ONE LEARNING RATE, MANY POSITIONAL STEPS: POSITION-AWARE ROPE PRECONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rotary positional embeddings (RoPE) assign different frequencies to two-dimensional query/key pairs, yet standard optimizers ignore this structure. We identify a potential *RoPE-induced learning-rate mismatch*: the same parameter-space turn can induce different positional steps across pairs, with sensitive pairs potentially limiting the stable global learning rate. We derive an analytic sensitivity for each pair and show that the same frequency-dependent structure appears in a RoPE-specific component of the local generalized Gauss–Newton geometry. Based on this, we propose *Position-Aware Preconditioning* (PA), which preserves the scale-changing part of a base-optimizer update and recalibrates only its turning part using RoPE information. PA requires no extra forward/backward pass or per-parameter optimizer state and retains first-order convergence guarantees for its gradient specialization. On C4 pretraining with 130M- and 350M-parameter LLaMA-style models, PA improves AdamW and Muon, reaching baseline training-loss levels with 13–27% fewer optimizer updates.

1 INTRODUCTION

Transformer attention decides which tokens should interact by comparing their query and key representations (Vaswani et al., 2017). Rotary positional embeddings (RoPE) add relative position to this comparison by grouping query and key coordinates into two-dimensional pairs and rotating each pair according to token position (Su et al., 2021). A simple way to picture each RoPE pair is as a clock hand: its length controls how strongly that pair contributes, its direction determines its current orientation, and its RoPE frequency determines how quickly the hand rotates as the distance between two tokens changes. Although the RoPE rotations themselves are fixed, the query and key projections that produce these clock hands are learned. Training therefore changes not only the content represented by the model, but also the vectors through which RoPE expresses relative position.

This creates a natural optimization problem. An optimizer changes model parameters, but what ultimately matters is how much those changes alter model behavior. AdamW (Loshchilov & Hutter, 2019) and Muon (Liu et al., 2025a), for example, adapt updates using gradient statistics or matrix structure, but do not use the RoPE frequency attached to each query/key pair. An update to a RoPE pair can either change the hand's length or turn it: the former mainly changes feature strength, whereas the latter changes how the pair participates in relative-position computation. Since different pairs rotate at different frequencies, the same turn need not produce the same positional effect. A common parameter-space update rule can therefore yield very different positional step sizes across RoPE pairs, creating a hidden imbalance in how aggressively they are updated.

Frequency differences alone, however, are not enough to justify modifying the optimizer. We first need to establish that comparable turns indeed produce substantially different positional effects across RoPE pairs, and to quantify how large this mismatch can be. Even then, such a geometric difference matters for optimization only if it also appears in the actual training objective and affects the step sizes that training can safely take. Only after these two connections are established does it become meaningful to ask how the mismatch should be corrected. Ideally, the correction should use the RoPE information already known from the architecture, while preserving the update learned by a strong base optimizer and adding little training overhead. This would turn RoPE's fixed frequency structure from a representation mechanism into useful optimization guidance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Contribution. We address these challenges in three steps: we first quantify the turn sensitivity of each RoPE pair, then connect this frequency-dependent structure to the local geometry of the training objective, and finally introduce Position-Aware Preconditioning (PA), a lightweight wrapper that preserves the base optimizer's local scale component while recalibrating its turn.

First, we identify a *RoPE-induced learning-rate mismatch*: the same parameter-space step can produce very different positional steps across RoPE pairs, making sensitive pairs potential training bottlenecks. For pair i with frequency θ_i, define $S_i = \mathbb{E}_{h\sim\pi}\left[4\sin^2\left(\frac{h\theta_i}{2}\right)\right]$, where h is relative token distance and π its distribution. A scale change leaves the scale-normalized positional behavior unchanged, whereas a small turn ϕ yields a squared positional-response change $\mathcal{O}(S_i\phi^2)$. Thus, S_i measures how strongly the same turn moves pair i in positional behavior, with more sensitive pairs potentially constraining a shared learning rate.

Second, we show that the sensitivity S_i also enters the actual training objective, so the mismatch is not an artifact of our positional measure. For pair i, the RoPE-modulated attention interaction contains the rotation $R(h\theta_i)$. We isolate the RoPE-specific part of its Jacobian and show that it carries the factor $2\sin(h\theta_i/2)$; averaging its squared effect over relative distances therefore recovers the same sensitivity S_i. Propagating this response via a generalized Gauss–Newton geometry (Martens, 2020) yields an S_i-dependent local loss-curvature component. So the architecture-known RoPE sensitivity directly informs the real optimization geometry, providing a principled signal for calibrating frequency-dependent updates.

Third, we propose *Position-Aware Preconditioning* (PA) to correct this RoPE-induced step-size imbalance while preserving the base optimizer's update. For pair i, PA keeps the scale component unchanged and rescales only the turn by $p_i = (1 + \kappa S_i)^{-1}$, where $\kappa \geq 0$ controls the correction strength. This changes the effective positional learning-rate coefficient from S_i to $S_i/(1+\kappa S_i)$, suppressing overly sensitive pairs and reducing their step-size bottleneck. PA is a lightweight wrapper for optimizers such as AdamW and Muon: it requires no extra forward/backward pass or persistent curvature state and retains first-order convergence guarantees for its gradient specialization.

Finally, we evaluate whether this architecture-aware correction translates into a meaningful training benefit. On C4 pretraining (Raffel et al., 2020) with 130M- and 350M-parameter LLaMA-style models (Touvron et al., 2023), PA improves both AdamW and Muon across the four reported model-optimizer settings, reaching the corresponding baseline training-loss levels with approximately 13–27% fewer optimizer updates, with modest per-step overhead and no measurable increase in peak allocated GPU memory.

1.1 Related Work

Optimization for large-scale language models. Adam (Kingma & Ba, 2015) and AdamW adapt updates using coordinate-wise gradient statistics and remain widely used for Transformer pretraining and fine-tuning (Zhao et al., 2024). Recent methods exploit richer parameter structure: Shampoo and related approaches use matrix-based statistics for preconditioning (Gupta et al., 2018; Frans et al., 2025; Liu et al., 2025b), SOAP combines matrix preconditioning with Adam-style adaptation (Vyas et al., 2024), and Muon applies Newton–Schulz orthogonalization to matrix-structured gradients. Other examples include SWAN (Ma et al., 2024), MARS (Yuan et al., 2024), MARS-M (Liu et al., 2025c), and Sophia (Liu et al., 2024). These mainly exploit statistical or algebraic parameter structure; our work instead uses functional structure specified directly by model architecture.

Rotary positional embeddings. RoPE was introduced in RoFormer as a positional encoding mechanism that applies position-dependent rotations to query and key representations, making their interaction depend on relative position (Su et al., 2021). It has since become standard in decoder-only model families such as LLaMA and Qwen (Touvron et al., 2023; Team, 2024). Prior work has mainly studied context extension and positional behavior, including Position Interpolation, YaRN, LongRoPE, and analyses of how attention uses different frequency bands (Chen et al., 2023; Peng et al., 2023; Ding et al., 2024; Barbero et al., 2024). Differently, we keep RoPE fixed and study how its known frequency structure shapes the optimization geometry of the trainable queries and keys.

2 RoPE-Induced Positional Step Imbalance

We begin with the first issue raised in Sec. 1: how much positional behavior does a parameter update change? Separating a change in length from a turn reveals a frequency-dependent response. We quantify this structure directly and briefly illustrate its possible optimization consequence.

From RoPE pairs to trainable parameters. Consider one attention head with even dimension d_h and let $N = d_h/2$. For token representations $\mathbf{x}_m, \mathbf{x}_n \in \mathbb{R}^{d_{\text{model}}}$, their query and key are $\mathbf{q}_m = \mathbf{W}_Q^\top \mathbf{x}_m$ and $\mathbf{k}_n = \mathbf{W}_K^\top \mathbf{x}_n$, where $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_{\text{model}} \times d_h}$. RoPE applies block-diagonal rotations $\widetilde{\mathbf{q}}_m = \mathbf{R}_m \mathbf{q}_m$ and $\widetilde{\mathbf{k}}_n = \mathbf{R}_n \mathbf{k}_n$, with $\mathbf{R}_m = \text{blkdiag}(\mathbf{R}(m\theta_1), \ldots, \mathbf{R}(m\theta_N)) \in \mathbb{R}^{d_h \times d_h}$ and $\mathbf{R}(\alpha) = [\cos\alpha, -\sin\alpha; \sin\alpha, \cos\alpha] \in \mathbb{R}^{2\times 2}$.

For the relative displacement $h := n - m$, the unscaled attention interaction satisfies $\widetilde{\mathbf{q}}_m^\top \widetilde{\mathbf{k}}_n = \mathbf{q}_m^\top \mathbf{R}_h \mathbf{k}_n$, where $\mathbf{R}_h = \mathbf{R}_m^\top \mathbf{R}_n = \text{blkdiag}(\mathbf{R}(h\theta_1), \ldots, \mathbf{R}(h\theta_N))$. Therefore, writing $\mathbf{q}_m = [(\mathbf{q}_m^{(1)})^\top, \ldots, (\mathbf{q}_m^{(N)})^\top]^\top$ and analogously for $\mathbf{k}_n$, with $\mathbf{q}_m^{(i)}, \mathbf{k}_n^{(i)} \in \mathbb{R}^2$, gives

$$\mathbf{q}_m^\top \mathbf{R}_h \mathbf{k}_n = \sum\nolimits_{i=1}^{N} \left(\mathbf{q}_m^{(i)}\right)^\top \mathbf{R}(h\theta_i) \mathbf{k}_n^{(i)}, \qquad \mathbf{q}_m^{(i)}, \mathbf{k}_n^{(i)} \in \mathbb{R}^2, \;\; h = n - m. \tag{1}$$

Each *RoPE pair* therefore refers to a two-dimensional output subspace, rather than to an individual trainable parameter. The trainable parameters that feed this subspace are the two-entry row slices

$$\mathbf{w}_{p,i,j} = ([\mathbf{W}_p]_{j,2i-1}, [\mathbf{W}_p]_{j,2i})^\top \in \mathbb{R}^2, \qquad p \in \{Q, K\}, \;\; j \in \{1, \cdots, d_{\text{model}}\}, \tag{2}$$

which we call *RoPE-aligned parameter cells*. For example, $\mathbf{q}_m^{(i)} = \sum_j x_{m,j} \mathbf{w}_{Q,i,j}$. Hence, all parameter cells associated with RoPE pair i inherit the same frequency θ_i. This shared frequency label provides the parameter-level structure used throughout the analysis below. For the standard schedule, $\theta_i = b^{-2(i-1)/d_h}$, $i = 1, \ldots, N$, with base $b > 1$ (Su et al., 2021).

Measuring the Positional Effect of a Turn. Consider a nonzero RoPE-aligned parameter cell $\mathbf{w} \in \mathbb{R}^2$ in Eq. (2). Let $r = \|\mathbf{w}\|_2$ and $\mathbf{u} = \mathbf{w}/r$. A local update can either change the cell's magnitude or turn its direction, which we represent as

$$\mathbf{w}(s, \phi) = (1 + s)\mathbf{R}(\phi)\mathbf{w}, \qquad (s > -1). \tag{3}$$

Here s controls scale and ϕ is the turning angle. To first order, these are precisely the radial and angular components of an arbitrary cell perturbation; the derivation is given in Appendix A.1.

We are interested in how the *direction* of this cell participates in RoPE, independently of its magnitude. So we normalize the cell and measure the change introduced by the relative-position rotation:

$$\boldsymbol{\psi}_{i,h}(\mathbf{w}) := \big(\mathbf{R}(h\theta_i) - \mathbf{I}\big)\mathbf{w}/\|\mathbf{w}\|_2. \tag{4}$$

Normalization removes the cell magnitude, while $\mathbf{R}(h\theta_i) - \mathbf{I}$ measures the RoPE-induced change in its direction. Thus, positive rescaling leaves $\boldsymbol{\psi}_{i,h}$ unchanged, whereas a turn remains because $\mathbf{w}(s, \phi)/\|\mathbf{w}(s, \phi)\|_2 = \mathbf{R}(\phi)\mathbf{u}$. For a fixed displacement h, $\boldsymbol{\psi}_{i,h}(\mathbf{w}(s, \phi)) - \boldsymbol{\psi}_{i,h}(\mathbf{w})$ therefore measures the positional-response change caused by the update, with its squared norm measuring the change size. Since token pairs occur at different relative displacements, we average this quantity over a distribution π, where $\pi(h)$ is the weight of displacement h. We define

$$D_i(\phi) := \mathbb{E}_{h\sim\pi}\left[\|\boldsymbol{\psi}_{i,h}(\mathbf{w}(s, \phi)) - \boldsymbol{\psi}_{i,h}(\mathbf{w})\|_2^2\right]. \tag{5}$$

So $D_i(\phi)$ is the average squared positional-response change produced by turning a cell of RoPE pair i. It depends only on ϕ, as positive rescaling is removed by normalization. The following proposition shows that this change is governed by the pair's RoPE frequency with its proof in Appendix A.1.

