# The Cost of Absolute Position: A Spread–Expressivity Tradeoff for Additive Positional Encodings

**Noah Mitchell*** **Isaac Gabriel*** **Alexander Wyatt***
University of Ottawa
{nmitc082, igabr085, awyat049}@uottawa.ca

## Abstract

Self-attention without positional information is permutation equivariant, therefore positional encodings are required whenever a transformer must distinguish input order. However, injecting absolute position can also weaken length generalization. We formalize this tension for additive positional encodings by introducing positional spread, the unnormalized centered variance of the positional embedding matrix. For additive PE in softmax attention with mean pooling, we prove an upper bound demonstrating that the equivariance breaking component of risk is controlled by $\|\widetilde{E}\|_F^2$, along with a worst-case lower-bound construction for self-separable additive PEs. We finally prove a spread-expressivity tradeoff: any additive PE that distinguishes absolute positions with minimum separation $\delta$ must have spread $\Omega(\delta^2 n)$. The fixed-length analysis suggests two regimes for length extrapolation: when the target is near permutation invariant, lower spread may help; when the task requires absolute position, additional spread reduction destroys expressivity. We test this prediction using controlled synthetic tasks, spread-regularized additive PE, and long context retrieval and reasoning tasks. Across low positional sensitivity tasks, spread strongly predicts OOD accuracy; on absolute position tasks, performance demonstrates the predicted non-monotone tradeoff. For score-based methods such as RoPE and ALiBi, we give a one-sided score-space displacement bound and a qualitative comparison, without extending the additive lower bound.

## 1 Introduction

Transformer architectures [Vaswani et al., 2017] rely on positional encodings to distinguish token order, because self-attention without positional information is permutation equivariant. This symmetry is useful: equivariant or nearly equivariant models can generalize well when the target function respects the same symmetry. But it is also limiting: a permutation-equivariant model cannot represent first-token prediction, absolute-position lookup, or any rule that depends on a distinguished index.

Positional encodings therefore play two roles. They give transformers access to order, but they also break a symmetry that can otherwise support generalization. This raises a basic question:

*What is the geometric cost of giving a transformer access to absolute position?*

We answer this question for additive positional encodings. Let $E \in \mathbb{R}^{n\times d}$ be an additive PE matrix, and let $\widetilde{E}$ denote the row-centered version of $E$. We define the *positional spread* of $E$ as

$$\|\widetilde{E}\|_F^2 = \sum_{i=1}^{n} \|e_i - \bar{e}\|_2^2.$$

*Equal contribution

This quantity has a simple interpretation: it is exactly one half of the expected squared displacement of $E$ under a random permutation. Thus spread measures how strongly the positional encoding resists permutation of the input sequence.

Our main theoretical result shows that this geometric quantity controls the cost of symmetry breaking. For additive PE in softmax attention with mean-pooling readout, we prove an upper bound showing that the equivariance-breaking component of risk is at most proportional to $\|\widetilde{E}\|_F^2$. We also give a worst-case lower-bound construction for self-separable additive PEs, showing that linear spread dependence cannot generally be removed. The upper-bound constant is uniform over PE classes with bounded operator norm.

The same quantity also exposes a fundamental tradeoff. If an additive PE can distinguish all absolute positions with minimum pairwise separation $\delta$, then its spread must satisfy

$$\|\widetilde{E}\|_F^2 \geq \Omega(\delta^2 n).$$

Consequently, absolute-position expressivity cannot be obtained for free: reducing spread improves the symmetry-based generalization bound, but reducing it too far removes the positional signal needed for absolute-position tasks.

This yields a task-dependent prediction. On tasks whose targets are close to permutation invariant, low-spread encodings should extrapolate better to longer sequences. On tasks that genuinely require absolute position, however, the best performance should occur at an intermediate level of spread: enough to encode position, but not so much that the model pays an unnecessary equivariance-breaking cost. We introduce MS-PE, a spread-regularized additive positional encoding, primarily as an experimental knob for tracing this tradeoff.

Although these bounds are specific to additive PE, many widely used positional schemes are not additive. RoPE [Su et al., 2024] and ALiBi [Press et al., 2022], for example, modify attention scores rather than adding vectors to token embeddings. To compare these methods with additive PE, we introduce a displacement framework that places additive and score-based schemes on a common symmetry-breaking axis. A one-sided score-space displacement bound extends the symmetry-breaking viewpoint, but does not give the additive lower-bound or expressivity results for relative-position methods.

Empirically, we validate the spread–expressivity view on controlled synthetic tasks and long-context retrieval/reasoning benchmarks. In low-positional- sensitivity regimes, spread strongly predicts out-of-distribution length accuracy. When absolute position is required, the relationship becomes non-monotone, as predicted by the tradeoff. Supplementary WikiText-103 results provide a larger-scale check of the qualitative displacement ordering.

**Contributions.**

1. **A geometric cost for additive positional encodings.** We introduce positional spread, $\|\widetilde{E}\|_F^2$, and show that it is exactly tied to the expected permutation displacement of an additive PE matrix.
2. **Spread bounds for equivariance breaking.** For additive PE in softmax attention with mean-pooling readout, we prove an upper bound showing that the equivariance-breaking component of risk is controlled by $\|\widetilde{E}\|_F^2$, and a worst-case lower-bound construction for self-separable additive PEs (Theorems 7 and 9). The upper bound extends to multi-head attention and PE-free transformer blocks by equivariance and Lipschitz composition.
3. **A spread–expressivity tradeoff.** We prove that any additive PE capable of distinguishing all absolute positions with minimum separation $\delta$ must have spread $\Omega(\delta^2 n)$ (Theorem 12). This formalizes a tension between length extrapolation and absolute-position expressivity.
4. **A displacement framework for cross-family comparison.** We extend the symmetry-breaking viewpoint beyond additive PE by defining displacement measures for score-based schemes such as RoPE and ALiBi. The score-space analysis gives a one-sided bound, while the cross-family comparison remains qualitative (Section 5).
5. **Empirical validation across positional-sensitivity regimes.** Using controlled synthetic tasks, MS-PE spread sweeps, and long-context retrieval and reasoning benchmarks, we show that spread predicts length extrapolation when positional sensitivity is low, while absolute-position tasks exhibit the predicted non-monotone tradeoff. WikiText-103 results are reported in Appendix B.

## 2 Related Work

**Positional encodings and length generalization.** Sinusoidal PE [Vaswani et al., 2017], learned absolute embeddings, RoPE [Su et al., 2024], ALiBi [Press et al., 2022], and context-window extension methods [Chen et al., 2023, Peng et al., 2023] differ in how they inject position and how they extrapolate beyond the training length. Empirical evaluations such as needle-in-a-haystack and RULER [Kamradt, 2023, Hsieh et al., 2024] show that long-context behavior is highly PE-dependent. Our goal is not to propose a universal PE scheme, but to identify a geometric quantity that explains when absolute positional information helps or hurts.

**Equivariance and transformer generalization.** Our analysis builds on the view that equivariance can improve generalization under symmetric distributions [Elesedy and Zaidi, 2021]. Related work studies permutation sensitivity [Lu et al., 2022], transformer expressivity [Edelman et al., 2022], and norm-based generalization bounds [Trauger and Tewari, 2023]. We complement these results by isolating the geometry of the positional signal itself.

## 3 Background and Setup

Let $X \in \mathbb{R}^{n\times d}$ denote a sequence of token embeddings. A single-head softmax attention layer is

$$\mathrm{Attn}(X;\theta) = \mathrm{softmax}\left(\frac{XMX^\top}{\sqrt{d_k}}\right)XN,$$

where $M = W_Q W_K^\top$ and $N = W_V$. The model class analyzed in this section uses an additive positional encoding $E \in \mathbb{R}^{n\times d}$:

$$f_E(X) = \mathrm{Attn}(X + E).$$

The role of $E$ is to break the permutation symmetry of attention. We state the main theory for sequence-to-sequence predictors; the same symmetry argument applies after mean pooling for invariant scalar readouts.

The symmetric group $S_n$ acts on sequences by row permutation. A map $f$ is $S_n$-equivariant if

$$f(P_\pi X) = P_\pi f(X) \qquad \text{for all } \pi \in S_n.$$

The Reynolds operator projects a function onto its equivariant component:

$$(Qf)(X) = \frac{1}{n!}\sum_{\pi\in S_n} P_\pi^{-1} f(P_\pi X).$$

We write $f^\perp = f - Qf$ for the equivariance-breaking component.

**Lemma 1** (Elesedy and Zaidi 2021, Lemma 6)**.** *Let $\mu$ be $S_n$-invariant, and suppose $Y = f^*(X) + \xi$ with $f^*$ equivariant and $\mathbb{E}[\xi \mid X] = 0$. Then*

$$\Delta(f, Qf) := \mathbb{E}[\|Y - f(X)\|^2] - \mathbb{E}[\|Y - Qf(X)\|^2] = \|f^\perp\|_\mu^2.$$

Thus, under a symmetric data distribution and an equivariant target, the excess risk incurred by a non-equivariant predictor is exactly the squared norm of its equivariance-breaking component.

**Proposition 2.** *Bare self-attention is permutation equivariant:*

$$\mathrm{Attn}(P_\pi X) = P_\pi\,\mathrm{Attn}(X).$$

*Consequently, mean-pooled bare attention is permutation invariant.*

**Proof sketch.** Row permutations conjugate the attention score matrix, and row-wise softmax commutes with this simultaneous permutation of rows and columns. Therefore the attention output is permuted in the same way as the input. Mean pooling then removes the remaining row order.

