number of column blocks. Since $B_c \geq \frac{M}{4d}$, we have $T_c \leq 1 + \frac{4Nd}{M}$. Hence, for a block of size $|I_j|$, the number of HBM accesses is $O\Big(|I_j|\, d \left(1 + \frac{|I_j|\, d}{M}\right)\Big)$. Summing over all blocks of a single sample:

$$\text{HBM accesses for one sample} \;\leq\; O\left(\sum_{j=1}^{t} |I_j|\, d\right) + O\left(\sum_{j=1}^{t} \frac{|I_j|^2\, d^2}{M}\right).$$

The first sum is $O(nd)$. For the second sum, taking expectation over $b \sim \Gamma(2, \sigma)$ and using $\mathbb{E}[b] = 2\sigma$ from Fact 3.3:

$$\mathbb{E}\left[\sum_{j=1}^{t} \frac{|I_j|^2\, d^2}{M}\right] \;\leq\; \frac{d^2}{M}\, \mathbb{E}\left[\max_j |I_j| \cdot \sum_{j=1}^{t} |I_j|\right] \;\leq\; \frac{d^2}{M} \cdot \mathbb{E}[b+1] \cdot n \;=\; O\left(\frac{nd^2\sigma}{M}\right).$$

**Combining the samples.** The outputs $o_{ir}$ and the values $m$, $\ell$ are written to HBM by FlashAttention itself, so the cost of these writes is already included in the bound for the local attention operations. Computing the rows $\widetilde{T}^{\star}_{i,*}$ for all $i$ by eq. (13) then requires reading them once for all $s$ samples, which takes $O(snd)$ HBM accesses.

Summing the cost of the local operations over all $s$ samples and adding the cost of combining them, the total expected number of HBM accesses is $O\Big(snd\Big(1 + \frac{d\sigma}{M}\Big)\Big)$. □

## B Additional Experiments and Experimental Details

Figure 3 depicts plots for additional $\sigma$ values, complementing the experiments described in subsection 6.1.

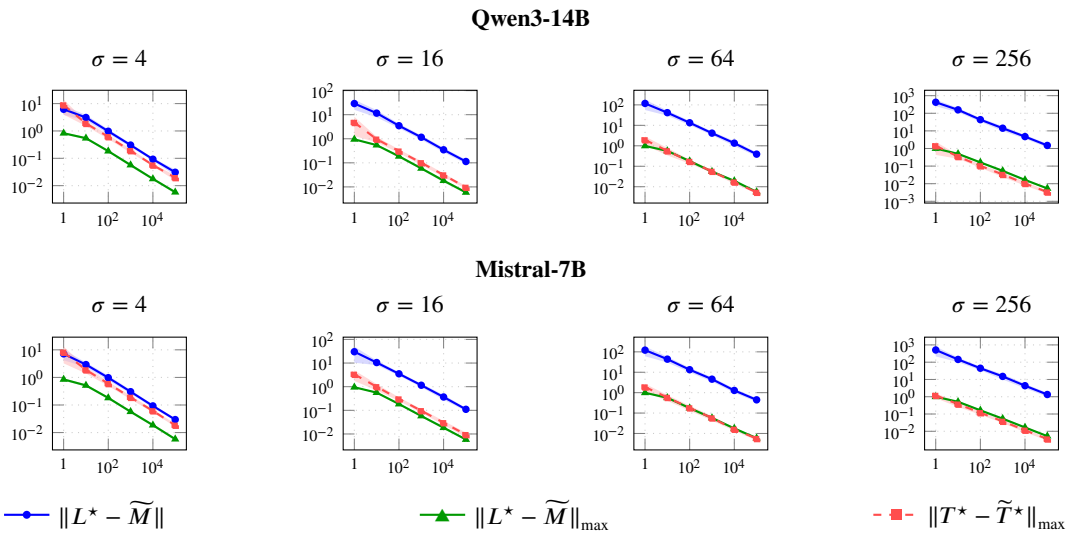


Figure 3: $\|L^{\star} - \widetilde{M}\|$, $\|L^{\star} - \widetilde{M}\|_{\max}$ and $\|T^{\star} - \widetilde{T}^{\star}\|_{\max}$ as functions of the sample size $s$, for four attention heads (with $\sigma$ values: $\sigma \in \{4, 16, 64, 256\}$), for the models Qwen3-14B and Mistral-7B. Both axes show values on a log scale. Shaded regions span [mean − std, mean + std] across inputs, and they do not appear if the standard deviation is negligible.

Similarly to the results in subsection 6.1, for both models, all norms decrease towards zero as the sample size increases.

*Experimental details.* Experiments on all models except pre-training Qwen3-4B were run on an NVIDIA H100 Tensor Core GPU (with 80GB VRAM). Experiments on Qwen3-4B were run on 8 NVIDIA B200 GPUs using Fully Sharded Data Parallel (FSDP) [44]. The AdamW optimizer [30]

was used for training, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The hyperparameters were selected via evaluations performed on three different values for each parameter, spanning conventional value ranges. For Qwen3-0.6B, a learning rate of $5 \times 10^{-4}$ is used, while for Mistral-7B the learning rate is $1 \times 10^{-4}$. The remaining hyperparameters are shared across both models: 5 epochs, a weight decay of 0.01, and a warmup ratio of 0.1. A batch size of 1 is used for training and evaluation in both models, as larger batch sizes would exceed GPU memory limits and would compromise context length. The results of Section 6.1 are averaged over 3 independent runs. Mistral-7B was fine-tuned using LoRA [21] with the following configuration: rank 16, $\alpha = 32$, dropout 0.05, no bias, targeting the query, key, value, and output projection matrices.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The theoretical results are explained in sections 2-4, and the experiment validation is explained in section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are described in section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The proofs contain the full set of their assumptions. They are complete and correct, as required.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses all relevant information to reproduce the main experimental results of the paper, as part of the Experimental Validation section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code to reproduce the paper's experiments, as well as its running instructions, are provided with the submission.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the relevant test details in the Experimental Validation section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports error bars, mean and standard deviation results as part of the experimental section. Error bars are omitted only for the Qwen3-4B pre-training experiment, since its large scale and the compute it required made it impossible for us to perform repetitions of this experiments in time.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The type of the used GPU and its memory capacity are described in the Experimental Validation section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper fully adheres to the CoE.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper describes theoretical and experimental results related to the attention mechanism in transformers - it does not introduce specific societal impacts beyond those inherent to all research in ML and AI.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper and its work uses only open source models and datasets and does not create the need for further custom safeguards. .

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The code, datasets and models used in the paper are all open source. They are credited and their licenses are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research that necessitates IRB or similar approvals.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.