Proposition 2.1 (The same turn has different positional effects)**.** *For every nonzero cell associated with pair i, every $s > -1$, and every $\phi \in \mathbb{R}$,*

$$D_i(\phi) = 4S_i \sin^2(\phi/2) = S_i\phi^2 + \mathcal{O}(S_i\phi^4), \qquad \textit{where} \;\; S_i := \mathbb{E}_{h\sim\pi}\left[4\sin^2\left(h\theta_i/2\right)\right]. \tag{6}$$

In particular, positive rescaling alone gives $D_i(0) = 0$.

Proposition 2.1 has a simple interpretation. Positive rescaling does not change the normalized positional response, but turning the cell does. For a small turn ϕ, the average squared response change is approximately $S_i\phi^2$, so the root-mean-square response change scales as $\sqrt{S_i}|\phi|$. Hence, the same parameter-space turn can produce different positional steps across RoPE pairs. The coefficient S_i is determined by the fixed RoPE frequency θ_i and the relative-distance distribution π. Because the rotary factor is bounded and periodic, S_i need not be strictly ordered by frequency alone.

Illustration example. For $S_i > 0$, let $z_i = \sqrt{S_i}\phi_i$ be a local positional coordinate. Even if a loss has the same curvature in z_i for every pair, its curvature in the turn coordinate ϕ_i scales with S_i. Thus, a more sensitive pair can impose a tighter restriction on a shared turn-coordinate step. Appendix A.2 gives the complete quadratic example and recurrence. This is an illustration, not a model of the full Transformer loss; Sec. 3 examines the actual attention-to-loss structure.

3 How Positional Sensitivity Enters the Training Loss

Sec. 2 measures a cell's normalized positional response. We now address the second issue raised in Sec. 1: does the same rotary structure appear when cells contribute to attention and the training loss? We first follow a cell turn into one attention score, then combine its effects across token distances, and finally weight these score changes by their effect on the loss.

3.1 How a Cell Turn Changes an Attention Score

A cell affects attention through its query or key. Fix a head's input activations and all parameters except its query/key projections. Using the cells in Eq. (2), pair i contributes the unscaled score

$$a_{mn}^{(i)} = (\mathbf{q}_m^{(i)})^\top \mathbf{R}(h\theta_i)\mathbf{k}_n^{(i)}, \text{ where } \mathbf{q}_m^{(i)} = \sum\nolimits_j x_{m,j}\mathbf{w}_{Q,i,j},\ \mathbf{k}_n^{(i)} = \sum\nolimits_\ell x_{n,\ell}\mathbf{w}_{K,i,\ell},\ h = n-m. \quad (7)$$

Thus, perturbing a cell changes the query or key contribution and hence the attention score.

For a nonzero cell, retain $r = \|\mathbf{w}\|_2$ and $\mathbf{u} = \mathbf{w}/r \in \mathbb{R}^2$ from Sec. 2, and set $\mathbf{e}_r = \mathbf{u}$ and $\mathbf{e}_\phi = \mathbf{J}\mathbf{u}$, where $\mathbf{J} = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ rotates a vector by $90°$. These are the local scale and turn directions. A cell perturbation is $\boldsymbol{\xi}_{p,i,j} = r_{p,i,j}(s_{p,i,j}\mathbf{e}_{r,p,i,j} + \phi_{p,i,j}\mathbf{e}_{\phi,p,i,j})$, with $p \in \{Q, K\}$ (see the meaning of indices i and j in Eq. (2)). Write $\delta a_{mn}^{(i)}$ for the first-order score change under these perturbations, rather than the exact score difference after a finite update. For example, turning only query cell $\mathbf{w}_{Q,i,j}$ gives $\delta\mathbf{q}_m^{(i)} = x_{m,j}\phi_{Q,i,j}\mathbf{J}\mathbf{w}_{Q,i,j}$. Splitting $\mathbf{R}(h\theta_i) = \mathbf{I} + (\mathbf{R}(h\theta_i) - \mathbf{I})$ yields

$$\frac{\partial a_{mn}^{(i)}}{\partial \phi_{Q,i,j}}\Big|_0 = x_{m,j}(\mathbf{J}\mathbf{w}_{Q,i,j})^\top \mathbf{R}(h\theta_i)\mathbf{k}_n^{(i)} = x_{m,j}(\mathbf{J}\mathbf{w}_{Q,i,j})^\top \mathbf{k}_n^{(i)} + x_{m,j}(\mathbf{J}\mathbf{w}_{Q,i,j})^\top \big(\mathbf{R}(h\theta_i) - \mathbf{I}\big)\mathbf{k}_n^{(i)}.$$

Here 0 means no additional turn at the current state. The first term in the split is the response without relative rotation; the second captures how RoPE changes that response at fixed inputs and parameters. The latter contains the same rotation-difference operator used in Eq. (4), now within an actual attention score derivative. The analogous key-cell derivative has the same structure. The complete decomposition over all query and key cells is stated as Lemma B.1 in Appendix B.1, where we denote its aggregate RoPE-specific turning component by $\delta a_{\mathrm{R},mn}^{(i)}$. In particular, $\delta a_{\mathrm{R},mn}^{(i)} = 0$ when $\mathbf{R}(h\theta_i) = \mathbf{I}$, although the total score derivative need not vanish.

Crucially, the same operator $\mathbf{R}(h\theta_i) - \mathbf{I}$ used in Sec. 2 now appears in the derivative of a real attention score, with cell magnitudes and query-key interactions retained. This connects the normalized positional measure to actual attention computation and provides the starting point for tracing its frequency dependence into the loss. We quantify how this contribution varies across token distances.

3.2 How RoPE Weights Changes Across Token Distances

Sec. 3.1 isolates the RoPE-specific change of one score. We now keep pair i fixed and ask how these changes depend on the relative distances between tokens. Collect its query- and key-cell turns into

$$\boldsymbol{\phi}_i = \big(\phi_{Q,i,1}, \ldots, \phi_{Q,i,d_{\mathrm{model}}}, \phi_{K,i,1}, \ldots, \phi_{K,i,d_{\mathrm{model}}}\big)^\top \in \mathbb{R}^{2d_{\mathrm{model}}}. \quad (8)$$

These $2d_{\mathrm{model}}$ entries are turning angles, not parameter values, and need not be equal. Let $\mathcal{I}$ contain the head's unmasked interactions and $n_{\mathrm{att}} = |\mathcal{I}| > 0$. At fixed inputs and parameters, Eq. (B.2) is linear in these angles. Its coefficients define $\mathbf{J}_{\mathrm{R},i} \in \mathbb{R}^{n_{\mathrm{att}} \times 2d_{\mathrm{model}}}$ through

$$\big[\mathbf{J}_{\mathrm{R},i}\boldsymbol{\phi}_i\big]_{(m,n)} = \delta a_{\mathrm{R},mn}^{(i)}, \qquad (m,n) \in \mathcal{I}. \quad (9)$$

Thus, for $\mathbf{J}_{\mathrm{R},i}$, each row corresponds to one attention interaction (m,n), and each column to one cell turn. The product $\mathbf{J}_{\mathrm{R},i}\boldsymbol{\phi}_i$ gives the RoPE-specific first-order score changes.

For a fixed displacement h, define $\mathcal{I}_h = \{(m,n) \in \mathcal{I} : n - m = h\}$ and $n_h = |\mathcal{I}_h|$. The submatrix $\mathbf{J}_{\mathrm{R},i,h}$ contains all rows of $\mathbf{J}_{\mathrm{R},i}$ corresponding to these interactions. Its n_h rows represent interactions at distance h, while its $2d_{\mathrm{model}}$ columns represent query- and key-cell turns. Thus, $\mathbf{J}_{\mathrm{R},i,h}\boldsymbol{\phi}_i$ gives the RoPE-specific first-order score changes for this distance group. Consider the coefficient multiplying the query-cell turn $\phi_{Q,i,j}$ in interaction (m,n). From Eq. (B.2),

$$[\mathbf{J}_{\mathrm{R},i,h}]_{(m,n),(Q,j)} = x_{m,j}(\mathbf{J}\mathbf{w}_{Q,i,j})^\top \big(\mathbf{R}(h\theta_i) - \mathbf{I}\big)\mathbf{k}_n^{(i)} \overset{①}{=} 2\sin(h\theta_i/2)[\mathbf{G}_{i,h}]_{(m,n),(Q,j)}$$

where $[\mathbf{G}_{i,h}]_{(m,n),(Q,j)} = x_{m,j}(\mathbf{J}\mathbf{w}_{Q,i,j})^\top \mathbf{R}(h\theta_i/2)\mathbf{J}\mathbf{k}_n^{(i)}$. Here equality ① uses $\mathbf{R}(\alpha) - \mathbf{I} = 2\sin(\alpha/2)\mathbf{R}(\alpha/2)\mathbf{J}$. This separates two components: a shared rotary factor $2\sin(h\theta_i/2)$, determined only by i and h, and the remaining coefficient defined above. The factor's magnitude gives the length scaling of the rotation difference; the remaining coefficient retains the input, cell, key, and length-preserving rotation $\mathbf{R}(h\theta_i/2)\mathbf{J}$.

Applying the same identity to every query- and key-cell coefficient defines $\mathbf{G}_{i,h}$ from the remaining expressions, with the same row and column ordering as $\mathbf{J}_{\mathrm{R},i,h}$. These entries are defined directly, not by division, so they remain well defined when the scalar factor is zero. To average across distances, set $\pi(h) = n_h/n_{\mathrm{att}}$. This weights each interaction equally, giving more frequent distances more weight. We also use π to evaluate S_i in Sec. 2; these are distance weights, not attention probabilities.

Lemma 3.1 (Rotary weighting of attention-score responses)**.** *With the above construction, every nonempty distance group and their aggregating over distances respectively satisfy*

$$\mathbf{J}_{\mathrm{R},i,h} \overset{①}{=} 2\sin(h\theta_i/2)\,\mathbf{G}_{i,h}, \qquad \mathbf{J}_{\mathrm{R},i}^\top \mathbf{J}_{\mathrm{R},i}/n_{\mathrm{att}} \overset{②}{=} \mathbb{E}_{h\sim\pi}\left[4\sin^2(h\theta_i/2)(\mathbf{G}_{i,h}^\top \mathbf{G}_{i,h}/n_h)\right]. \tag{10}$$

The proof is in Appendix B.2. Lemma B.1 isolates RoPE's contribution to a score's turning response, while Lemma 3.1 shows how frequency and distance shape it. In Eq. ①, all response coefficients at distance h share the rotary factor $2\sin(h\theta_i/2)$, while $\mathbf{G}_{i,h}$ retains the inputs, parameters, and remaining rotation. Eq. ② averages the corresponding squared responses across distances with weights $4\sin^2(h\theta_i/2)$, whose mean is S_i. Thus, the rotary weighting identified in Sec. 2 also appears in actual attention derivatives. Since $\mathbf{G}_{i,h}$ still depends on the model, inputs, and frequency, S_i alone does not determine score sensitivity. We next examine how the loss weights these score changes.

3.3 From Attention-Score Changes to Loss Curvature

Sec. 3.2 shows how RoPE weights score changes. To connect this result to training, we must also account for the downstream loss: equal-sized score changes need not affect it equally. Let $\mathbf{a}$ collect the total unscaled scores $a_{mn} = \sum_i a_{mn}^{(i)}$, ordered as the rows of $\mathbf{J}_{\mathrm{R},i}$, where $a_{mn}^{(i)}$ is given in Eq. (7). Let $\mathbf{y}(\mathbf{a})$ be the output logits with all other inputs fixed, and let $\ell(\mathbf{y})$ be a twice continuously differentiable convex output loss, such as cross-entropy. Its local weighting of score changes is

$$\mathbf{B}_a := (\partial\mathbf{y}/\partial\mathbf{a})^\top \nabla_{\mathbf{y}}^2 \ell\,(\partial\mathbf{y}/\partial\mathbf{a}) \succeq 0. \tag{11}$$

The derivative maps score changes to logit changes, and the middle factor weights their quadratic effect on the output loss. Thus, $\frac{1}{2}\delta\mathbf{a}^\top \mathbf{B}_a \delta\mathbf{a}$ measures the quadratic cost of a score change $\delta\mathbf{a}$ in the linearized score-to-output model, where $\delta\mathbf{a}$ is given in Eq. (B.2). This generalized Gauss–Newton (GGN) construction retains output-loss curvature, not the full nonlinear task Hessian (Schraudolph, 2002; Martens, 2020). The usual $1/\sqrt{d_h}$ attention scaling is included in the downstream map.