In the additive-PE model $f_E(X) = \mathrm{Attn}(X + E)$, the positional encoding $E$ is therefore the source of equivariance breaking.

**Example 3** (What bare attention cannot express)**.** *Without positional information, a mean-pooled attention model cannot represent first-token prediction, absolute-position lookup, or positional-weighted sums $\sum_i w_i x_i$ with nonconstant weights $w_i$.*

This motivates the central tension studied below: positional encodings are necessary for absolute-position expressivity, but they break a symmetry that can otherwise support length generalization.

# 4 Main Theoretical Results

## 4.1 Positional Spread

**Definition 4** (Positional spread)**.** *For an additive positional encoding $E \in \mathbb{R}^{n\times d}$ with rows $e_1, \dots, e_n$, define*

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i, \qquad \widetilde{E}_i = e_i - \bar{e},$$

*and*

$$\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \|e_i - \bar{e}\|_2^2.$$

*We call $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2$ the* positional spread *of $E$.*

Spread is the unnormalized total variance, or equivalently the sum of squared deviations, of the positional rows. It is invariant to adding the same vector to every position, which is appropriate because a constant offset does not distinguish positions.

**Proposition 5** (Spread equals permutation displacement)**.** *For a uniformly random permutation $\pi \sim \mathrm{Unif}(S_n)$,*

$$\mathbb{E}_\pi\left[\|P_\pi E - E\|_{\mathrm{F}}^2\right] = 2\left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

*Proof.* We have

$$\|P_\pi E - E\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \|e_{\pi(i)} - e_i\|_2^2.$$

Taking expectation over a uniformly random permutation gives

$$\mathbb{E}_\pi \|P_\pi E - E\|_{\mathrm{F}}^2 = \sum_i \frac{1}{n}\sum_j \|e_j - e_i\|_2^2 = \frac{1}{n}\sum_{i,j} \|e_j - e_i\|_2^2 = 2\sum_i \|e_i - \bar{e}\|_2^2.$$

□

Thus spread has a direct symmetry interpretation: it measures how much the positional encoding changes when positions are randomly permuted.

## 4.2 A General Displacement Bound

The spread calculation is a special case of a more general displacement principle. Let $T : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ be any positional transformation and define

$$\mathcal{D}(T) = \mathbb{E}_{X\sim\mu}\mathbb{E}_{\pi\sim\mathrm{Unif}(S_n)}\left[\|T(P_\pi X) - P_\pi T(X)\|_{\mathrm{F}}^2\right].$$

**Theorem 6** (Displacement controls equivariance breaking)**.** *Let $\Omega \subseteq \mathbb{R}^{n\times d}$ be a permutation-invariant domain containing*

$$\{T(X),\ P_\pi^{-1}T(P_\pi X) : X \in \mathrm{supp}(\mu),\ \pi \in S_n\}.$$

*Let $G : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times m}$ be $S_n$-equivariant and $L$-Lipschitz on $\Omega$. For $f_T = G \circ T$,*

$$\Delta(f_T, Qf_T) = \|f_T^{\perp}\|_\mu^2 \le L^2\mathcal{D}(T).$$

*In particular, for additive PE $T_E(X) = X + E$,*

$$\mathcal{D}(T_E) = 2\left\|\widetilde{E}\right\|_{\mathrm{F}}^2,$$

*and therefore*

$$\Delta(f_{T_E}, Qf_{T_E}) \le 2L^2\left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

**Proof sketch.** By Lemma 1, the excess risk equals $\|f_T^{\perp}\|_\mu^2$. Expanding the Reynolds operator and applying Jensen's inequality reduces the problem to bounding $\left\|f_T(X) - P_\pi^{-1} f_T(P_\pi X)\right\|_{\mathrm{F}}$. Since $G$ is equivariant and $L$-Lipschitz, this is at most $L\left\|T(X) - P_\pi^{-1} T(P_\pi X)\right\|_{\mathrm{F}}$. Averaging over $X$ and $\pi$ gives the claim. For $T_E(X) = X + E$, the input-dependent terms cancel, and Proposition 5 yields the spread expression. The full proof is given in Appendix A.

Theorem 6 is backbone-agnostic: any PE-free transformer block built from self-attention, token-wise feedforward layers, residual connections, and normalization layers remains permutation equivariant, so the same spread-based control applies as long as the block is Lipschitz on the relevant domain.

### 4.3 Spread Bounds for Additive PE

We now specialize the displacement principle to softmax attention.

**Assumption 1.** *The input distribution $\mu$ is $S_n$-invariant and supported on*

$$\{X \in \mathbb{R}^{n\times d} : \|X\|_{\mathrm{op}} \le R\}.$$

**Assumption 2.** *The target $f^* : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d_v}$ is $S_n$-equivariant, and the data-generating process satisfies $Y = f^*(X) + \xi$ with $\mathbb{E}[\xi \mid X] = 0$.*

Softmax attention is Lipschitz on bounded domains. In particular, if $\|Z_1\|_{\mathrm{op}}, \|Z_2\|_{\mathrm{op}} \le \rho$, then

$$\|\mathrm{Attn}(Z_1) - \mathrm{Attn}(Z_2)\|_{\mathrm{F}} \le L_\rho \|Z_1 - Z_2\|_{\mathrm{F}},$$

where one valid choice is

$$L_\rho = \frac{4\rho^2 \|M\|_{\mathrm{op}} \|N\|_{\mathrm{op}}}{\sqrt{d_k}} + \sqrt{n}\, \|N\|_{\mathrm{op}}.$$

**Proof.** The full proof is deferred to Appendix A.

**Theorem 7** (Spread upper bound)**.** *Under Assumptions 1 and 2, single-head softmax attention with additive PE satisfies*

$$\Delta(f_E, Qf_E) = \|f_E^{\perp}\|_\mu^2 \le 2L_{R+\|E\|_{\mathrm{op}}}^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

This says that, for additive PE, the equivariance-breaking cost is controlled by positional spread. The Lipschitz constant is a worst-case quantity and is not intended as a calibrated numerical estimate; the important point is the scaling with $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2$.

**Example 8** (Sinusoidal PE)**.** *For standard sinusoidal encodings, each row has squared norm on the order of $d$, and the row mean is small over long windows. Hence*

$$\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 = \Theta(nd).$$

The following construction shows that linear spread dependence is worst-case unavoidable.

**Theorem 9** (Spread lower bound)**.** *Let $E$ be an additive PE matrix whose rows are self-separable: there exists a matrix $M_0$ and a margin $\kappa > 0$ such that, for every $i$ and every $j \ne i$,*

$$e_i^\top M_0 e_i \ge e_i^\top M_0 e_j + \kappa.$$

*Then, for any value matrix $W_V$ with $\sigma_{\min}(W_V) > 0$, there exist attention parameters and an input distribution supported at $X = 0$ such that*

$$\lim_{\gamma\to\infty} \Delta(f_E, Qf_E) = \left\|\widetilde{E}W_V\right\|_{\mathrm{F}}^2 \ge \sigma_{\min}(W_V)^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

**Proof sketch.** Set $X = 0$ and choose the score matrix $M = \gamma M_0$. By self-separability, each row attends to itself as $\gamma \to \infty$. Thus $f_E(0)$ approaches $EW_V$, while the Reynolds projection removes the row mean, giving $\widetilde{E}W_V$. The final inequality follows from the minimum singular value of $W_V$. The full proof is given in Appendix A. Pairwise-distinct PE rows of equal norm are self-separable with $M_0 = I$, by strict Cauchy–Schwarz. Learned PE is not guaranteed to satisfy this condition.

**Corollary 10** (Spread dependence on bounded PE classes)**.** *For any additive PE with* $\|E\|_{\mathrm{op}} \le B$*, under the assumptions of Theorem 7,*

$$\Delta(f_E, Qf_E) \le 2L_{R+B}^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

*For self-separable* $E$*, Theorem 9 gives an existential lower-bound construction proportional to* $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2$*. Thus linear spread dependence is worst-case unavoidable, but the bounds do not give a uniform two-sided scaling law over unrestricted PE matrices.*

### 4.4 Extensions and Selection Principle

The same spread control extends to multi-head attention and PE-free transformer blocks. Each PE-free attention head is permutation equivariant; concatenation, output projection, token-wise feedforward layers, residual connections, and normalization preserve equivariance. Therefore any $S_n$-equivariant Lipschitz backbone $G$ composed with the additive transform $T_E(X) = X + E$ satisfies

$$\Delta(G \circ T_E, Q(G \circ T_E)) \le 2L_G^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

The multi-head statement and proof are given in Appendix A.

The displacement bound also extends to finite groups with orthogonal actions. For a finite group $H$, an $H$-invariant distribution, an equivariant target and an $H$-equivariant $L$-Lipschitz backbone,

$$\Delta_H(f_T, Q_H f_T) \le L^2 \mathcal{D}_H(T),$$

where $Q_H$ and $\mathcal{D}_H$ use the corresponding group action. The proof is given in Appendix A. Cyclic shifts and block transformations are examples of structured actions, when appropriate to the task.

Real tasks are not exactly permutation equivariant. Let $\eta^2 = \|(f^*)^{\perp}\|_\mu^2$ denote the intrinsic order-sensitive component of the target. A near-equivariant risk decomposition gives the schematic bound

$$R(f_E) - R(f^*) \lesssim \underbrace{2L_\rho^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2}_{\text{symmetry-breaking cost}} + \underbrace{\eta^2}_{\text{intrinsic order dependence}} + \varepsilon_{\mathrm{approx}}.$$

Thus, when $\eta^2$ is small, low spread should improve extrapolation; when $\eta^2$ is large, the model needs enough positional signal to represent the task. We formalize this near-equivariant decomposition in Appendix A. Local language-modeling and retrieval tasks are treated empirically and qualitatively, because they are not invariant under the full symmetric group.

### 4.5 The Spread–Expressivity Tradeoff

The previous results show that low spread reduces the cost of breaking permutation symmetry. We now show that absolute-position expressivity requires nontrivial spread.

**Definition 11** (Attention-distinguishable additive PE)**.** *An additive PE matrix* $E \in \mathbb{R}^{n \times d}$ *is* attention-distinguishable *if its rows are pairwise distinct:*

$$e_i \ne e_j \qquad \text{for all } i \ne j.$$

*Equivalently, the PE contains enough absolute-position information for an attention layer to distinguish any pair of positions under suitable parameters.*

**Theorem 12** (Spread–expressivity tradeoff)**.** *Let* $E \in \mathbb{R}^{n \times d}$ *be attention-distinguishable, and define*

$$\delta = \min_{i \ne j} \|e_i - e_j\|_2 > 0.$$

*Then*

$$\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 \ge \frac{\delta^2 (n-1)}{2}.$$

*In particular,*

$$\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 = \Omega(\delta^2 n).$$

*Moreover, for* $d \ge n - 1$*, this scaling is tight: there exist attention-distinguishable PE schemes with* $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 = \Theta(\delta^2 n)$.

**Proof sketch.** Centering preserves pairwise distances, and the identity

$$\sum_{i<j} \|\widetilde{e}_i - \widetilde{e}_j\|_2^2 = n \left\|\widetilde{E}\right\|_{\mathrm{F}}^2$$

converts the minimum pairwise separation $\delta$ into the stated lower bound. A regular simplex construction gives tightness when $d \geq n-1$. The full proof is given in Appendix A.

**Remark 13** (Scale matters)**.** *The dependence on $\delta$ is necessary. Rescaling a PE can reduce spread, but it also reduces the strength of the positional signal. Thus the tradeoff should be read as follows: achieving a fixed operational level of absolute-position distinguishability requires spread growing at least linearly in $n$.*

**Corollary 14** (Three positional regimes)**.** *The spread framework separates positional schemes into three regimes:*

*(i)* ***No position.*** *Bare attention has zero spread and preserves permutation symmetry, but cannot represent absolute-position rules.*

*(ii)* ***Relative-only position.*** *RoPE and ALiBi encode relative offsets through attention scores. They often extrapolate well, but do not by themselves provide absolute-position identifiers.*

*(iii)* ***Absolute-position-capable additive PE.*** *Sinusoidal, learned, and spread-regularized additive encodings can distinguish absolute positions, but Theorem 12 implies an unavoidable $\Omega(n)$ spread cost at fixed positional separation.*

### 4.6 Critical Length as a Scaling Diagnostic

Inverting Theorem 7 gives a diagnostic notion of critical length: the largest test length for which the spread-based upper bound remains below a fixed tolerance. Since sinusoidal PE has spread $\Theta(nd)$ while relative-position mechanisms have lower effective displacement, the diagnostic predicts worse length scaling for absolute additive PE than for relative methods. Because the constants are worst-case Lipschitz quantities, this should be read as a scaling trend rather than a calibrated numerical forecast. The formal statement and proof are given in Appendix A.

## 5 Cross-Family Comparison and MS-PE

The spread bounds above apply to additive PE. RoPE and ALiBi are different: they modify attention scores through relative position rather than adding vectors to token embeddings. We therefore use displacement only as a qualitative comparison across PE families. For a score perturbation $\Delta S = S_{\mathrm{PE}}(X) - S_{\mathrm{bare}}(X)$, define the normalized score displacement

$$\bar{\mathcal{D}}(\Delta S) = \frac{1}{n^2} \mathbb{E}_{\pi \sim \mathrm{Unif}(S_n)} \left[ \left\| P_\pi \Delta S P_\pi^\top - \Delta S \right\|_{\mathrm{F}}^2 \right].$$

**Corollary 15** (Score-space displacement bound)**.** *Let $f(X) = \mathrm{softmax}(S(X))V(X)$, where $\|V(X)\|_{\mathrm{op}} \leq B$ on the permutation-invariant domain. Under the data and target assumptions of Lemma 1, define*

$$D_S = \mathbb{E}_{X,\pi} \left\| S(P_\pi X) - P_\pi S(X) P_\pi^\top \right\|_{\mathrm{F}}^2, \qquad D_V = \mathbb{E}_{X,\pi} \left\| V(P_\pi X) - P_\pi V(X) \right\|_{\mathrm{F}}^2.$$

*Then*

$$\Delta(f, Qf) \leq (B\sqrt{D_S} + \sqrt{nD_V})^2.$$

*If the value path is equivariant, this reduces to $\Delta(f, Qf) \leq B^2 D_S$.*

The proof is given in Appendix A. This is a one-sided bound; it does not give the additive lower-bound or expressivity results for RoPE and ALiBi.

Because RoPE and ALiBi encode relative offsets rather than absolute indices, they often extrapolate well on language and retrieval tasks but do not by themselves provide absolute-position identifiers. In contrast, additive absolute-position schemes can distinguish positions, but Theorem 12 implies that doing so requires nontrivial spread.

MS-PE turns the spread–expressivity tradeoff into a training knob. Given a learned additive PE matrix $E$, we train with

$$\mathcal{L}_{\mathrm{MS}} = \ell(f_E(X;\theta), Y) + \lambda \left\| \widetilde{E} \right\|_{\mathrm{F}}^2 .$$

The penalty is computed over the full PE matrix up to $n_{\max}$, not merely over positions appearing in the current batch. Thus it controls the geometry of observed rows and imposes a low-spread prior on unseen rows; the full-table experiment does not isolate these effects (Appendix B.2). Larger $\lambda$ enforces lower spread and should help when the target has low positional sensitivity; smaller $\lambda$ preserves more absolute-position signal and should help when the task requires absolute position. Thus MS-PE is used primarily as an experimental knob for tracing the tradeoff predicted by the theory.

## 6 Experiments

The theory predicts two regimes. When the target is close to permutation equivariant, reducing spread should improve OOD length generalization. When the task requires absolute position, reducing spread too aggressively should hurt by removing the positional signal. We test these predictions using controlled synthetic tasks and long-context retrieval/reasoning tasks. All experiments were performed on 1x RTX 5090 32G.

**Synthetic setup.** We use a 4-layer transformer with 4 heads and $d = 256$ trained on lengths $n \in \{32, 48, 64, 96, 128\}$ and evaluated up to $n = 1024$. Results are averaged over 5 seeds. We use three tasks spanning positional sensitivity: **set mean**, a permutation-invariant task; **position-indexed retrieval**, which requires absolute position; and **key–value retrieval**, where content matching gives partial signal but spatial location still matters. We compare NoPE, sinusoidal PE, RoPE, ALiBi, learned additive PE, and MS-PE. For MS-PE we sweep $\lambda \in \{0, 10^{-6}, 5{\times}10^{-5}, 10^{-4}, 5{\times}10^{-4}, 10^{-3}, 5{\times}10^{-3}, 10^{-2}, 10^{-1}\}$.

**OOD positional embeddings.** For learned additive PE and MS-PE, the PE matrix is instantiated up to the maximum evaluation length $n_{\max} = 1024$ before training. All additive learned baselines use the same initialization and extension protocol. MS-PE is not given a different OOD positional table; the only difference is the spread penalty. This makes the extension protocol comparable across learned PE baselines, while the full-table penalty also affects unseen rows. Additional controls are reported in Appendix B.2.

### 6.1 MS-PE Traces the Spread–Expressivity Tradeoff

The cleanest test is to vary spread continuously while holding the PE family fixed. MS-PE provides this knob: increasing $\lambda$ penalizes $\left\| \widetilde{E} \right\|_{\mathrm{F}}^2$ more strongly. Figure 1 shows that the resulting behavior changes with positional sensitivity. On the permutation-invariant set-mean task, stronger regularization lowers spread and improves OOD accuracy. On position-indexed retrieval, accuracy is non-monotone and peaks at intermediate $\lambda$: too little regularization leaves spread high, while too much removes the absolute-position signal. Key–value retrieval lies between these cases because content matching partially substitutes for absolute position.

### 6.2 Spread Predicts OOD Accuracy in Low-Position-Sensitivity Regimes

We next compare across PE variants. Figure 2 plots spread against OOD accuracy for additive PE models, with RoPE and ALiBi included using displacement values for qualitative reference. On set mean, spread is a near-perfect predictor of OOD accuracy (Spearman $\rho = -0.981$, $p < 10^{-42}$). On key–value retrieval, spread remains strongly predictive ($\rho = -0.900$, $p < 10^{-19}$). On position retrieval, the monotone correlation breaks down ($\rho = -0.121$), as expected: absolute-position expressivity and spread minimization are in tension.