Recall that for the turns $\boldsymbol{\phi}_i$ in Eq. (8), $\mathbf{J}_{\mathrm{R},i}\boldsymbol{\phi}_i$ is the isolated RoPE-specific score change as shown in Sec. 3.2. Applying the loss weighting gives

$$\mathbf{H}_{\mathrm{R},i} := \mathbf{J}_{\mathrm{R},i}^\top \mathbf{B}_a \mathbf{J}_{\mathrm{R},i}. \tag{12}$$

The quantity $\boldsymbol{\phi}_i^\top \mathbf{H}_{\mathrm{R},i}\boldsymbol{\phi}_i$ measures the squared size of the RoPE-specific score changes after weighting them by the local loss curvature. We bound it in two stages: Γ_i bounds the average squared response of $\mathbf{G}_{i,h}$ to a unit-length turn vector within each distance group, after the explicit rotary factor has been separated; β bounds the subsequent loss weighting by $\mathbf{B}_a$. These constants are used only in the analysis and need not be estimated to compute the parameter update.

Proposition 3.2 (Loss-weighted RoPE response and its sharpness)**.** *Suppose* $0 \preceq \mathbf{B}_a \preceq \beta\mathbf{I}$ *and* $\mathbf{G}_{i,h}^\top \mathbf{G}_{i,h}/n_h \preceq \Gamma_i\mathbf{I}$ *for every nonempty distance group, where* $\beta, \Gamma_i \geq 0$*. The matrix inequalities hold in every quadratic-form direction. Then every turn vector satisfies*

$$0 \leq \boldsymbol{\phi}_i^\top \mathbf{H}_{\mathrm{R},i}\boldsymbol{\phi}_i \leq n_{\mathrm{att}}\beta\Gamma_i S_i \|\boldsymbol{\phi}_i\|_2^2. \tag{13}$$

Moreover, the linear dependence on S_i is order-sharp: a family of RoPE attention models with cross-entropy loss matches this bound up to constant factors as $S_i \to 0^+$.

See the upper-bound proof in Appendix B.3.1. Proposition 3.2 connects the rotary structure in Lemma 3.1 to the training objective. The quantity $\frac{1}{2}\boldsymbol{\phi}_i^\top \mathbf{H}_{\mathrm{R},i}\boldsymbol{\phi}_i$ measures the quadratic cost of

Algorithm 1 Position-Aware Preconditioning (ideal nonzero-cell form)

Require: Parameters $\mathbf{W}_0$, base optimizer $\mathcal{A}$, T updates, precomputed multipliers $\{p_i\} \subset (0, 1]$ (default: Eq. (17); all cells nonzero, relative scale $s > -1$)

for $t = 0, \dots, T-1$ **do**
  Form $\boldsymbol{\Delta}_t^{\text{base}}$ before decoupled decay; retain the pre-update parameters
  for each RoPE-aligned cell $\mathbf{w}_{p,i,j}$ in $\mathbf{W}_Q, \mathbf{W}_K$ **do**
    $\mathbf{w} \leftarrow \mathbf{w}_{p,i,j}$, $r \leftarrow \|\mathbf{w}\|_2$, $\mathbf{e}_r \leftarrow \mathbf{w}/r$, $\mathbf{e}_\phi \leftarrow \mathbf{J}\mathbf{e}_r$
    $s \leftarrow \langle \boldsymbol{\Delta}_{p,i,j}^{\text{base}}, \mathbf{e}_r \rangle / r$, $\phi \leftarrow p_i \langle \boldsymbol{\Delta}_{p,i,j}^{\text{base}}, \mathbf{e}_\phi \rangle / r$
    $\widehat{\mathbf{w}}_{p,i,j} \leftarrow (1+s)\mathbf{R}(\phi)\mathbf{w}$
  end for
  Apply the base increments unchanged to all other parameters
  Apply the separate decay convention, when used, and commit $\mathbf{W}_{t+1}$ (Appendix C.2)
end for

the isolated RoPE-specific response under the local GGN model. The bound separates the turn magnitude $\|\boldsymbol{\phi}_i\|_2^2$ from the known rotary factor S_i and the model- and data-dependent factors β, Γ_i, showing that the same frequency structure also enters the loss-weighted response.

The full Q/K-block GGN additionally accounts for interactions among scale and turn changes, which may reinforce or cancel one another. Proposition B.4 retains S_i-weighted turning terms under its stated boundedness assumptions, while the construction in Appendix B.3.2 attains the same S_i order along a full Q/K-block direction. This motivates calibrating turns by S_i rather than treating them uniformly in Sec. 4.

4 Position-Aware Preconditioning

Secs. 2 and 3 identify a computable rotary factor in positional responses and in a loss-weighted response bound. We now use this information to choose updates: preserve the base optimizer's scale change and calibrate its turn. We first derive the correction from a local quadratic model, then examine when it relaxes a step-size restriction and establish convergence of the finite update.

4.1 From a Local Quadratic Model to the PA Update

Sec. 2 shows that equal turns can produce unequal positional changes, while positive rescaling leaves the normalized response unchanged. Sec. 3 traces the same rotary weighting into loss-weighted attention responses. PA therefore uses different turning multipliers across pairs to compensate for their unequal positional amplification, rather than uniformly shrinking every update.

Let $f(\mathbf{W})$ be the training objective. For a nonzero cell $\mathbf{w}$ of pair i, recall $r = \|\mathbf{w}\|_2$, $\mathbf{e}_r = \mathbf{w}/r$, and $\mathbf{e}_\phi = \mathbf{J}\mathbf{e}_r$ as its local scale and turn directions. Let $\mathbf{g} = \nabla_{\mathbf{w}} f(\mathbf{W})$ be its gradient and $\boldsymbol{\xi} \in \mathbb{R}^2$ a candidate parameter increment. To balance predicted loss reduction against step size, use

$$m_i(\boldsymbol{\xi}) = \mathbf{g}^\top \boldsymbol{\xi} + \frac{1}{2\eta}\|\boldsymbol{\xi}\|_2^2 + \frac{1}{2\eta}\kappa S_i \langle \boldsymbol{\xi}, \mathbf{e}_\phi \rangle^2, \qquad \eta > 0, \quad \kappa \geq 0. \tag{14}$$

The first term predicts the loss change, while the second controls the step and alone recovers ordinary gradient descent. The last term increases the relative turning cost from 1 to $1+\kappa S_i$ without affecting scale, so more sensitive pairs are corrected more strongly. This is a chosen quadratic step model, not the exact task Hessian or an added training loss. Here $\langle \boldsymbol{\xi}, \mathbf{e}_\phi \rangle = r\phi$ is a turning displacement, not an angle. Appendix A.3 gives the equivalent relative-coordinate form, where the positional cost becomes $\kappa S_i \phi^2/2$. Minimizing Eq. (14) gives

$$\boldsymbol{\xi}^\star = -\eta \mathbf{P}_i(\mathbf{w})\mathbf{g}, \qquad \mathbf{P}_i(\mathbf{w}) = \mathbf{e}_r\mathbf{e}_r^\top + p_i \mathbf{e}_\phi \mathbf{e}_\phi^\top, \quad \text{where} \quad p_i = p_i^{\text{ana}} := (1+\kappa S_i)^{-1}. \tag{15}$$

Thus, scale retains step coefficient η, while turn uses ηp_i^{ana}. The baseline keeps p_i^{ana} near one when κS_i is small, avoiding excessive correction of weakly sensitive directions. As the local basis follows the current cell, PA calibrates its turn rather than scaling the pair uniformly. Section 4.2 quantifies how this correction balances effective positional learning rates.

Applying PA to the base proposal. In practice, the base optimizer first forms its signed Cartesian increment $\mathbf{\Delta}^{\text{base}}$, including learning-rate scaling but excluding decoupled weight decay. PA then applies $\boldsymbol{\xi}^{\text{PA}} = \mathbf{P}_i \mathbf{\Delta}^{\text{base}}$. For the analytic multiplier p_i, this is the unique minimal-change correction with an S_i-weighted turning penalty (Appendix A.3), even when the proposal is not a raw gradient. So PA preserves how the base optimizer constructs its proposal and its local scale component, while modifying the turn. The same operator form supports any selected multiplier $0 < p_i \leq 1$ below.

Realizing the finite update. Extract the proposal's relative scale s_A and turn coordinates ϕ_A below. Then the implementation preserves s_A and applies the corrected turn directly to update $\mathbf{w}$:

$$s_A = \langle \mathbf{\Delta}^{\text{base}}, \mathbf{e}_r \rangle / r, \quad \phi_A = \langle \mathbf{\Delta}^{\text{base}}, \mathbf{e}_\phi \rangle / r, \quad \widehat{\mathbf{w}} = (1 + s_A)\mathbf{R}(p_i \phi_A)\mathbf{w}. \tag{16}$$

For $s_A > -1$, the cell's length changes by $1 + s_A$ and its direction by $p_i \phi_A$. For small relative steps, $\widehat{\mathbf{w}} = \mathbf{w} + \mathbf{P}_i \mathbf{\Delta}^{\text{base}} + \mathcal{O}\big(\|\mathbf{\Delta}^{\text{base}}\|_2^2 / r\big)$. The tangent correction is exact; the finite realization differs from Cartesian addition at second order. Consequently, $p_i = 1$ recovers the additive base update only to first order. The convergence proof retains this finite-map difference.

Choosing the practical multipliers. For small relevant rotary angles, $S_i \approx \mathbb{E}_\pi[h^2]\theta_i^2$. Absorbing the common distance factor into the correction strength gives the practical profile

$$p_i^{\text{LF}} = 1/(1 + \kappa \theta_i^2). \tag{17}$$

The rescaled strength is again denoted by κ, but need not have the same numerical value as in the analytic rule. Using this profile across the full frequency grid is a practical surrogate, not a uniform approximation to S_i. The exact response-factor calibration above concerns analytic PA; Section 4.2 assesses other profiles through their actual multipliers. Both profiles are precomputed. Appendix C gives the expansion, strength conversion, and separate PA-Log variant.

Putting the steps together. Algorithm 1 summarizes the update. Each cached p_i is shared by all cells of RoPE pair i; parameters outside the RoPE-aligned query/key cells retain their base increments, with decoupled decay applied separately. Beyond the cached profile, PA adds only linear arithmetic in the Q/K parameters, with no extra forward/backward pass or per-parameter optimizer state. Appendix C.2 details the pairing, radius safeguards, and decay convention.

4.2 Balancing Effective Positional Learning Rates

Sec. 4.1 specifies how PA corrects a turn. We now ask whether this correction makes the effective positional learning rates more balanced across RoPE pairs. To isolate the rotary effect, compare local response-gradient updates with the same turn-coordinate learning rate $\eta_\phi > 0$. In the linearized positional coordinate $z_i = \sqrt{S_i}\phi_i$ from Section 2, the coefficient multiplying the local loss gradient is the *effective positional learning rate*. Without correction and with analytic PA, respectively, it is

$$\eta_{i,\text{base}}^{\text{pos}} = \eta_\phi S_i, \qquad \eta_{i,\text{PA}}^{\text{pos}} = \eta_\phi p_i^{\text{ana}} S_i = \eta_\phi S_i / (1 + \kappa S_i). \tag{18}$$

Thus PA uses unequal turn multipliers to reduce the unequal positional rates created by RoPE. Appendix A.2 gives the chain-rule derivation.

Proposition 4.1 (More balanced effective positional learning rates)**.** *For fixed sensitivities, a common $\eta_\phi > 0$, and analytic PA with $\kappa \geq 0$, any two pairs with $S_i \geq S_j > 0$ satisfy*

$$1 \leq \eta_{i,\text{PA}}^{\text{pos}} / \eta_{j,\text{PA}}^{\text{pos}} = S_i / S_j \cdot (1 + \kappa S_j)/(1 + \kappa S_i) \leq S_i / S_j = \eta_{i,\text{base}}^{\text{pos}} / \eta_{j,\text{base}}^{\text{pos}}. \tag{19}$$

The upper inequality is strict when $\kappa > 0$ and $S_i > S_j$.