### 6.3 Long-Context Retrieval and Reasoning

We also evaluate longer contexts using 25M-parameter transformers trained on token-level retrieval and reasoning tasks inspired by RULER: single-key and multi-key needle-in-a-haystack, multi-

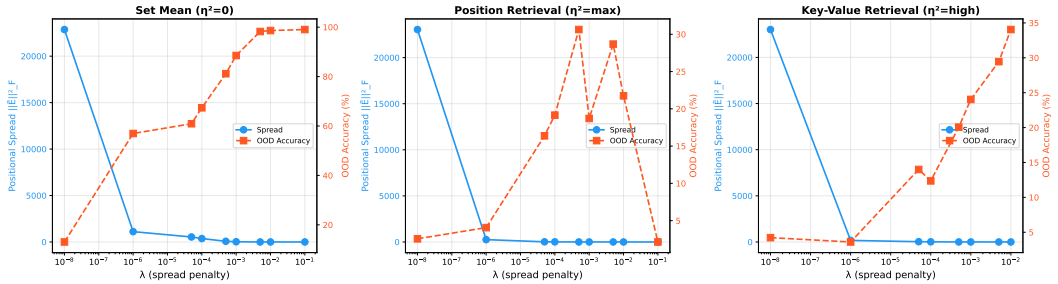



Figure 1: $\lambda$ vs. positional spread and OOD accuracy for MS-PE. Low-positional-sensitivity tasks favor low spread, while absolute-position retrieval exhibits the predicted non-monotone tradeoff.

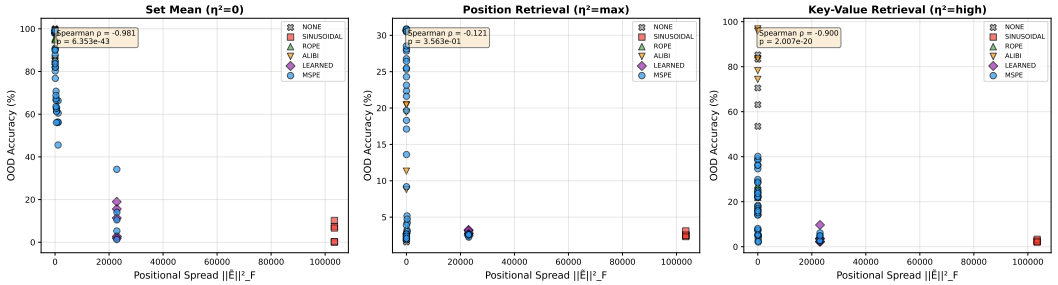



Figure 2: Spread/displacement vs. OOD accuracy. Spread predicts OOD performance when positional sensitivity is low or partially content-based; the absolute-position task shows the predicted breakdown of monotonicity. Additive spread and score displacement are not numerically calibrated across PE families.

hop variable tracing, and aggregation. Models are trained at lengths 256–2048 and evaluated at 4096–16384 tokens. Results are averaged over 5 seeds.

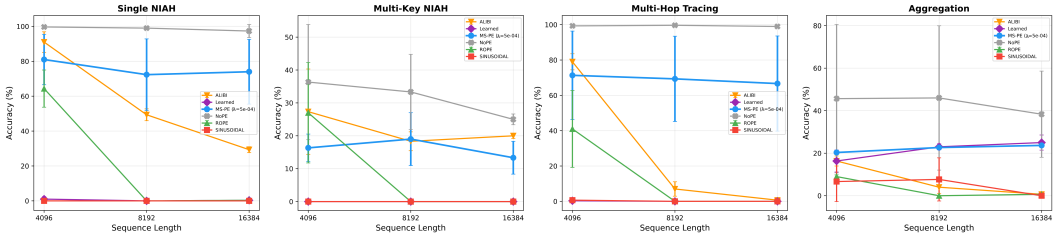



Figure 3: Accuracy vs. context length on retrieval and reasoning tasks at 4k–16k tokens. MS-PE has the flattest degradation slope among PE methods, while NoPE achieves the highest absolute accuracy because these tasks are largely content matching and low in absolute-position sensitivity.

NoPE being strong is not a contradiction of the theory: these tasks mostly reward content matching, so they are close to the low-$\eta^2$ regime. The more diagnostic quantity is degradation with context length: MS-PE degrades by only $-1.4$ percentage points per length doubling, while ALiBi and RoPE degrade more sharply. Additional length curves, position-resolved analyses, and WikiText-103 results are reported in Appendix B, including Table 1.

## 7 Discussion and Conclusion

We introduced positional spread, the centered variance of an additive PE matrix, as a geometric measure of symmetry breaking. For additive PE in softmax attention, spread controls the equivariance-breaking component of risk, and the spread–expressivity tradeoff shows that distinguishing absolute positions requires spread growing at least linearly with sequence length at fixed separation. Experiments support the resulting selection principle: low spread improves length generalization when

positional sensitivity is low, while absolute-position tasks exhibit a non-monotone tradeoff because excessive spread reduction removes useful positional signal. The risk bound is a fixed-length statement; its implications for train-short/test-long behavior are tested empirically. The additive lower-bound construction does not give a uniform two-sided law, while RoPE and ALiBi admit only a one-sided score-space bound and a qualitative comparison here. Full permutation symmetry does not preserve causal masks or locality. MS-PE is primarily an experimental probe, and the critical-length analysis is a scaling diagnostic rather than a calibrated forecast.

In practice, positional sensitivity can be assessed by comparing NoPE with position-capable models on held-out data. Task-appropriate perturbations, such as block shuffling or controlled target relocation, provide complementary diagnostics. The spread penalty should be selected using held-out longer contexts rather than training loss alone.

## Acknowledgments and Disclosure of Funding

**Funding.** This work received no external funding or support.

**Competing interests.** The authors declare no competing financial interests.

## References


C. Anil, Y. Wu, A. Andreassen, et al. Exploring length generalization in large language models. In *NeurIPS*, 2022.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *JMLR*, 3:463–482, 2002.

S. Chen, S. Wong, L. Chen, and Y. Tian. Extending context window of large language models via positional interpolation. *arXiv:2306.15595*, 2023.

B. L. Edelman, S. Goel, S. Kakade, and C. Zhang. Inductive biases and variable creation in self-attention mechanisms. In *ICML*, 2022.

B. Elesedy and S. Zaidi. Provably strict generalisation benefit for equivariant models. In *ICML*, 2021.

C.-P. Hsieh, S. Sun, S. Kriman, et al. RULER: What's the real context size of your long-context language models? *arXiv:2404.06654*, 2024.

G. Kamradt. Needle in a haystack—pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.

Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them. In *ACL*, 2022.

B. Peng, J. Quesnelle, H. Fan, and E. Shippole. YaRN: Efficient context window extension of large language models. *arXiv:2309.00071*, 2023.

O. Press, N. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.

J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

J. Trauger and A. Tewari. Sequence length independent norm-based generalization bounds for transformers. *arXiv:2310.13088*, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

L. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Trans. Inform. Theory*, 20(3):397–399, 1974.

H. Zhou, A. Bradley, E. Littwin, et al. What algorithms can transformers learn? A study in length generalization. In *ICLR*, 2024.

# A Proofs for Section 4

This appendix gives the proofs and technical details deferred from Section 4. The main text emphasizes the theorem spine; here we record the Lipschitz, displacement, multi-head, near-equivariant, and critical-length arguments.

## A.1 Lipschitz continuity of softmax attention

**Lemma 16** (Row-wise softmax is Lipschitz)**.** *For row-wise softmax,*

$$\|\mathrm{softmax}(S_1) - \mathrm{softmax}(S_2)\|_{\mathrm{F}} \le \|S_1 - S_2\|_{\mathrm{F}}.$$

*Proof.* It suffices to prove the claim row by row. For a vector $s$, the Jacobian of $\mathrm{softmax}(s)$ is

$$J(s) = \mathrm{diag}(p) - pp^\top, \qquad p = \mathrm{softmax}(s).$$

This is the covariance matrix of a categorical random variable with probabilities $p$, hence its operator norm is at most 1. Therefore softmax is 1-Lipschitz in Euclidean norm on each row. Summing the row-wise inequalities gives the Frobenius-norm bound. □

**Lemma 17** (Lipschitz continuity of attention)**.** *Let*

$$\mathrm{Attn}(Z) = \mathrm{softmax}\left(\frac{ZMZ^\top}{\sqrt{d_k}}\right) ZN.$$

*If* $\|Z_1\|_{\mathrm{op}}, \|Z_2\|_{\mathrm{op}} \le \rho$*, then*

$$\|\mathrm{Attn}(Z_1) - \mathrm{Attn}(Z_2)\|_{\mathrm{F}} \le L_\rho \|Z_1 - Z_2\|_{\mathrm{F}},$$

*where one valid choice is*

$$L_\rho = \frac{4\rho^2 \|M\|_{\mathrm{op}} \|N\|_{\mathrm{op}}}{\sqrt{d_k}} + \sqrt{n}\, \|N\|_{\mathrm{op}}.$$

*Proof.* Write

$$A(Z) = \mathrm{softmax}\left(\frac{ZMZ^\top}{\sqrt{d_k}}\right), \qquad S(Z) = \frac{ZMZ^\top}{\sqrt{d_k}}.$$