More sensitive pairs receive stronger attenuation, bringing their rate ratios closer to one without reversing their order. For pairs with $\kappa S_i \gg 1$, the rates approach the common scale η_ϕ / κ. PA therefore reduces the RoPE-induced imbalance, rather than making every rate exactly equal at finite strength. Stronger correction also reduces the absolute rates, so greater balance alone does not imply faster training. Appendix A.3.3 proves the result and quantifies when the rates are nearly equal.

Connection to a shared training step. Under Assumption A.4, the local finite-update loss bound has background curvature $L_0 > 0$ and an additional turning coefficient ρS_i, with $\rho \geq 0$. Choosing $\kappa = \rho / L_0$ gives

$$p_i^{\text{ana}}(L_0 + \rho S_i) = L_0 \qquad \text{for every pair } i. \tag{20}$$

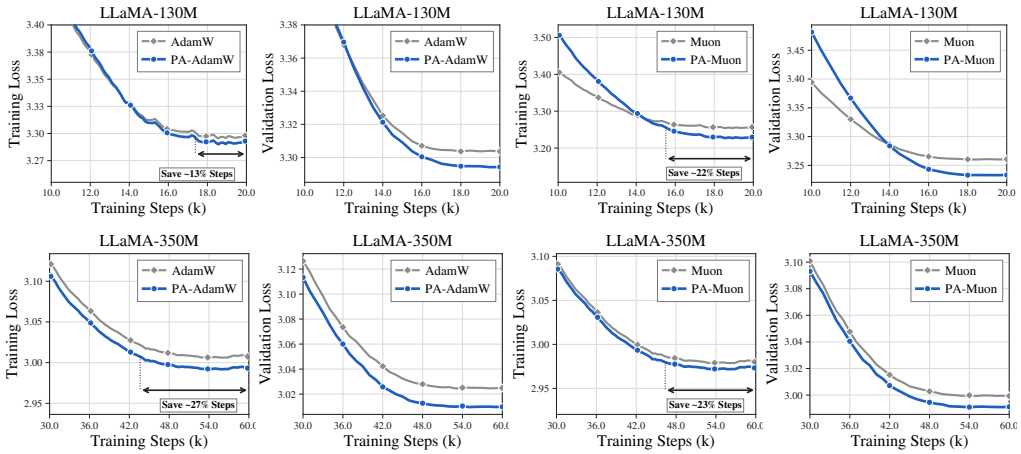


Figure 1: Training and evaluation curves for 130M and 350M LLaMA-style models optimized with AdamW and Muon using the standard PA correction. From left to right, the columns show AdamW training, AdamW evaluation, Muon training, and Muon evaluation; the top and bottom rows correspond to 130M and 350M models.

All turning directions then have the same coefficient in the preconditioned upper bound. For gradient updates without decoupled decay, Proposition A.5 relaxes the sufficient global step-size bound from $1/(L_0 + \rho S_{\max})$ to $1/L_0$, where $S_{\max} := \max_i S_i$, relative to $p_i = 1$ with the same finite map. This removes the additional rotary restriction in that model, provided the step stays within the assumed neighborhood; it does not assert equal true curvatures or guaranteed acceleration.

The rate comparison holds at matched η_ϕ: Cartesian gradient descent has $\eta_\phi = \eta/r^2$, so cell norms and base-optimizer scaling can introduce other differences. The practical frequency profile in Eq. (17) approximates analytic PA for small relevant rotary angles with the corresponding strength conversion, but need not satisfy the same ratio guarantee over the full frequency grid (Appendix C).

4.3 Convergence Guarantees

The conditional result explains a possible benefit. We separately establish convergence of the ideal finite update with $\mathbf{\Delta}_t^{\text{base}} = -\eta \nabla f(\mathbf{W}_t)$. See its proof in Appendix D.

Theorem 4.2 (First-order convergence of PA)**.** *Let f be continuously differentiable, bounded below by $f_{\inf}$, and have an L-Lipschitz gradient with $L > 0$. Fix $p_i \in [p_{\min}, 1]$ with $p_{\min} > 0$, and assume every visited RoPE-aligned cell has norm at least $r_{\min} > 0$. An explicit threshold $\eta_0 > 0$ ensures that, for $0 < \eta \le \eta_0$ and every $T \ge 1$,*

$$\min_{0 \le t < T} \|\nabla f(\mathbf{W}_t)\|_2^2 \le 2[f(\mathbf{W}_0) - f_{\inf}]/(\eta p_{\min} T). \tag{21}$$

Theorem 4.2 shows that PA preserves the standard $O(1/T)$ first-order stationarity rate. This is a safety guarantee rather than a claim of faster training; any speed benefit is instead tied to the RoPE-dependent conditioning identified in Secs. 2–3 and evaluated in Sec. 5. Appendix D extends the analysis to unbiased stochastic gradients and, under explicit alignment and second-moment conditions, more general optimizer directions, with the latter remaining conditional for AdamW or Muon.

5 Experiments

We evaluate whether PA accelerates pretraining with low additional cost on 130M- and 350M-parameter LLaMA-style models (Touvron et al., 2023) and Qwen2.5-0.5B (Team, 2024) pretrained on C4 (Raffel et al., 2020). We apply PA to AdamW and Muon using the θ_i^2 profile in Algorithm 1 and the PA-Log variant. Full configurations, hyperparameters, training schedules, and evaluation protocols are given in Appendix E.

5.1 Pretraining Performance and Training Cost

Faster pretraining progress. Figure 1 compares the standard PA update in Algorithm 1 with its corresponding base optimizer on 130M- and 350M-parameter LLaMA-style models. Across AdamW

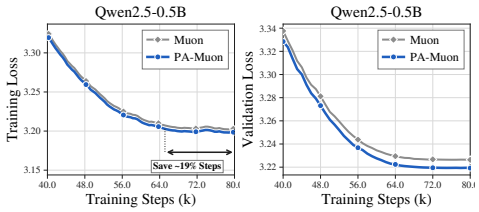


Figure 2: Training and validation curves for Qwen2.5-0.5B optimized with Muon and PA-Muon using the standard PA.

Table 1: Per-step cost and estimated wall-clock saving to the same training loss.

Model	Optimizer	Step time (s)		Peak mem. (GiB)	Time saving
		Base	PA	Base / PA	to same loss
LLaMA-130M	AdamW	0.760	0.806	30.83 / 30.83	7.7%
LLaMA-130M	Muon	0.820	0.840	30.83 / 30.83	20.1%
LLaMA-350M	AdamW	0.629	0.657	32.65 / 32.65	23.8%
LLaMA-350M	Muon	0.663	0.668	31.66 / 31.66	22.4%
Qwen2.5-0.5B	Muon	0.751	0.763	59.82 / 59.82	17.7%

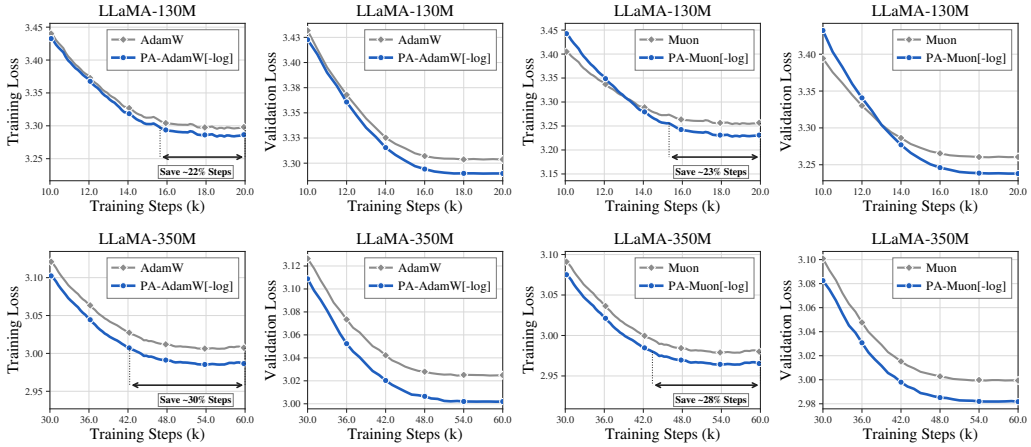


Figure 3: Training and evaluation curves for 130M and 350M LLaMA-style models optimized with AdamW and Muon using PA-Log. From left to right, the columns show AdamW training, AdamW evaluation, Muon training, and Muon evaluation.

and Muon, PA reaches the corresponding baseline training-loss level with approximately 13%–27% fewer optimizer steps, while the validation curves show the same overall improvement. We further evaluate PA on Qwen2.5-0.5B in Figure 2, where PA-Muon reaches the corresponding Muon training loss with approximately 19% fewer steps. Together, these results show consistent faster pretraining progress across model scales, model families, and substantially different base optimizers.

Training cost. The faster convergence comes with modest additional per-step computation and no measurable increase in peak allocated GPU memory. As shown in Table 1, PA increases per-step runtime by approximately 0.8%–6.1% across the measured settings. Combining these measured per-step runtimes with the step reductions observed in the corresponding training curves gives an estimated wall-clock saving of 7.7%–23.8% to reach the same training loss.

5.2 A Log-Smoothed PA Variant

The standard PA multiplier follows from the low-frequency approximation and uses the θ_i^2 frequency profile. Since RoPE frequencies follow a geometric schedule, this profile can vary sharply across rotary pairs. We therefore also consider a logarithmically smoothed variant, PA-Log, which preserves the same scale-turn correction while compressing the dynamic range through a normalized log-frequency coordinate; the exact profile is given in Appendix C.1. Figure 3 reports the corresponding results for PA-AdamW-Log and PA-Muon-Log. Across both model scales and base optimizers, PA-Log reaches the corresponding baseline training-loss level with approximately 22%-30% fewer optimizer steps and generally improves over the standard θ_i^2 profile, suggesting additional benefit.

6 Conclusion

We identify a RoPE-induced positional-step mismatch and show that the same frequency-dependent structure enters the local training-loss geometry. Position-Aware Preconditioning preserves the base optimizer's scale update while recalibrating its turn. Across LLaMA-style and Qwen pretraining, PA reaches the reported loss targets sooner with modest overhead and no measurable cost increase.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity, grammar, and presentation of the manuscript. All AI-assisted content was carefully reviewed and verified by the authors. All scientific ideas, methodological designs, formulation, theoretical analysis, experimental results, and conclusions were developed and validated by the authors. We take full responsibility for the final content of this work.

REFERENCES

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 7432–7439, 2020.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Kevin Frans, Pieter Abbeel, and Sergey Levine. What really matters in matrix-whitening optimizers? *arXiv preprint arXiv:2510.25000*, 2025.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pp. 552–561, 2012.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International conference on learning representations*, volume 2024, pp. 1621–1650, 2024.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025a.

Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. Cosmos: A hybrid adaptive optimizer for memory-efficient training of llms. *arXiv preprint arXiv:2502.17410*, 2025b.

Yifeng Liu, Angela Yuan, and Quanquan Gu. Mars-m: When variance reduction meets matrices. *arXiv preprint arXiv:2510.21800*, 2025c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Chao Ma, Wenbo Gong, Meyer Scetbon, and Edward Meeds. Swan: Preprocessing sgd enables adam-level performance on llm training with significant memory reduction. *arXiv preprint arXiv:2412.13148*, 2024.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

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. URL `https://jmlr.org/papers/v21/20-074.html`.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Nicol N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002. doi: 10.1162/08997660260028683.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Huizhuo Yuan, Yifeng Liu, Shuang Wu, Xun Zhou, and Quanquan Gu. Mars: Unleashing the power of variance reduction for training large models. *arXiv preprint arXiv:2411.10438*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for language models. *arXiv preprint arXiv:2407.07972*, 2024.

A Positional Response and Local Optimization Models

A.1 Exact response change and the meaning of the coordinates

Before analyzing the positional response induced by a RoPE-aligned parameter cell, we first justify the local scale–turn parameterization used throughout the paper. The purpose of this construction is to connect an arbitrary Cartesian perturbation of a two-dimensional parameter cell to two geometrically meaningful local coordinates: a radial change in magnitude and an angular turn in direction. We then realize these local coordinates through the finite map $\mathbf{w}(s,\phi) = (1+s)\mathbf{R}(\phi)\mathbf{w}$. The following proposition shows that every infinitesimal cell perturbation admits a unique radial–angular decomposition and that the finite scale–turn map reproduces this perturbation to first order.

Proposition A.1 (Local scale–turn coordinates and finite realization)**.** *For any nonzero cell $\mathbf{w} \in \mathbb{R}^2$, let $r = \|\mathbf{w}\|_2$, $\mathbf{e}_r = \mathbf{w}/r$, and $\mathbf{e}_\phi = \mathbf{J}\mathbf{e}_r$, where $\mathbf{J} = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$. Then $\{\mathbf{e}_r, \mathbf{e}_\phi\}$ is an orthonormal basis of $\mathbb{R}^2$. Consequently, any perturbation $\boldsymbol{\xi} \in \mathbb{R}^2$ admits the unique representation $\boldsymbol{\xi} = r(s\mathbf{e}_r + \phi\mathbf{e}_\phi)$, where $s = \langle \boldsymbol{\xi}, \mathbf{e}_r\rangle/r$ and $\phi = \langle \boldsymbol{\xi}, \mathbf{e}_\phi\rangle/r$.*

Moreover, the finite scale–turn map $\mathbf{w}(s,\phi) = (1+s)\mathbf{R}(\phi)\mathbf{w}$ satisfies, for sufficiently small s and ϕ,

$$\mathbf{w}(s,\phi) = \mathbf{w} + \boldsymbol{\xi} + \mathcal{O}\big(r\big(|s\phi| + \phi^2\big)\big)\,. \tag{A.1}$$

Hence, its first-order increment coincides with $\boldsymbol{\xi}$.

Proof. Since $\mathbf{J}^\top\mathbf{J} = \mathbf{I}$ and $\mathbf{e}_r^\top\mathbf{J}\mathbf{e}_r = 0$, $\mathbf{e}_\phi = \mathbf{J}\mathbf{e}_r$ is a unit vector orthogonal to $\mathbf{e}_r$. Therefore, $\{\mathbf{e}_r, \mathbf{e}_\phi\}$ forms an orthonormal basis of $\mathbb{R}^2$. Every $\boldsymbol{\xi} \in \mathbb{R}^2$ therefore decomposes uniquely as $\boldsymbol{\xi} = \langle \boldsymbol{\xi}, \mathbf{e}_r\rangle\mathbf{e}_r + \langle \boldsymbol{\xi}, \mathbf{e}_\phi\rangle\mathbf{e}_\phi = r(s\mathbf{e}_r + \phi\mathbf{e}_\phi)$.

Expanding $\mathbf{R}(\phi) = \mathbf{I} + \phi\mathbf{J} + \mathcal{O}(\phi^2)$ in $(1+s)\mathbf{R}(\phi)\mathbf{w}$ gives linear terms $sr\mathbf{e}_r + \phi r\mathbf{e}_\phi = \boldsymbol{\xi}$ and remainder $\mathcal{O}(r(|s\phi| + \phi^2))$ for bounded small s, ϕ. This proves Eq. (A.1). □

Equation (A.1) shows that the scale–turn map is a finite realization of the local tangent perturbation rather than an additional modeling assumption. The remainder term arises only from the nonlinear finite rotation and therefore does not omit any first-order component of the original Cartesian perturbation.

We next prove Proposition 2.1. By Proposition A.1, s and ϕ represent the local radial and angular components of a cell perturbation. Because the positional response is normalized by the cell magnitude, the positive radial factor $1+s$ cancels exactly. Hence, the proof reduces to quantifying how the angular turn ϕ interacts with the rotary transformation $\mathbf{R}(h\theta_i)$.

Proof of Proposition 2.1. For $s > -1$, normalization removes the positive scale:

$$\frac{(1+s)\mathbf{R}(\phi)\mathbf{w}}{\|(1+s)\mathbf{R}(\phi)\mathbf{w}\|_2} = \mathbf{R}(\phi)\mathbf{e}_r.$$

Therefore the response difference is

$$\boldsymbol{\psi}_{i,h}(\mathbf{w}(s,\phi)) - \boldsymbol{\psi}_{i,h}(\mathbf{w}) = (\mathbf{R}(h\theta_i) - \mathbf{I})(\mathbf{R}(\phi) - \mathbf{I})\mathbf{e}_r.$$

For every real α,

$$(\mathbf{R}(\alpha) - \mathbf{I})^\top(\mathbf{R}(\alpha) - \mathbf{I}) = 4\sin^2(\alpha/2)\mathbf{I}. \tag{A.2}$$

Applying this identity twice and using $\|\mathbf{e}_r\|_2 = 1$ gives

$$\|\boldsymbol{\psi}_{i,h}(\mathbf{w}(s,\phi)) - \boldsymbol{\psi}_{i,h}(\mathbf{w})\|_2^2 = 16\sin^2(h\theta_i/2)\sin^2(\phi/2).$$

Taking the expectation over h gives $D_i(\phi) = 4S_i\sin^2(\phi/2)$, the exact identity in Eq. (6). Since $4\sin^2(\phi/2) = \phi^2 + \mathcal{O}(\phi^4)$, its small-turn expansion is $S_i\phi^2 + \mathcal{O}(S_i\phi^4)$. In particular, a pure positive rescaling leaves the normalized response unchanged. A negative rescaling is not covered: it reverses the direction, which is why $s > -1$ is required. □

The result is about a *vector-valued normalized response*. It does not say that the response norm changes with orientation, nor that every attention score changes when two cells turn together. The response vector can rotate while its norm stays fixed; query/key interactions can also cancel such turns.

A.2 From a response change to an effective positional learning rate

This subsection defines the effective positional learning rate used in Section 4.2. It concerns the local response model, not the entire attention-to-loss map. Fix a nonzero cell $\mathbf{w}$ of pair i and the displacement distribution π. Stack the normalized responses as

$$\mathbf{\Psi}_i(\phi) := \mathrm{col}_h\big(\sqrt{\pi(h)}\,\boldsymbol{\psi}_{i,h}(\mathbf{R}(\phi)\mathbf{w})\big), \qquad \mathbf{b}_i := \left.\frac{\partial \mathbf{\Psi}_i}{\partial \phi}\right|_{\phi=0}.$$

Here $\mathbf{b}_i$ is the first-order change of the stacked response per unit turn, and $\phi = 0$ denotes no additional turn at the current cell.

Lemma A.2 (Local response Jacobian scale)**.** *The response derivative satisfies*

$$\|\mathbf{b}_i\|_2^2 = S_i. \tag{A.3}$$

Proof. The distance-h block of $\mathbf{b}_i$ is $\sqrt{\pi(h)}(\mathbf{R}(h\theta_i) - \mathbf{I})\mathbf{J}\mathbf{e}_r$. Since $\|\mathbf{J}\mathbf{e}_r\|_2 = 1$, the rotation identity in Eq. (A.2) gives its squared norm as $\pi(h)4\sin^2(h\theta_i/2)$. Summing over h yields S_i. □

For $S_i > 0$, let $\mathbf{t}_i = \mathbf{b}_i/\sqrt{S_i}$ be the unit response direction reached by turning this cell. With $z_i = \sqrt{S_i}\phi_i$,

$$\mathbf{\Psi}_i(\phi_i) = \mathbf{\Psi}_i(0) + \mathbf{t}_i z_i + \mathcal{O}(\phi_i^2).$$

Thus z_i measures signed movement along the *linearized* response direction. For a differentiable loss ℓ on this response, define $\ell_i^{\mathrm{loc}}(z) := \ell(\mathbf{\Psi}_i(0) + \mathbf{t}_i z)$. No quadratic-loss assumption is needed for the following local update.

Corollary A.3 (Local effective turn-coordinate step)**.** *Fix $p_i \in (0, 1]$. A turn-coordinate gradient proposal with rate $\eta_\phi > 0$, followed by turn scaling by p_i, gives*

$$\Delta\phi_i = -\eta_\phi p_i \left.\frac{d\,\ell(\mathbf{\Psi}_i(\phi))}{d\phi}\right|_0, \qquad \Delta z_i = -\underbrace{\eta_\phi p_i S_i}_{\eta_i^{\mathrm{pos}}}(\ell_i^{\mathrm{loc}})'(0). \tag{A.4}$$

The second equality is exact for the linearized coordinate; the actual response change has an $\mathcal{O}((\Delta\phi_i)^2)$ remainder. For $S_i = 0$, the first-order normalized response is zero and we set $\eta_i^{\mathrm{pos}} = 0$ without defining $\mathbf{t}_i$.

Proof. At the current response, the chain rule gives

$$\left.\frac{d\,\ell(\mathbf{\Psi}_i(\phi))}{d\phi}\right|_0 = \mathbf{b}_i^\top \nabla_{\mathbf{\Psi}}\ell = \sqrt{S_i}\,(\ell_i^{\mathrm{loc}})'(0).$$

Multiply the turn update by $\sqrt{S_i}$ to obtain Eq. (A.4). Equivalently,

$$\Delta\mathbf{\Psi}_i = -\eta_\phi p_i \mathbf{b}_i\mathbf{b}_i^\top \nabla_{\mathbf{\Psi}}\ell + \mathcal{O}((\Delta\phi_i)^2).$$

For $S_i > 0$, $\mathbf{b}_i\mathbf{b}_i^\top = S_i\mathbf{t}_i\mathbf{t}_i^\top$, so its only nonzero eigenvalue is S_i; for $S_i = 0$, it is the zero operator. This proves both the effective-rate interpretation and the degenerate case. □

Cartesian rates and cell norms. For a scalar cell loss and a Cartesian gradient proposal $\Delta\mathbf{w} = -\eta\nabla_{\mathbf{w}}\ell$, the extracted turn is

$$\Delta\phi^{\mathrm{base}} = \frac{\langle \Delta\mathbf{w}, \mathbf{e}_\phi\rangle}{r} = -\frac{\eta}{r^2}\frac{\partial \ell}{\partial \phi}, \qquad \eta_\phi = \frac{\eta}{r^2}. \tag{A.5}$$

Indeed, $\partial\ell/\partial\phi = r\langle\nabla_{\mathbf{w}}\ell, \mathbf{e}_\phi\rangle$. PA multiplies this local turn by p_i. Consequently, for the normalized response model its effective rate is $(\eta/r^2)p_iS_i$. A shared Cartesian rate need not give a shared η_ϕ; different cell norms can introduce variation beyond the rotary factor. These coordinates refer to the local proposal, not the exact angle of an additive endpoint. A general adaptive proposal need not be a scalar multiple of the local loss gradient, so the corollary is not an unconditional effective-rate formula for AdamW or Muon on the full task loss.

The corresponding Cartesian response metric. Write the stacked normalized response as a function of the cell, $\mathbf{\Psi}_i(\mathbf{w}) := \mathrm{col}_h(\sqrt{\pi(h)}\,\boldsymbol{\psi}_{i,h}(\mathbf{w}))$, and let $\mathbf{A}_i = \partial\mathbf{\Psi}_i(\mathbf{w})/\partial\mathbf{w}$. Normalization removes the radial derivative, while a Cartesian increment $\boldsymbol{\xi}$ has local turn $\langle\boldsymbol{\xi}, \mathbf{e}_\phi\rangle/r$. Hence $\mathbf{A}_i = \mathbf{b}_i\mathbf{e}_\phi^\top/r$, and

$$\mathbf{A}_i^\top\mathbf{A}_i = \frac{S_i}{r^2}\mathbf{e}_\phi\mathbf{e}_\phi^\top, \qquad \|\mathbf{A}_i\boldsymbol{\xi}\|_2^2 = \frac{S_i}{r^2}\langle\boldsymbol{\xi}, \mathbf{e}_\phi\rangle^2. \tag{A.6}$$

The factor r^{-2} distinguishes equal Cartesian displacements from equal turns. The scale-balanced step cost in Appendix A.3 normalizes both fidelity and penalty by the same radius.

Relation to the illustrative quadratic loss. For the illustration in Sec. 2, let $z_i^\star = \sqrt{S_i}\phi_i^\star$ and consider $\ell_i(\phi_i) = \frac{a}{2}(z_i - z_i^\star)^2 = \frac{aS_i}{2}(\phi_i - \phi_i^\star)^2$ with $a > 0$. Thus, $\partial^2\ell_i/\partial z_i^2 = a$, while $\partial^2\ell_i/\partial\phi_i^2 = aS_i$. The local linear response model gives

$$z_i^{t+1} - z_i^\star = (1 - a\eta_i^{\mathrm{pos}})(z_i^t - z_i^\star).$$

Thus the effective rate multiplies curvature in the usual gradient-descent error recurrence. Without PA, contraction across the $S_i > 0$ directions requires $0 < \eta_\phi < 2/(aS_{\max})$, where $S_{\max} = \max_i S_i > 0$. With PA, replace S_i by p_iS_i. This is an illustration in the stated quadratic model, not a claim that the full Transformer objective has that form.