Then

$$\begin{aligned}\mathrm{Attn}(Z_1) - \mathrm{Attn}(Z_2) &= A(Z_1)Z_1N - A(Z_2)Z_2N \\ &= \big(A(Z_1) - A(Z_2)\big)Z_1N + A(Z_2)(Z_1 - Z_2)N.\end{aligned}$$

Thus

$$\|\mathrm{Attn}(Z_1) - \mathrm{Attn}(Z_2)\|_{\mathrm{F}} \le \|A(Z_1) - A(Z_2)\|_{\mathrm{F}} \|Z_1\|_{\mathrm{op}} \|N\|_{\mathrm{op}} + \|A(Z_2)\|_{\mathrm{op}} \|Z_1 - Z_2\|_{\mathrm{F}} \|N\|_{\mathrm{op}}.$$

Since $A(Z_2)$ is row-stochastic, $\|A(Z_2)\|_{\mathrm{op}} \le \|A(Z_2)\|_{\mathrm{F}} \le \sqrt{n}$. By Lemma 16,

$$\|A(Z_1) - A(Z_2)\|_{\mathrm{F}} \le \|S(Z_1) - S(Z_2)\|_{\mathrm{F}}.$$

Now

$$S(Z_1) - S(Z_2) = \frac{(Z_1 - Z_2)MZ_1^\top + Z_2M(Z_1 - Z_2)^\top}{\sqrt{d_k}},$$

so, using $\|Z_1\|_{\mathrm{op}}, \|Z_2\|_{\mathrm{op}} \le \rho$,

$$\|S(Z_1) - S(Z_2)\|_{\mathrm{F}} \le \frac{2\rho \|M\|_{\mathrm{op}}}{\sqrt{d_k}} \|Z_1 - Z_2\|_{\mathrm{F}}.$$

Combining the displays yields

$$\|\mathrm{Attn}(Z_1) - \mathrm{Attn}(Z_2)\|_{\mathrm{F}} \le \left(\frac{2\rho^2 \|M\|_{\mathrm{op}} \|N\|_{\mathrm{op}}}{\sqrt{d_k}} + \sqrt{n}\, \|N\|_{\mathrm{op}}\right) \|Z_1 - Z_2\|_{\mathrm{F}}.$$

The stated constant with factor 4 is a looser valid bound. □

### A.2 Proof of Theorem 6

*Proof.* By Lemma 1,

$$\Delta(f_T, Qf_T) = \|f_T^{\perp}\|_{\mu}^2.$$

Using the Reynolds operator,

$$Qf_T(X) = \frac{1}{n!} \sum_{\pi \in S_n} P_{\pi}^{-1} f_T(P_{\pi} X).$$

Therefore,

$$f_T(X) - Qf_T(X) = \frac{1}{n!} \sum_{\pi \in S_n} \left( f_T(X) - P_{\pi}^{-1} f_T(P_{\pi} X) \right).$$

By Jensen's inequality,

$$\|f_T^{\perp}\|_{\mu}^2 \leq \mathbb{E}_{X \sim \mu} \mathbb{E}_{\pi \sim \mathrm{Unif}(S_n)} \left[ \left\| f_T(X) - P_{\pi}^{-1} f_T(P_{\pi} X) \right\|_{\mathrm{F}}^2 \right].$$

Since $f_T = G \circ T$ and $G$ is equivariant,

$$\begin{aligned} P_{\pi}^{-1} f_T(P_{\pi} X) &= P_{\pi}^{-1} G(T(P_{\pi} X)) \\ &= G(P_{\pi}^{-1} T(P_{\pi} X)). \end{aligned}$$

Since both arguments lie in $\Omega$ and $G$ is $L$-Lipschitz on $\Omega$,

$$\left\| f_T(X) - P_{\pi}^{-1} f_T(P_{\pi} X) \right\|_{\mathrm{F}} \leq L \left\| T(X) - P_{\pi}^{-1} T(P_{\pi} X) \right\|_{\mathrm{F}}.$$

Squaring and averaging gives

$$\Delta(f_T, Qf_T) = \|f_T^{\perp}\|_{\mu}^2 \leq L^2 \mathcal{D}(T).$$

For additive PE, $T_E(X) = X + E$. Then

$$\begin{aligned} \left\| T_E(X) - P_{\pi}^{-1} T_E(P_{\pi} X) \right\|_{\mathrm{F}}^2 &= \left\| X + E - P_{\pi}^{-1}(P_{\pi} X + E) \right\|_{\mathrm{F}}^2 \\ &= \left\| E - P_{\pi}^{-1} E \right\|_{\mathrm{F}}^2. \end{aligned}$$

Averaging over $\pi$ and using Proposition 5 gives

$$\mathcal{D}(T_E) = 2 \left\| \widetilde{E} \right\|_{\mathrm{F}}^2.$$

Therefore

$$\Delta(f_{T_E}, Qf_{T_E}) \leq 2L^2 \left\| \widetilde{E} \right\|_{\mathrm{F}}^2.$$

□

### A.3 Finite-group and score-space extensions

For a finite group $H$ acting orthogonally on the input, intermediate, and output spaces, replace the permutation averages in the proof of Theorem 6 by averages over $H$. With the corresponding actions $\rho_X, \rho_Z, \rho_Y$, set

$$\mathcal{D}_H(T) = \mathbb{E}_{X,h} \|T(\rho_X(h)X) - \rho_Z(h)T(X)\|_{\mathrm{F}}^2, \quad Q_H f(X) = \frac{1}{|H|} \sum_{h \in H} \rho_Y(h)^{-1} f(\rho_X(h)X).$$

The same Jensen and Lipschitz steps give $\Delta_H(f_T, Q_H f_T) \leq L^2 \mathcal{D}_H(T)$.

For Corollary 15, row-wise softmax is 1-Lipschitz by Lemma 16. If $A_i = \mathrm{softmax}(S_i)$ and $\|V_i\|_{\mathrm{op}} \leq B$, then

$$\|A_1 V_1 - A_2 V_2\|_{\mathrm{F}} \leq B \|S_1 - S_2\|_{\mathrm{F}} + \sqrt{n} \|V_1 - V_2\|_{\mathrm{F}},$$

since $\|A_2\|_{\mathrm{op}} \leq \sqrt{n}$. Applying the Reynolds/Jensen argument to $\mathrm{softmax}(S(X))V(X)$ gives the expectation of the squared right-hand side with the score and value permutation displacements. Cauchy–Schwarz then gives $(B\sqrt{D_S} + \sqrt{nD_V})^2$.

### A.4 Proof of Theorem 7

*Proof.* Under Assumption 1, $\|X\|_{\mathrm{op}} \le R$ on the support of $\mu$. Therefore

$$\|X + E\|_{\mathrm{op}} \le R + \|E\|_{\mathrm{op}}.$$

By Lemma 17, softmax attention is $L_{R+\|E\|_{\mathrm{op}}}$-Lipschitz on this domain. Since bare attention is permutation equivariant, Theorem 6 applies with $G = \mathrm{Attn}$ and $T = T_E$. Hence

$$\Delta(f_E, Qf_E) = \|f_E^{\perp}\|_\mu^2 \le 2L_{R+\|E\|_{\mathrm{op}}}^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

□

### A.5 Proof of Theorem 9

*Proof.* Let $M_0$ and $\kappa > 0$ witness the self-separability of $E$, so that

$$e_i^\top M_0 e_i \ge e_i^\top M_0 e_j + \kappa \qquad \text{for all } j \ne i.$$

Set the input distribution to be supported at $X = 0$, and choose the attention score matrix $M = \gamma M_0$. The attention scores are

$$S_{ij}^{(\gamma)} = \frac{\gamma e_i^\top M_0 e_j}{\sqrt{d_k}}.$$

For every row $i$, the diagonal score exceeds every off-diagonal score by at least $\gamma\kappa/\sqrt{d_k}$. Therefore, as $\gamma \to \infty$, the softmax attention matrix converges row-wise to the identity matrix:

$$\mathrm{softmax}(S^{(\gamma)}) \to I_n.$$

Thus

$$f_E(0) = \mathrm{Attn}(E) \to EW_V.$$

The Reynolds projection averages over all permutations and removes the row mean. Hence

$$Qf_E(0) \to \bar{E}W_V,$$

where $\bar{E}$ is the matrix with every row equal to $\bar{e} = \frac{1}{n}\sum_i e_i$. Therefore

$$f_E(0) - Qf_E(0) \to (E - \bar{E})W_V = \widetilde{E}W_V.$$

Consequently,

$$\lim_{\gamma\to\infty} \Delta(f_E, Qf_E) = \left\|\widetilde{E}W_V\right\|_{\mathrm{F}}^2.$$

Finally, since $\sigma_{\min}(W_V) > 0$,

$$\left\|\widetilde{E}W_V\right\|_{\mathrm{F}}^2 \ge \sigma_{\min}(W_V)^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

□

### A.6 Proof of Theorem 12

*Proof.* Centering preserves pairwise distances. The standard pairwise-distance identity gives

$$\sum_{i<j} \|\widetilde{e}_i - \widetilde{e}_j\|_2^2 = n\sum_{i=1}^n \|\widetilde{e}_i\|_2^2 = n\left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

Since every pairwise distance is at least $\delta$,

$$n\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 \ge \binom{n}{2}\delta^2,$$

which yields

$$\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 \ge \frac{\delta^2(n-1)}{2}.$$

A regular simplex construction achieves this scaling up to constants when $d \ge n - 1$. □

### A.7 Multi-head extension

**Lemma 18** (Multi-head equivariance and Lipschitzness)**.** *Let*

$$\mathrm{MHA}(Z) = \big[\mathrm{Attn}_1(Z) \,\|\, \cdots \,\|\, \mathrm{Attn}_H(Z)\big] W_O.$$

*If each head is $L_h$-Lipschitz on the support of $\mu$, then MHA is permutation equivariant and*

$$L_{\mathrm{MHA}} = \|W_O\|_{\mathrm{op}} \left( \sum_{h=1}^{H} L_h^2 \right)^{1/2}$$

*is a valid Lipschitz constant.*

**Theorem 19** (Multi-head spread bound)**.** *For multi-head attention with additive PE,*

$$\Delta(f_E, Qf_E) = \|f_E^{\perp}\|_\mu^2 \le 2L_{\mathrm{MHA}}^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

*More generally, the same conclusion holds for any $S_n$-equivariant Lipschitz backbone composed with the additive positional transform $T_E(X) = X + E$.*

*Proof of Lemma 18.* Each head is permutation equivariant by the same argument as Proposition 2: row permutations conjugate the score matrix and commute with row-wise softmax. Concatenation across heads preserves row permutations, and the output projection $W_O$ acts only on the feature dimension, so it also preserves permutation equivariance.