A.3 Equivalent local models and the finite PA correction

This subsection connects the quadratic step model in Section 4.1 to the original minimal-change correction. The different coordinates describe the same local operator; the finite endpoint is specified separately. Throughout, the nonzero cell $\mathbf{w}$, its radius r, the orthonormal basis $\mathbf{e}_r, \mathbf{e}_\phi$, and the sensitivity S_i are fixed.

A.3.1 Cartesian, normalized, and scale–turn formulations

For any signed base increment $\boldsymbol{\Delta}^{\mathrm{base}}$, consider

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^2}\left\{\tfrac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{\Delta}^{\mathrm{base}}\|_2^2 + \tfrac{\kappa}{2}S_i\langle\boldsymbol{\xi}, \mathbf{e}_\phi\rangle^2\right\}. \tag{A.7}$$

Its Hessian is $\mathbf{I} + \kappa S_i\mathbf{e}_\phi\mathbf{e}_\phi^\top \succ 0$. The first-order condition therefore gives the unique solution $\boldsymbol{\xi}^\star = \mathbf{P}_i\boldsymbol{\Delta}^{\mathrm{base}}$, where $\mathbf{P}_i$ has multiplier p_i^{ana} in Eq. (15).

To recover the normalized-vector form, write $\mathbf{d} = \boldsymbol{\xi}/r$ and $\mathbf{d}^{\mathrm{base}} = \boldsymbol{\Delta}^{\mathrm{base}}/r$. Dividing Eq. (A.7)'s objective by r^2 gives

$$\min_{\mathbf{d}}\left\{\tfrac{1}{2}\|\mathbf{d} - \mathbf{d}^{\mathrm{base}}\|_2^2 + \tfrac{\kappa}{2}S_i\|(\mathbf{I} - \mathbf{e}_r\mathbf{e}_r^\top)\mathbf{d}\|_2^2\right\}. \tag{A.8}$$

Next set $\mathbf{d} = s\mathbf{e}_r + \phi\mathbf{e}_\phi$ and $\mathbf{d}^{\mathrm{base}} = s_A\mathbf{e}_r + \phi_A\mathbf{e}_\phi$. Orthonormality turns this into

$$\min_{s,\phi}\left\{\tfrac{1}{2}(s - s_A)^2 + \tfrac{1}{2}(\phi - \phi_A)^2 + \tfrac{\kappa}{2}S_i\phi^2\right\}, \qquad s^\star = s_A, \quad \phi^\star = p_i^{\mathrm{ana}}\phi_A. \tag{A.9}$$

Thus, all three formulations preserve the scale coordinate and attenuate the turn by the same amount. In relative coordinates, $S_i\phi^2$ is the leading squared normalized-response change. In Cartesian coordinates this quantity is $S_i\langle\boldsymbol{\xi}, \mathbf{e}_\phi\rangle^2/r^2$; Eq. (A.7) removes the same factor r^{-2} from both fidelity and penalty terms. It is not a literal unnormalized response metric; compare Eq. (A.6).

For a raw gradient proposal $\boldsymbol{\Delta}^{\mathrm{base}} = -\eta\mathbf{g}$, completing the square gives

$$\eta m_i(\boldsymbol{\xi}) = \tfrac{1}{2}\|\boldsymbol{\xi} + \eta\mathbf{g}\|_2^2 + \tfrac{\kappa}{2}S_i\langle\boldsymbol{\xi}, \mathbf{e}_\phi\rangle^2 - \tfrac{\eta^2}{2}\|\mathbf{g}\|_2^2.$$

Hence Eq. (14) has exactly the same minimizer. For completeness, the derivatives at zero of $\widetilde{f}(s, \phi) = f((1 + s)\mathbf{R}(\phi)\mathbf{w})$, with other parameters fixed, are $g_s = r\langle\mathbf{g}, \mathbf{e}_r\rangle$ and $g_\phi = r\langle\mathbf{g}, \mathbf{e}_\phi\rangle$. With $\eta_c = \eta/r^2$, the same step model becomes

$$m_i\big(r(s\mathbf{e}_r + \phi\mathbf{e}_\phi)\big) = g_s s + g_\phi\phi + \frac{s^2 + \phi^2 + \kappa S_i\phi^2}{2\eta_c}. \tag{A.10}$$

A shared Cartesian rate η therefore need not give a shared coordinate rate η_c. This derivation explains the chosen quadratic update cost; it does not identify it with the task Hessian or prove a loss upper bound. For a general base proposal, Eq. (A.7) remains a minimal-change problem without assuming the proposal is a gradient.

A.3.2 THE LOCAL INCREMENT AND ITS FINITE REALIZATION

For any fixed $0 < p_i \leq 1$, the extracted coordinates satisfy $\mathbf{\Delta}^{\text{base}} = r(s_A\mathbf{e}_r + \phi_A\mathbf{e}_\phi)$. The corrected local increment is exactly $r(s_A\mathbf{e}_r + p_i\phi_A\mathbf{e}_\phi) = \mathbf{P}_i\mathbf{\Delta}^{\text{base}}$. Expanding the finite map in Eq. (16) yields

$$\widehat{\mathbf{w}} - \mathbf{w} - \mathbf{P}_i\mathbf{\Delta}^{\text{base}} = \mathcal{O}\big(r\big(|s_A p_i \phi_A| + p_i^2\phi_A^2\big)\big) = \mathcal{O}\left(\frac{\|\mathbf{\Delta}^{\text{base}}\|_2^2}{r}\right). \tag{A.11}$$

The estimate holds for small relative steps, uniformly over $0 < p_i \leq 1$. Thus $p_i = 1$ removes attenuation, but retains the second-order difference from Cartesian addition. For example, $\mathbf{w} = (1, 0)^\top$ and $\mathbf{\Delta}^{\text{base}} = (0, \epsilon)^\top$ give the endpoint $(\cos\epsilon, \sin\epsilon)^\top$, not $(1, \epsilon)^\top$. The unconstrained local problem chooses coordinates; their positive-scale interpretation requires $s_A > -1$. The gradient theorem ensures a stricter local scale margin, while implementation safeguards have a separate scope.

A.3.3 WHY ANALYTIC PA BALANCES EFFECTIVE POSITIONAL LEARNING RATES

We compare one cell from each pair at the same local turn-coordinate rate $\eta_\phi > 0$. This holds the non-RoPE scaling fixed and isolates the imbalance induced by S_i. Taking $p_i = 1$ or $p_i = p_i^{\text{ana}}$ in Corollary A.3 gives the two rates in Eq. (18). Their ratio comparison does not require Assumption A.4; that additional condition is used only for the finite-update loss claim below.

Proof of Proposition 4.1. For $S_i \geq S_j > 0$, canceling the common η_ϕ gives the middle identity in Eq. (19). The function $F_\kappa(S) = S/(1+\kappa S)$ is increasing on $S \geq 0$, since $F'_\kappa(S) = (1+\kappa S)^{-2} > 0$. Hence the corrected ratio is at least one. Also $(1 + \kappa S_j)/(1 + \kappa S_i) \leq 1$, with strict inequality when $\kappa > 0$ and $S_i > S_j$. This proves the upper comparison. For equal sensitivities the ratio is one; for $\kappa = 0$ PA leaves both rates unchanged. If $S_j = 0$, both its rates are zero, so no ratio with that denominator is asserted. □

When are the rates nearly equal? For $\kappa > 0$ and $S_i > 0$,

$$\eta_{i,\text{PA}}^{\text{pos}} = \frac{\eta_\phi}{\kappa}\left(1 - \frac{1}{1 + \kappa S_i}\right). \tag{A.12}$$

This gives a quantitative version of the common-scale interpretation. For any $\tau > 0$, consider a nonempty finite group of pairs with $\kappa S_i \geq \tau$. Their rates satisfy

$$\frac{\eta_\phi}{\kappa}\frac{\tau}{1+\tau} \leq \eta_{i,\text{PA}}^{\text{pos}} < \frac{\eta_\phi}{\kappa}, \qquad \frac{\max_i \eta_{i,\text{PA}}^{\text{pos}}}{\min_i \eta_{i,\text{PA}}^{\text{pos}}} \leq 1 + \frac{1}{\tau}, \tag{A.13}$$

where the extrema are over that group. The first bounds follow directly from Eq. (A.12); dividing them gives the ratio bound. For example, $\tau = 10$ bounds the rate ratio by 1.1. For any fixed positive sensitivities, pairwise rate ratios approach one as $\kappa \to \infty$, but every absolute rate tends to zero when η_ϕ is fixed. Rate balance alone therefore does not imply rapid progress.

Prescribed turns are a different comparison. The effective rate describes a gradient-driven local response. For a fixed base proposal with relative scale $s_A > -1$ and turn ϕ_A, the actual finite update instead has the exact normalized-response change

$$D_i(p_i\phi_A) = 4S_i \sin^2(p_i\phi_A/2) = S_i p_i^2\phi_A^2 + \mathcal{O}(S_i p_i^4 \phi_A^4). \tag{A.14}$$

This follows from Proposition 2.1. The squared movement of a prescribed turn contains p_i^2; the local response-gradient rate contains p_i because its coefficient already includes both forward and backward response derivatives. These are different quantities. The finite sine expression is periodic, so its attenuation interpretation is local, not a monotonicity claim for arbitrary large turns.

What the comparison does not assume. Equal η_ϕ is a comparison condition, not a change to the implemented optimizer. For a cell with radius r_c, Cartesian gradient descent gives $\eta_{\phi,c} = \eta/r_c^2$ by Eq. (A.5). With unequal base coordinate rates, the corrected ratio also contains $\eta_{\phi,i}/\eta_{\phi,j}$, so its distance from one need not shrink. Arbitrary adaptive proposals and the full attention interaction introduce further structure not described by the normalized-response calculation.

The ratio guarantee uses the analytic multiplier, not an arbitrary decreasing frequency profile. The practical θ_i^2 profile approximates it when $S_i \approx \mu_2\theta_i^2$, $\mu_2 = \mathbb{E}_\pi[h^2]$, and the strengths are matched by $\kappa_{\mathrm{LF}} = \mu_2\kappa_{\mathrm{ana}}$ (Appendix C). Outside this regime, or for the separate PA-Log profile, no exact ratio guarantee is asserted. The finite-update loss analysis below nevertheless applies to any fixed multipliers in $(0, 1]$ under its own explicit assumptions.

A.4 Conditional step-size improvement for the finite update

The response-rate comparison in Section 4.2 isolates the rotary factor. A claim about the actual training loss additionally needs a local loss bound. Here we state that condition and retain the finite update, rather than replacing it by its first-order additive approximation.

Let $\mathcal{C}$ index all PA cells, including their layer, head, and Q/K identifiers, and let $i(c)$ identify the sensitivity and multiplier used by cell c. A full Cartesian increment $\boldsymbol{\xi}$ contains cell increments $\boldsymbol{\xi}_c$ and a remaining block $\boldsymbol{\xi}_{\mathrm{rest}}$. At the current parameters $\mathbf{W}$, assume all PA cells have radii $r_c > 0$ and extract

$$s_c = \frac{\langle \boldsymbol{\xi}_c, \mathbf{e}_{r,c}\rangle}{r_c}, \qquad \phi_c = \frac{\langle \boldsymbol{\xi}_c, \mathbf{e}_{\phi,c}\rangle}{r_c}.$$

The finite realization of this increment is

$$\begin{aligned} \big[\mathrm{Retr}_{\mathbf{W}}(\boldsymbol{\xi})\big]_c &= (1 + s_c)\mathbf{R}(\phi_c)\mathbf{w}_c, \quad c \in \mathcal{C}, \quad s_c > -1, \\ \big[\mathrm{Retr}_{\mathbf{W}}(\boldsymbol{\xi})\big]_{\mathrm{rest}} &= \mathbf{W}_{\mathrm{rest}} + \boldsymbol{\xi}_{\mathrm{rest}}. \end{aligned} \tag{A.15}$$

For selected fixed $p_i \in (0, 1]$, let $\mathbf{P}_{\mathbf{W}}$ have cell blocks $\mathbf{e}_{r,c}\mathbf{e}_{r,c}^\top + p_{i(c)}\mathbf{e}_{\phi,c}\mathbf{e}_{\phi,c}^\top$ and the identity elsewhere. Then $\mathrm{Retr}_{\mathbf{W}}(-\eta\mathbf{P}_{\mathbf{W}}\nabla f(\mathbf{W}))$ is exactly the ideal gradient specialization of Algorithm 1, without decoupled decay or additional safeguards. The turn multiplier is applied through $\mathbf{P}_{\mathbf{W}}$ once; the finite map applies no second copy of it.