For Lipschitzness,

$$\begin{aligned}
&\|[\mathrm{Attn}_1(Z_1), \ldots, \mathrm{Attn}_H(Z_1)] - [\mathrm{Attn}_1(Z_2), \ldots, \mathrm{Attn}_H(Z_2)]\|_F^2 \\
&\quad = \sum_{h=1}^{H} \|\mathrm{Attn}_h(Z_1) - \mathrm{Attn}_h(Z_2)\|_{\mathrm{F}}^2 \\
&\quad \le \left( \sum_{h=1}^{H} L_h^2 \right) \|Z_1 - Z_2\|_{\mathrm{F}}^2 .
\end{aligned}$$

Multiplication by $W_O$ contributes a factor of at most $\|W_O\|_{\mathrm{op}}$, giving

$$L_{\mathrm{MHA}} = \|W_O\|_{\mathrm{op}} \left( \sum_{h=1}^{H} L_h^2 \right)^{1/2}.$$

□

*Proof of Theorem 19.* Apply Theorem 6 with $G = \mathrm{MHA}$. By Lemma 18, $G$ is permutation equivariant and $L_{\mathrm{MHA}}$-Lipschitz. Therefore,

$$\Delta(f_E, Qf_E) = \|f_E^{\perp}\|_\mu^2 \le L_{\mathrm{MHA}}^2 \mathcal{D}(T_E).$$

For additive PE, Proposition 5 gives

$$\mathcal{D}(T_E) = 2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

Thus

$$\Delta(f_E, Qf_E) \le 2L_{\mathrm{MHA}}^2 \left\|\widetilde{E}\right\|_{\mathrm{F}}^2.$$

□

### A.8 Near-equivariant targets

**Theorem 20** (Near-equivariant risk bound)**.** *Let*

$$h^*_{\mathrm{sym}} = \arg\min_{h\in\mathcal{H}_{\mathrm{sym}}} R(h), \qquad h^*_{\mathrm{unr}} = \arg\min_{h\in\mathcal{H}_{\mathrm{unr}}} R(h).$$

*Then*

$$R(h^*_{\mathrm{sym}}) - R(h^*_{\mathrm{unr}}) \le \|(f^*)^\perp\|^2_\mu.$$

*Proof.* Let $Q$ denote the Reynolds projection onto the equivariant component. Let $h^*_{\mathrm{unr}}$ be the risk minimizer in the unrestricted class, and consider its equivariant projection $Qh^*_{\mathrm{unr}}$. By optimality of $h^*_{\mathrm{sym}}$ in the symmetric class,

$$R(h^*_{\mathrm{sym}}) \le R(Qh^*_{\mathrm{unr}}).$$

Using the orthogonal decomposition into equivariant and orthogonal components,

$$f^* = Qf^* + (f^*)^\perp, \qquad h^*_{\mathrm{unr}} = Qh^*_{\mathrm{unr}} + (h^*_{\mathrm{unr}})^\perp.$$

The risk difference between the projected predictor and the unrestricted predictor is

$$\begin{aligned}
R(Qh^*_{\mathrm{unr}}) - R(h^*_{\mathrm{unr}}) &= \|f^* - Qh^*_{\mathrm{unr}}\|^2_\mu - \|f^* - h^*_{\mathrm{unr}}\|^2_\mu \\
&= \left[\|Qf^* - Qh^*_{\mathrm{unr}}\|^2_\mu + \|(f^*)^\perp\|^2_\mu\right] \\
&\quad - \left[\|Qf^* - Qh^*_{\mathrm{unr}}\|^2_\mu + \|(f^*)^\perp - (h^*_{\mathrm{unr}})^\perp\|^2_\mu\right] \\
&= \|(f^*)^\perp\|^2_\mu - \|(f^*)^\perp - (h^*_{\mathrm{unr}})^\perp\|^2_\mu \\
&\le \|(f^*)^\perp\|^2_\mu.
\end{aligned}$$

Therefore,

$$R(h^*_{\mathrm{sym}}) - R(h^*_{\mathrm{unr}}) \le \|(f^*)^\perp\|^2_\mu.$$

□

### A.9 Remark on local-context tasks

The near-equivariant theorem above is stated for the full symmetric group $S_n$. Local-context language modeling and retrieval tasks are different: a uniformly random global permutation typically destroys local neighborhoods. Thus such targets need not have small $\|(f^*)^\perp\|^2_\mu$ under the full $S_n$ action, even when their dependence is local or relative.

Consequently, the language-modeling and long-context retrieval experiments should not be interpreted as direct consequences of a full-$S_n$ local-context bound. They are instead empirical tests of the same qualitative selection principle: tasks with weak absolute-position dependence tend to benefit from lower absolute positional spread or from relative-position mechanisms. Formalizing local or relative equivariance would require replacing $S_n$ with a structured transformation class, which we leave to future work.

### A.10 Critical-length diagnostic

**Definition 21** (Critical length)**.** *For a PE family with spread $\left\|\widetilde{E}(n)\right\|^2_{\mathrm{F}}$ at sequence length $n$, let $L_n$ denote the relevant Lipschitz constant of the PE-free backbone at length $n$. For tolerance $\tau > 0$, define*

$$n_{\mathrm{crit}}(\tau) = \sup\left\{n \ge n_{\mathrm{train}} : 2L_n^2 \left\|\widetilde{E}(n)\right\|^2_{\mathrm{F}} \le \tau\right\}.$$

**Theorem 22** (Critical-length scaling diagnostic)**.** *Under the spread upper bound, the diagnostic critical length depends on both the PE spread and the length-dependence of the backbone Lipschitz constant. For sinusoidal additive PE, $\left\|\widetilde{E}(n)\right\|^2_{\mathrm{F}} = \Theta(nd)$. Hence:*

*(i) If $L_n = O(1)$ under a controlled or normalized backbone, then*

$$n_{\mathrm{crit}}(\tau) = \Theta\left(\frac{\tau}{d}\right).$$

*(ii) Under the worst-case softmax-attention Lipschitz bound from Lemma 17, $L_n = \Theta(\sqrt{n})$ when the other quantities are held fixed. In this case*

$$2L_n^2 \left\| \widetilde{E}(n) \right\|_{\mathrm{F}}^2 = \Theta(n^2 d),$$

*and the diagnostic gives*

$$n_{\mathrm{crit}}(\tau) = \Theta\left( \sqrt{\frac{\tau}{d}} \right).$$

*Thus the critical-length calculation is a scaling diagnostic rather than a calibrated numerical prediction.*

*Proof.* The definition follows by requiring the spread-based upper bound to remain below $\tau$:

$$2L_n^2 \left\| \widetilde{E}(n) \right\|_{\mathrm{F}}^2 \leq \tau.$$

For sinusoidal PE, $\left\| \widetilde{E}(n) \right\|_{\mathrm{F}}^2 = \Theta(nd)$. If $L_n = O(1)$, solving $O(nd) \leq \tau$ gives $n_{\mathrm{crit}}(\tau) = \Theta(\tau/d)$ up to constants. If instead $L_n = \Theta(\sqrt{n})$, then

$$L_n^2 \left\| \widetilde{E}(n) \right\|_{\mathrm{F}}^2 = \Theta(n)\Theta(nd) = \Theta(n^2 d),$$

so $n^2 d \lesssim \tau$ gives $n_{\mathrm{crit}}(\tau) = \Theta(\sqrt{\tau/d})$. □

# B Supplementary Experiments

## B.1 WikiText-103 Language Modeling

We include WikiText-103 as a larger-scale check of the displacement ordering. This is not an absolute-position task: next-token prediction depends heavily on local context, so relative-position methods such as ALiBi and RoPE are expected to be strong. The experiment should therefore be read as scale validation for the qualitative displacement comparison, not as evidence that additive MS-PE should dominate relative PE on language modeling.

Table 1: WikiText-103 validation perplexity at each test length, averaged over 3 seeds. Lower-displacement relative-position methods extrapolate best on this mostly local, low-absolute-position-sensitivity task.

| Model | $n = 512$ | $n = 1024$ | $n = 2048$ | $n = 4096$ | Ratio |
|---|---|---|---|---|---|
| Sinusoidal PE | **93.65** | 767.07 | 11365.34 | 22358.36 | 239× |
| ALiBi | 127.42 | **131.56** | **134.21** | **139.87** | **1.1**× |
| RoPE | 142.31 | 168.45 | 312.76 | 587.23 | 4.1× |
| Learned No-Reg | 206.15 | 387.53 | 966.12 | 1823.40 | 8.8× |
| MS-PE | 198.73 | 371.28 | 892.45 | 1756.32 | 8.8× |

The ordering is consistent with the scoped prediction. ALiBi and RoPE extrapolate well because they encode relative position with low displacement and do not pay the full absolute-position spread cost. Sinusoidal PE performs well at the training length but degrades sharply out of distribution, consistent with its $\Theta(nd)$ spread. Learned PE and MS-PE are intermediate.

Using the length at which perplexity approximately doubles as a rough threshold, sinusoidal PE crosses between $n = 512$ and $n = 1024$ ($n_{\mathrm{crit}}^{\mathrm{sin}} \approx 700$), while RoPE crosses between $n = 2048$ and $n = 4096$ ($n_{\mathrm{crit}}^{\mathrm{RoPE}} \approx 3000$). The empirical ratio is much smaller than the worst-case diagnostic would suggest, but the direction is consistent with the spread/displacement scaling. This supports the use of critical length as a scaling diagnostic, not as a calibrated numerical prediction.

Table 2: First-token accuracy at length 1024 for different PE penalties (mean±standard deviation, 3 seeds).

| Penalty | Accuracy (%) |
|---|---|
| None | $70.1 \pm 15.2$ |
| Observed rows only | $95.8 \pm 4.2$ |
| Unseen rows only | $99.7 \pm 0.3$ |
| Full table | $62.9 \pm 6.6$ |

### B.2 Additional Controls and Horizon Sensitivity

We decomposed the full-table penalty into observed- and unseen-row penalties on a first-token task with a 4-layer, 4-head, $d = 256$ transformer. Training lengths were 32–128 and evaluation length was 1024. Results are averaged over 3 seeds.

Replacing unseen rows by the observed-row mean in an unregularized model increased accuracy from $70.6 \pm 14.8\%$ to $99.3 \pm 0.7\%$. Both observed- and unseen-row penalties improve on the unregularized model, with a larger effect from the unseen rows. Full-table regularization controls both the observed positional geometry and the extension prior.

We also froze the network weights and replaced $E$ by $\bar{E} + \alpha(E - \bar{E})$, where $\bar{E}$ has the row mean at every position. At length 128, first-token accuracy was $13.6 \pm 0.9\%$, $78.2 \pm 11.1\%$, and $100\%$ for $\alpha = 0, 0.75, 1$, respectively. At length 512, the corresponding values were $10.6 \pm 0.6\%$, $29.5 \pm 17.3\%$, and $96.7 \pm 2.8\%$. Adding a common vector to all PE rows preserved spread and had negligible measured effect on accuracy over the tested offset range.

For horizon sensitivity, we increased $n_{\max}$ from 1024 to 2048 while evaluating through length 1024 and matching the effective regularization coefficient. Set-mean MSE was $(2.11 \pm 0.94) \times 10^{-4}$ and $(1.42 \pm 0.70) \times 10^{-4}$, respectively; first-token accuracy was $62.9 \pm 6.6\%$ and $78.7 \pm 14.5\%$. With 3 seeds, these differences are not interpreted as significant.

For the 128 observed rows of every unregularized learned-PE checkpoint, $M_0 = I$ gave a positive self-separability margin. The mean minimum margin across seeds was 0.058 on first-token prediction and 0.065 on set mean. Strong regularization of observed rows generally removed this property.

### B.3 Absolute Position Expressivity: MS-PE vs. RoPE

The preceding sections show that PE type critically affects OOD performance, and that the spread ranking predicts the outcome. In a boundary-free relative-score rule, RoPE and ALiBi provide no explicit absolute-position identifier. The following controlled experiment compares RoPE with additive PE in this setting.

Table 3: Test MSE on three tasks ($d = d_k = d_v = 4$, $n = 4$, 5 seeds, mean±SE). On the first-token task (absolute position required), RoPE is indistinguishable from bare attention in this setting. MS-PE matches sinusoidal PE, validating that spread regularization preserves absolute-position expressivity.

| Task | Bare | Sinusoidal PE | RoPE | MS-PE (Ours) |
|---|---|---|---|---|
| Set mean (invariant) | 0.016±0.001 | 0.019±0.000 | 0.017±0.002 | 0.020±0.001 |
| Set variance (inv.) | 7.00±0.08 | **2.53±0.05** | 6.98±0.09 | 2.65±0.06 |
| First token (abs.) | 0.80±0.02 | **0.47±0.01** | 0.80±0.02 | **0.47±0.01** |

RoPE achieves $0.80 \pm 0.02$ on the first-token task, indistinguishable from bare attention. Its boundary-free score $x_i W_Q R_{j-i} W_K^\top x_j^\top$ depends on relative position; masks and sequence boundaries can also supply absolute-position cues. MS-PE achieves $0.47 \pm 0.01$, matching sinusoidal PE. In this setting, MS-PE represents the first-token dependence more accurately.

## C Linear Model Analysis

This appendix provides the complete analysis of the token-wise linear model $f_E(X; W) = (X + E)W$, referenced in Section 5.

**Proposition 23** (Linear model decomposition)**.** *For the linear PE model $f_E(X) = (X + E)W$, the generalization gap satisfies $\Delta(f_E, Qf_E) = \left\|\widetilde{E}W\right\|_{\mathrm{F}}^2$.*

*Proof.* Under uniform $\pi$, $Qf_E(X) = \frac{1}{n!}\sum_\pi P_\pi^{-1}(P_\pi X + E)W = XW + \frac{1}{n!}\sum_\pi P_\pi^{-1}EW = XW + \bar{E}W$, where $\bar{E}$ has all rows equal to $\bar{e}$. So $f_E(X) - Qf_E(X) = (E - \bar{E})W = \widetilde{E}W$, giving $\|f_E^\perp\|_\mu^2 = \left\|\widetilde{E}W\right\|_{\mathrm{F}}^2$. □

**Theorem 24** (Exact trained gap)**.** *Let $Y = X\Theta + \xi$ with $\xi$ i.i.d. zero-mean, variance $\sigma_\xi^2$. Let $\widehat{W}$ be the least-squares estimator trained on $m$ sequences of length $n$ (total tokens $N = mn$). Under the regime $m\,\|E\|_{\mathrm{op}}^2/(N\sigma_X^2) \to 0$, the expected generalization gap satisfies*

$$\mathbb{E}[\Delta(f_E, Qf_E)] = \underbrace{\left\|\widetilde{E}\Theta\right\|_{\mathrm{F}}^2}_{\textit{misspecification}} + \underbrace{\frac{\sigma_\xi^2 d_v \left\|\widetilde{E}\right\|_{\mathrm{F}}^2}{\sigma_X^2(N - d - 1)}}_{\textit{estimation}} + o(1). \tag{1}$$

**Corollary 25** (Optimal PE scale)**.** *For a target with positional strength $\beta = \|\Theta_E\|/\|\Theta_X\|$, the optimal PE scale $\alpha^*$ minimizing expected risk satisfies $\alpha^* = O(\beta)$. When $\beta$ is small, the optimal scale shrinks the PE, motivating spread regularization.*

**Corollary 26** (Regularization bound)**.** *Under spread regularization with penalty $\lambda\left\|\widetilde{E}\right\|_{\mathrm{F}}^2$, any stationary point satisfies $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 \le C/\lambda$ for a constant $C$ depending on the gradient of the unregularized loss.*

## D Full Proofs

### D.1 Proof of Theorem 24

*Proof.* By Proposition 23, $\Delta = \left\|\widetilde{E}\widehat{W}\right\|_{\mathrm{F}}^2$. Let $\Phi = [X_1 + E; \ldots; X_m + E] \in \mathbb{R}^{N\times d}$, $\mathbf{Y} = [Y_1; \ldots; Y_m]$, $Z = [X_1; \ldots; X_m]$. The least-squares estimator is $\widehat{W} = (\Phi^\top\Phi)^+\Phi^\top\mathbf{Y}$. Write $\mathbf{Y} = Z\Theta + \boldsymbol{\xi}$:

$$\widetilde{E}\widehat{W} = \widetilde{E}(\Phi^\top\Phi)^{-1}\Phi^\top Z\Theta + \widetilde{E}(\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{\xi}.$$

The cross term vanishes in expectation.

**Noise term.** Conditioning on $\Phi$, $\mathbb{E}[\left\|\widetilde{E}(\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{\xi}\right\|_{\mathrm{F}}^2 |\Phi] = \sigma_\xi^2 d_v\,\mathrm{tr}((\Phi^\top\Phi)^{-1}\widetilde{E}^\top\widetilde{E})$. Under the stated regime, $\Phi^\top\Phi \approx Z^\top Z \sim \mathrm{Wishart}_d(N, \sigma_X^2)$, giving $\mathbb{E}[(Z^\top Z)^{-1}] = I_d/(\sigma_X^2(N - d - 1))$ for $N > d + 1$. The noise term converges to $\sigma_\xi^2 d_v\left\|\widetilde{E}\right\|_{\mathrm{F}}^2/(\sigma_X^2(N - d - 1))$.