Assumption A.4 (Local finite-map quadratic upper bound)**.** At the current $\mathbf{W}$, there exist $L_0 > 0$, $\rho \geq 0$, and a specified neighborhood of zero such that every admissible $\boldsymbol{\xi}$ in that neighborhood satisfies

$$\begin{aligned} f(\mathrm{Retr}_{\mathbf{W}}(\boldsymbol{\xi})) \leq f(\mathbf{W}) + \langle \nabla f(\mathbf{W}), \boldsymbol{\xi}\rangle + \frac{L_0}{2}\|\boldsymbol{\xi}\|_2^2 \\ + \frac{\rho}{2}\sum_{c\in\mathcal{C}} S_{i(c)}\langle \boldsymbol{\xi}_c, \mathbf{e}_{\phi,c}\rangle^2. \end{aligned} \tag{A.16}$$

Here admissibility requires $s_c > -1$ for every cell.

The common background term has coefficient L_0; the turn direction of cell c has total coefficient $L_0 + \rho S_{i(c)}$. These coefficients describe an upper model, not exact Hessian eigenvalues. The condition covers the finite map's nonlinear effect and is not implied by the isolated GGN bound alone. Other responses, cell norms, cross terms, and non-GGN curvature can also matter. The constant ρ is not the downstream weight bound β. Neither L_0 nor ρ is needed to run the practical algorithm.

Proposition A.5 (Conditional step-size benefit)**.** *Under Assumption A.4, fix $p_i \in (0, 1]$ and set*

$$L(p) := \max\{L_0, \max_i p_i(L_0 + \rho S_i)\}. \tag{A.17}$$

For the ideal gradient update without decoupled decay, write $\mathbf{g} = \nabla f(\mathbf{W})$ and $\mathbf{P} = \mathbf{P}_{\mathbf{W}}$. Whenever $-\eta\mathbf{P}\mathbf{g}$ is admissible and lies in the assumed neighborhood,

$$f(\mathrm{Retr}_{\mathbf{W}}(-\eta\mathbf{P}\mathbf{g})) \leq f(\mathbf{W}) - \eta\big(1 - \eta L(p)/2\big)\mathbf{g}^\top\mathbf{P}\mathbf{g}. \tag{A.18}$$

In particular, $0 < \eta \leq 1/L(p)$ guarantees a decrease of at least $\eta\mathbf{g}^\top\mathbf{P}\mathbf{g}/2$. For $p_i = p_i^{\mathrm{ana}}$ and $\kappa = \rho/L_0$, $L(p) = L_0$; for $p_i = 1$, $L(p) = L_0 + \rho S_{\max}$, where $S_{\max} := \max_i S_i$.

Proof. Let $\mathbf{H}_0$ denote the quadratic matrix on the right-hand side of Eq. (A.16). In the local orthonormal scale–turn basis, its coefficients are L_0 on scale and non-PA directions and $L_0 + \rho S_{i(c)}$ on turns. The corresponding coefficients of $\mathbf{P}$ are 1 and $p_{i(c)}$. Thus $\mathbf{P}^{1/2}\mathbf{H}_0\mathbf{P}^{1/2}$ has coefficients L_0 and $p_{i(c)}(L_0 + \rho S_{i(c)})$; these are the preconditioned upper-model coefficients referred to in the main text. Equivalently, direction by direction,

$$\mathbf{P}\mathbf{H}_0\mathbf{P} \preceq L(p)\mathbf{P}.$$

Substituting $\boldsymbol{\xi} = -\eta \mathbf{P}\mathbf{g}$ in the assumed bound gives

$$f(\mathrm{Retr}_{\mathbf{W}}(-\eta \mathbf{P}\mathbf{g})) \le f(\mathbf{W}) - \eta \mathbf{g}^\top \mathbf{P}\mathbf{g} + \frac{\eta^2}{2} \mathbf{g}^\top \mathbf{P}\mathbf{H}_0\mathbf{P}\mathbf{g},$$

which proves Eq. (A.18). For $\eta \le 1/L(p)$, $1 - \eta L(p)/2 \ge 1/2$. Since $\mathbf{P}$ is positive definite, the stated decrease is strictly positive unless $\mathbf{g} = 0$. The preconditioner is evaluated at the fixed current state, so its derivative is not needed for this one-step bound.

For analytic PA, substituting p_i^{ana} gives

$$L_\kappa = \max\left\{ L_0, \max_i \frac{L_0 + \rho S_i}{1 + \kappa S_i} \right\}. \tag{A.19}$$

At $\kappa = \rho/L_0$, every turn-direction ratio equals L_0, as stated in Eq. (20). At $\kappa = 0$, the maximum is $L_0 + \rho S_{\max}$. This proves the comparison, including $\rho = 0$ and zero-sensitivity directions. □

The equality at the matched strength removes frequency variation from the *preconditioned quadratic upper model*. It does not make all actual loss curvatures equal, nor does it make the response-only rates $\eta_\phi p_i^{\mathrm{ana}} S_i$ exactly equal. The two conclusions concern different quantities: the rate-ratio result holds without the loss-bound assumption, while the present proposition uses that additional assumption to guarantee finite-step loss decrease.

For the practical profile, the same proposition uses $L(p) = \max\{L_0, \max_i (L_0 + \rho S_i)/(1 + \kappa\theta_i^2)\}$, which need not equal L_0. The comparison at $p_i = 1$ uses the same finite scale–rotation map, not Cartesian addition. A smaller $L(p)$ relaxes one sufficient step-size restriction, but changes $\mathbf{g}^\top \mathbf{P}\mathbf{g}$ and does not remove the neighborhood requirement. It therefore does not, by itself, prove faster convergence for an arbitrary base optimizer or practical training run.

B The RoPE-Specific Attention Derivative and Its Scope

We fix the chosen head's input activations and all parameters except its Q/K projections. The first two subsections derive the attention response; the remaining subsections establish its loss weighting, include all Q/K perturbations, analyze cancellations, and establish sharpness with an explicit attention model. Every parameter cell used in scale–turn coordinates is nonzero. All score vectors and Jacobians use a fixed order of the unmasked interactions in $\mathcal{I}$.

B.1 From cell turns to the attention-score derivative

For interaction (m, n), signed displacement $h = n - m$, and RoPE pair i,

$$a_{mn}^{(i)} = \sum_{j,\ell} x_{m,j} x_{n,\ell} \mathbf{w}_{Q,i,j}^\top \mathbf{R}(h\theta_i) \mathbf{w}_{K,i,\ell}. \tag{B.1}$$

Set $\mathbf{R}_{i,h} = \mathbf{R}(h\theta_i)$ and, to shorten the formulas, $\mathbf{e}_{Q,j} = \mathbf{e}_{r,Q,i,j}$, $\mathbf{e}_{K,\ell} = \mathbf{e}_{r,K,i,\ell}$, $\mathbf{e}_{p,j}^\perp = \mathbf{J}\mathbf{e}_{p,j}$, and $c_{mn,j\ell} = x_{m,j} x_{n,\ell} r_{Q,i,j} r_{K,i,\ell}$. Let $\delta a_{mn}^{(i)}$ be the derivative of the score when each Q/K cell is perturbed by $\varepsilon \boldsymbol{\xi}_{p,i,j}$, at $\varepsilon = 0$.

Sec. 3.1 shows, through a single query-cell turn, how the rotation-difference operator $\mathbf{R}(h\theta_i) - \mathbf{I}$ enters an actual attention-score derivative. We now extend this calculation to all query and key cells and give the complete decomposition of the resulting score response.

Lemma B.1 (RoPE-specific turning response of one score)**.** *For the local perturbations above,* $\delta a_{mn}^{(i)} = \delta a_{\mathrm{rem},mn}^{(i)} + \delta a_{\mathrm{R},mn}^{(i)}$, *where*

$$\begin{aligned} \delta a_{\mathrm{R},mn}^{(i)} = &\sum\nolimits_j x_{m,j} \phi_{Q,i,j} (\mathbf{J}\mathbf{w}_{Q,i,j})^\top \big(\mathbf{R}(h\theta_i) - \mathbf{I}\big) \mathbf{k}_n^{(i)} \\ &+ \sum\nolimits_\ell x_{n,\ell} \phi_{K,i,\ell} (\mathbf{q}_m^{(i)})^\top \big(\mathbf{R}(h\theta_i) - \mathbf{I}\big) \mathbf{J}\mathbf{w}_{K,i,\ell}. \end{aligned} \tag{B.2}$$

The remainder collects all scale terms and the turn terms with the identity rotation. In particular, $\delta a_{\mathrm{R},mn}^{(i)} = 0$ *when* $\mathbf{R}(h\theta_i) = \mathbf{I}$.

Lemma B.1 identifies how RoPE changes an attention score's response to parameter cell turns: $\delta a^{(i)}_{\mathrm{R},mn}$ compares the first-order score changes induced by the same turns with and without relative rotation. In Eq. (B.2), the first sum collects query-cell turns, and the second collects key-cell turns. When $\mathbf{R}(h\theta_i) = \mathbf{I}$, this difference vanishes, although the total score change need not: the remainder still contains scale changes and ordinary turning effects. Crucially, the same operator $\mathbf{R}(h\theta_i) - \mathbf{I}$ used in Sec. 2 now appears in the derivative of an actual attention score, with cell magnitudes and query-key interactions retained.

Proof of Lemma B.1. Differentiating the score gives

$$\delta a^{(i)}_{mn} = \sum_{j,\ell} x_{m,j} x_{n,\ell} \big[\boldsymbol{\xi}^\top_{Q,i,j} \mathbf{R}_{i,h} \mathbf{w}_{K,i,\ell} + \mathbf{w}^\top_{Q,i,j} \mathbf{R}_{i,h} \boldsymbol{\xi}_{K,i,\ell} \big]. \tag{B.3}$$

Substituting $\mathbf{w} = r\mathbf{e}_r$ and $\boldsymbol{\xi} = r(s\mathbf{e}_r + \phi\mathbf{e}_\phi)$, and splitting $\mathbf{R}_{i,h} = \mathbf{I} + (\mathbf{R}_{i,h} - \mathbf{I})$ only in the turn terms, gives the expanded decomposition

$$\delta a^{(i)}_{mn} = \delta a^{(i)}_{\mathrm{rem},mn} + \delta a^{(i)}_{\mathrm{R},mn}, \tag{B.4}$$

where

$$\begin{aligned} \delta a^{(i)}_{\mathrm{rem},mn} = \sum_{j,\ell} c_{mn,j\ell} \big[& (s_{Q,i,j} + s_{K,i,\ell}) \mathbf{e}^\top_{Q,j} \mathbf{R}_{i,h} \mathbf{e}_{K,\ell} \\ & + \phi_{Q,i,j} (\mathbf{e}^\perp_{Q,j})^\top \mathbf{e}_{K,\ell} + \phi_{K,i,\ell} \mathbf{e}^\top_{Q,j} \mathbf{e}^\perp_{K,\ell} \big], \end{aligned} \tag{B.5}$$

$$\begin{aligned} \delta a^{(i)}_{\mathrm{R},mn} = \sum_{j,\ell} c_{mn,j\ell} \big[& \phi_{Q,i,j} (\mathbf{e}^\perp_{Q,j})^\top (\mathbf{R}_{i,h} - \mathbf{I}) \mathbf{e}_{K,\ell} \\ & + \phi_{K,i,\ell} \mathbf{e}^\top_{Q,j} (\mathbf{R}_{i,h} - \mathbf{I}) \mathbf{e}^\perp_{K,\ell} \big]. \end{aligned} \tag{B.6}$$

In particular, $\delta a^{(i)}_{\mathrm{R},mn} = 0$ when $\mathbf{R}_{i,h} = \mathbf{I}$.

The scale terms keep $\mathbf{R}_{i,h}$, while the identity-turn terms enter the remainder. Finally, use $r_{p,i,j}\mathbf{e}_{\phi,p,i,j} = \mathbf{J}\mathbf{w}_{p,i,j}$ and sum the unchanged opposite-side cells back into $\mathbf{k}^{(i)}_n$ or $\mathbf{q}^{(i)}_m$. This gives Eq. (B.2). No first-order term is omitted; products of query and key perturbations are second order. The remainder is not a RoPE-free score derivative. □

B.2 Matrix entries and proof of rotary weighting

This subsection expands the coefficient construction used in Lemma 3.1. Retain the fixed order of interactions in $\mathcal{I}$, and let $\mathcal{I}_h = \{(m,n) \in \mathcal{I} : n - m = h\}$ and $n_h = |\mathcal{I}_h|$. The row submatrix $\mathbf{J}_{\mathrm{R},i,h}$ keeps exactly these interactions, without changing the column order: query-cell turns first, followed by key-cell turns, as in $\boldsymbol{\phi}_i$.