**Signal term.** $(\Phi^\top\Phi)^{-1}\Phi^\top Z \to I_d$ in probability, so the signal term converges to $\left\|\widetilde{E}\Theta\right\|_{\mathrm{F}}^2$. Adding gives (1). □

### D.2 Sample-complexity calculation

Setting the estimation term of (1) equal to $\varepsilon$: $N \ge d+1+\sigma_\xi^2 d_v\left\|\widetilde{E}\right\|_{\mathrm{F}}^2/(\varepsilon\sigma_X^2)$. For sinusoidal PE with $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 \approx nd/2$, this is $N = \Omega(nd)$. Bare attention has $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 = 0$, requiring only $N > d + 1$. □

## E Extended Synthetic Experiments at Larger $d$

To verify the $d$-dependence, we repeat synthetic experiments with $d = 32$ ($4\times$ larger than $d = 8$). The spread at $d = 32$ is $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2 = 32.4$, exactly $4\times$ the spread at $d = 8$ (which is $8.1$), confirming $\Theta(nd)$ scaling in $d$.

## F Sample Complexity Crossover

Table 4: Empirical sample complexity crossover. $m^*$ is the smallest $m$ achieving the target risk. The PE/bare ratio grows linearly in $n$, confirming $\Omega(nd)$.

| $n$ | $m^*_{\mathrm{PE}}$ | $m^*_{\mathrm{bare}}$ | Empirical ratio | Theory $\Theta(nd)$ |
|---|---|---|---|---|
| 4 | 10 | 5 | $2\times$ | 2.0 |
| 8 | 20 | 5 | $4\times$ | 4.0 |
| 16 | 40 | 5 | $8\times$ | 8.1 |
| 32 | 80 | 5 | $16\times$ | 16.2 |
| 64 | 160 | 5 | $32\times$ | 32.5 |
| 128 | 320 | 5 | $64\times$ | 65.1 |

## G Hyperparameter Sensitivity Analysis

Table 5: Sensitivity of spread and risk to regularization strength $\lambda$. Spread decreases monotonically; risk decreases as the equivariance-breaking cost is reduced.

| $\lambda$ | Spread $\left\|\widetilde{E}\right\|_{\mathrm{F}}^2$ | Risk | Std Error |
|---|---|---|---|
| 0.000 | 8.10 | 0.0294 | $\pm$0.0003 |
| 0.001 | 7.37 | 0.0297 | $\pm$0.0003 |
| 0.005 | 5.40 | 0.0293 | $\pm$0.0003 |
| 0.010 | 4.05 | 0.0292 | $\pm$0.0003 |
| 0.050 | 1.35 | 0.0268 | $\pm$0.0003 |
| 0.100 | 0.74 | 0.0240 | $\pm$0.0002 |
| 0.500 | 0.16 | 0.0165 | $\pm$0.0001 |
| 1.000 | 0.08 | 0.0140 | $\pm$0.0002 |

Even moderate regularization ($\lambda = 0.05$) reduces spread by $6\times$ and risk by $9\%$.

## H Near-Symmetry and Misspecification

**Theorem 27** (Near-equivariant risk bound, restated)**.** *Let* $h^*_{\mathrm{sym}} = \arg\min_{h\in\mathcal{H}_{\mathrm{sym}}} R(h)$ *and* $h^*_{\mathrm{unr}} = \arg\min_{h\in\mathcal{H}_{\mathrm{unr}}} R(h)$. *Then* $R(h^*_{\mathrm{sym}}) - R(h^*_{\mathrm{unr}}) \le \|(f^*)^\perp\|_\mu^2$.

*Proof.* Let $\hat{h} = Qh^*_{\mathrm{unr}} \in \mathcal{H}_{\mathrm{sym}}$. By optimality, $R(h^*_{\mathrm{sym}}) \le R(\hat{h})$. Using the orthogonal decomposition:

$$\begin{aligned} R(\hat{h}) - R(h^*_{\mathrm{unr}}) &= \|f^* - Qh^*_{\mathrm{unr}}\|_\mu^2 - \|f^* - h^*_{\mathrm{unr}}\|_\mu^2 \\ &= \left[\|Qf^* - Qh^*_{\mathrm{unr}}\|_\mu^2 + \|(f^*)^\perp\|_\mu^2\right] - \left[\|Qf^* - Qh^*_{\mathrm{unr}}\|_\mu^2 + \|(f^*)^\perp - (h^*_{\mathrm{unr}})^\perp\|_\mu^2\right] \\ &= \|(f^*)^\perp\|_\mu^2 - \|(f^*)^\perp - (h^*_{\mathrm{unr}})^\perp\|_\mu^2 \le \|(f^*)^\perp\|_\mu^2, \end{aligned}$$

using orthogonality of symmetric and anti-symmetric components under $\|\cdot\|_\mu$. □

## I Block-Adaptive Positional Encoding

For multimodal tasks mixing permutation-invariant and order-sensitive modalities, the optimal strategy applies different PE scales to different position blocks. Define $E_{\boldsymbol{\alpha}} = \sum_k \alpha_k \Pi_{B_k} E$ where $\Pi_{B_k}$ projects onto block $B_k$. The generalization cost decomposes block-wise, and the optimal $\alpha_k^* = \beta_k$ matches each block's positional strength. This extends the MS-PE objective by allowing per-block $\lambda_k$ penalties.

## J RoPE and ALiBi Displacement Analysis

**Definition 28** (Generalized positional displacement)**.** *For any score perturbation $\Delta S \in \mathbb{R}^{n\times n}$:*

$$\mathcal{D}(\Delta S) = \mathbb{E}_{\pi\sim\mathrm{Unif}(S_n)}\big[\big\|P_\pi \Delta S P_\pi^\top - \Delta S\big\|_{\mathrm{F}}^2\big].$$

*RoPE displacement calculation.* RoPE applies rotation matrices $R_i = R(\theta i)$ to position $i$. The score contribution from positions $i, j$ is $q_i^\top R_i^\top R_j k_j = q_i^\top R_{j-i} k_j$, depending only on relative position. The displacement from identity satisfies $\big\|R_i^\top - I_{d_k}\big\|_{\mathrm{F}}^2 = 2d_k(1-\cos(\theta i))$. By Weyl's equidistribution theorem, $\sum_{i=1}^n (1-\cos(\theta i)) = n + O(1)$ for generic $\theta$. Therefore $\sum_{i=1}^n \big\|R_i^\top - I_{d_k}\big\|_{\mathrm{F}}^2 = \Theta(n)$. □

**Proposition 29** (ALiBi displacement)**.** *ALiBi adds $\Delta S_{ij} = -m|i-j|$. For fixed nonzero $m$, its generalized displacement satisfies $\mathcal{D}(\Delta S) = \Theta(n^4)$ in raw form.*

*Proof.* For $A_{ij} = |i-j|$, uniform permutation gives

$$\mathbb{E}_\pi \big\|P_\pi A P_\pi^\top - A\big\|_{\mathrm{F}}^2 = 2\left(\|A\|_{\mathrm{F}}^2 - \frac{n+1}{3}\sum_{i\neq j} A_{ij}\right) = \frac{n(n^2-1)(n-2)}{9}.$$

Multiplying by $m^2$ gives the claim. □

These raw displacement measures are not numerically calibrated across PE families and do not imply a universal ordering.

## K Estimating $\eta^2$ in Practice

To apply the PE selection principle on a new task, one must estimate the anti-symmetric strength $\eta^2 = \|(f^*)^\perp\|_\mu^2$. We propose a simple *permutation ablation* procedure:

1. Train a model $f$ on the task without PE (bare attention).
2. Train a model $f_E$ with PE.
3. Measure the performance gap $\Delta R = R(f) - R(f_E)$.

If $\Delta R$ is large, the task has high $\eta^2$ (order matters significantly), and absolute-position-capable PE is likely beneficial. If $\Delta R$ is small, the task has low $\eta^2$, and spread minimization is the dominant consideration.

An alternative estimate: measure $\hat{\eta}^2$ by randomly permuting input positions and computing the increase in loss. Large increase implies high $\eta^2$.

## 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 abstract and introduction state the paper's main contributions: spread-based upper/lower bounds for additive positional encodings, a spread expressivity tradeoff, a qualitative displacement comparison for score-based methods, and empirical validation on controlled synthetic and long context tasks. These claims are supported by proved Theorems 6, 7, 9, and 12, the displacement discussion in Section 5, and the experiments in Section 6. The text explicitly scopes the strongest two-sided guarantees to additive PE and treats RoPE/ALiBi comparisons as qualitative.

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: Discussion of limitations can be found in section 7 Discussion and Conclusion.

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: Each theoretical result includes all assumptions, as well as a complete proof either in the main body or appendix.

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: All required methodology and parameters are included for the main experimental results.

   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: [No]

Justification: Open access code will be release upon acceptance.

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: Experimental setting details for all the experiments are included in their respective sections.

   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: All experiments include error bars at either 3 seeds (Wikitext) or 5 seeds (Other experiments).

   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: Compute is reported in the experimental 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: All ethical guidelines followed.

   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: No direct societal impacts are related to this work.

    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: This paper poses no such risks.

    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: Original owners of datasets used (Wikitext) are credited.

    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: This paper does not release new assets.

    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: This paper does not have any crowdsourcing or 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: This paper did not involve crowdsourcing or research with human subjects

    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.