Explicit matrix entries. For $(m,n) \in \mathcal{I}_h$, the coefficients of $\phi_{Q,i,j}$ and $\phi_{K,i,\ell}$ in Eq. (B.2) are

$$[\mathbf{J}_{\mathrm{R},i,h}]_{(m,n),(Q,j)} = x_{m,j} (\mathbf{J}\mathbf{w}_{Q,i,j})^\top (\mathbf{R}(h\theta_i) - \mathbf{I}) \mathbf{k}^{(i)}_n, \tag{B.7}$$

$$[\mathbf{J}_{\mathrm{R},i,h}]_{(m,n),(K,\ell)} = x_{n,\ell} (\mathbf{q}^{(i)}_m)^\top (\mathbf{R}(h\theta_i) - \mathbf{I}) \mathbf{J}\mathbf{w}_{K,i,\ell}. \tag{B.8}$$

Factoring out the common rotary scalar leaves the coefficients

$$[\mathbf{G}_{i,h}]_{(m,n),(Q,j)} = x_{m,j} (\mathbf{J}\mathbf{w}_{Q,i,j})^\top \mathbf{R}(h\theta_i/2) \mathbf{J}\mathbf{k}^{(i)}_n, \tag{B.9}$$

$$[\mathbf{G}_{i,h}]_{(m,n),(K,\ell)} = x_{n,\ell} (\mathbf{q}^{(i)}_m)^\top \mathbf{R}(h\theta_i/2) \mathbf{J} (\mathbf{J}\mathbf{w}_{K,i,\ell}). \tag{B.10}$$

Both matrices have shape $n_h \times 2d_{\mathrm{model}}$. Since $\mathbf{J}\mathbf{w}_{p,i,j} = r_{p,i,j}\mathbf{e}_{\phi,p,i,j}$, these expressions are also the cell-magnitude and unit-direction form of the same coefficients. In particular, $\mathbf{G}_{i,h}$ retains the cell norms; normalization from Sec. 2 has not been silently applied to the actual attention score.

Proof of Lemma 3.1. The exact planar identity

$$\mathbf{R}(\alpha) - \mathbf{I} = 2\sin(\alpha/2)\, \mathbf{R}(\alpha/2)\mathbf{J} \tag{B.11}$$

follows by expanding $\mathbf{R}(\alpha)$. Substitute $\alpha = h\theta_i$ into Eqs. (B.7)–(B.8). Every entry becomes $2\sin(h\theta_i/2)$ times the corresponding entry in Eqs. (B.9)–(B.10). This proves Eq. (10) without dividing by the scalar, including when it is zero.

Different displacement groups contain disjoint rows and exhaust $\mathcal{I}$. Thus

$$\begin{aligned}\mathbf{J}_{\mathrm{R},i}^{\top}\mathbf{J}_{\mathrm{R},i} &= \sum_{h:n_h>0} \mathbf{J}_{\mathrm{R},i,h}^{\top}\mathbf{J}_{\mathrm{R},i,h} \\ &= \sum_{h:n_h>0} 4\sin^2(h\theta_i/2)\,\mathbf{G}_{i,h}^{\top}\mathbf{G}_{i,h}.\end{aligned} \tag{B.12}$$

Dividing by n_{att} and using $\pi(h) = n_h/n_{\mathrm{att}}$ gives Eq. (10). Equivalently, for any joint turn vector,

$$\frac{\|\mathbf{J}_{\mathrm{R},i}\boldsymbol{\phi}_i\|_2^2}{n_{\mathrm{att}}} = \sum_{h:n_h>0} \pi(h)\,4\sin^2(h\theta_i/2)\frac{\|\mathbf{G}_{i,h}\boldsymbol{\phi}_i\|_2^2}{n_h}.$$

This is the scalar average-squared-response interpretation of the matrix identity in the main text. □

To identify the mean rotary weight with S_i, use the same π in Sec. 2. Opposite signed displacements have identical squared weights, though their score derivatives can differ. Included self-interactions have zero rotary weight but affect the normalization. The entries of $\mathbf{G}_{i,h}$ still contain a frequency-dependent rotation, so the identity neither makes this matrix frequency independent nor factors its average into S_i times an unweighted average geometry.

B.3 Loss weighting, pairwise response bound, and sharpness

Use $\mathbf{a}$, $\mathbf{y}(\mathbf{a})$ and the twice continuously differentiable convex output loss ℓ from Sec. 3.3. Write $\mathbf{J}_y = \partial\mathbf{y}/\partial\mathbf{a}$ for brevity and use $\mathbf{B}_a$ as defined in Eq. (11). The downstream map includes the usual attention-score scaling. For a fixed score change $\delta\mathbf{a}$, Taylor expansion along the linearized output change gives

$$\begin{aligned}\ell(\mathbf{y} + \varepsilon\mathbf{J}_y\delta\mathbf{a}) = \ell(\mathbf{y}) &+ \varepsilon(\nabla_{\mathbf{y}}\ell)^{\top}\mathbf{J}_y\delta\mathbf{a} \\ &+ \frac{\varepsilon^2}{2}\delta\mathbf{a}^{\top}\mathbf{B}_a\delta\mathbf{a} + o(\varepsilon^2).\end{aligned} \tag{B.13}$$

This is the quadratic weighting of the *linearized* network response, not the full nonlinear loss change. Substituting $\delta\mathbf{a} = \mathbf{J}_{\mathrm{R},i}\boldsymbol{\phi}_i$ recovers $\mathbf{H}_{\mathrm{R},i}$ in Eq. (12). It acts on one pair's turn coordinates, not all Cartesian Q/K perturbations.

We first prove the upper-bound part of Proposition 3.2, and then establish its order-sharpness by an explicit softmax-attention construction.

B.3.1 Upper-bound proof

Upper-bound part of Proposition 3.2. Apply the assumed Γ_i bound to the nonnegative weighted average in Lemma 3.1:

$$\frac{\mathbf{J}_{\mathrm{R},i}^{\top}\mathbf{J}_{\mathrm{R},i}}{n_{\mathrm{att}}} \preceq \Gamma_i\mathbb{E}_{h\sim\pi}[4\sin^2(h\theta_i/2)]\mathbf{I} = \Gamma_i S_i\mathbf{I}. \tag{B.14}$$

Then $0 \preceq \mathbf{B}_a \preceq \beta\mathbf{I}$ gives, for every turn vector,

$$\begin{aligned}0 \le \boldsymbol{\phi}_i^{\top}\mathbf{H}_{\mathrm{R},i}\boldsymbol{\phi}_i &= (\mathbf{J}_{\mathrm{R},i}\boldsymbol{\phi}_i)^{\top}\mathbf{B}_a(\mathbf{J}_{\mathrm{R},i}\boldsymbol{\phi}_i) \\ &\le \beta\|\mathbf{J}_{\mathrm{R},i}\boldsymbol{\phi}_i\|_2^2 \le n_{\mathrm{att}}\beta\Gamma_i S_i\|\boldsymbol{\phi}_i\|_2^2.\end{aligned} \tag{B.15}$$

This proves Eq. (13); equivalently, $0 \preceq \mathbf{H}_{\mathrm{R},i} \preceq n_{\mathrm{att}}\beta\Gamma_i S_i\mathbf{I}$. If $S_i = 0$, the nonnegative rotary weights vanish on every group with positive $\pi(h)$, so $\mathbf{J}_{\mathrm{R},i} = 0$ and the bound holds with equality at zero. □

The loss weight may couple scores at different distances: applying $\mathbf{B}_a \preceq \beta\mathbf{I}$ before distance averaging requires no block-diagonal assumption. The factors β, Γ_i retain model- and data-dependent effects, so the bound does not order total curvature by S_i.

B.3.2 ATTAINING THE LINEAR SENSITIVITY SCALE WITH SOFTMAX ATTENTION

The upper bound would be less informative if every admissible response vanished much faster than S_i. We rule this out by constructing a causal attention head with ordinary softmax attention and cross-entropy loss. The projections, inputs, head dimension, and loss stay fixed across the family; only the RoPE base varies. Within each member, frequencies remain fixed when parameters are differentiated. This is a sharpness construction, not a model of every training checkpoint or a modification of the PA algorithm.

Proposition B.2 (Order-sharpness in a causal attention-to-loss model)**.** *There is a two-token causal attention family with $d_{\mathrm{model}} = d_h = 4$, standard geometric RoPE frequencies, fixed linear projections, and binary cross-entropy loss such that, for pair $i = 2$, $S_2 \to 0^+$. The bounds in Proposition 3.2 hold uniformly with $\beta = 1/128$ and $\Gamma_2 = 2$. A fixed unit turn vector $\mathbf{v}_2$ and its corresponding unit Cartesian Q/K direction $\boldsymbol{\xi}_*$ satisfy*

$$\frac{\mathbf{v}_2^\top \mathbf{H}_{\mathrm{R},2}\mathbf{v}_2}{n_{\mathrm{att}}S_2} = \frac{\boldsymbol{\xi}_*^\top \mathbf{H}_{\mathrm{GN}}\boldsymbol{\xi}_*}{n_{\mathrm{att}}S_2} \longrightarrow c_* > 0. \tag{B.16}$$

Thus no uniform upper bound over this class can replace its linear S_i scale by $o(S_i)$ while keeping the other constants bounded independently of S_i.

Proof. A fixed attention model. Let $0 < \varepsilon < 1$ and set the RoPE base to $b = \varepsilon^{-2}$. The standard schedule $\theta_i = b^{-2(i-1)/d_h}$ gives $\theta_1 = 1$ and $\theta_2 = \varepsilon$. Choose inputs $\mathbf{x}_1 = (1,0,0,0)^\top$ and $\mathbf{x}_2 = (0,1,0,0)^\top$, and let

$$\mathbf{W}_Q = \mathbf{W}_K = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

All Q/K parameter cells have unit norm. In both pairs and at both positions, $\mathbf{q}_m^{(i)} = \mathbf{k}_m^{(i)} = (1,0)^\top$. The fixed value projection $\mathbf{W}_V = \mathrm{diag}(1,0,0,0)$ produces $\mathbf{v}_1^{\mathrm{val}} = (1,0,0,0)^\top$ and $\mathbf{v}_2^{\mathrm{val}} = \mathbf{0}$. A fixed readout uses the first component of token 2's attention output as a scalar classification logit. The score scale is $c = 1/\sqrt{d_h} = 1/2$.

A single parameter-cell turn. Order the unmasked interactions as $(1,1),(2,1),(2,2)$, so $n_{\mathrm{att}} = 3$. Turn only query cell $\mathbf{w}_{Q,2,2}$ through an angle ϕ, keeping its norm fixed. Its derivative at zero is $\mathbf{J}(1,0)^\top = (0,1)^\top$, while all other parameter derivatives are zero. The total unscaled scores are

$$a_{11} = 2, \qquad a_{21}(\phi) = \cos 1 + \cos(\varepsilon + \phi), \qquad a_{22}(\phi) = 1 + \cos\phi. \tag{B.17}$$

The sign follows from $h = 1 - 2 = -1$ in the past-token interaction. Their derivative vector at $\phi = 0$ is $(0, -\sin\varepsilon, 0)^\top$. Here the query and key of the perturbed pair are aligned at the current state, so every identity-rotation turning term in this direction is zero. There is no scale change and no turn in the other pair. Hence the entire score derivative is RoPE-specific:

$$\mathbf{J}_a\boldsymbol{\xi}_* = \mathbf{J}_{\mathrm{R},2}\mathbf{v}_2 = (0, -\sin\varepsilon, 0)^\top, \qquad \mathbf{J}_{\mathrm{rem}}\boldsymbol{\xi}_* = 0, \tag{B.18}$$

where $\mathbf{v}_2 = (0,1,0,0,0,0,0,0)^\top$ selects the query-cell turn in the established Q-then-K order, and $\boldsymbol{\xi}_*$ has only $\boldsymbol{\xi}_{Q,2,2} = (0,1)^\top$ nonzero. Both directions have unit norm.

Actual softmax and output-loss weighting. Let $\sigma(t) = (1 + e^{-t})^{-1}$. Token 2's attention probability on token 1, and thus the scalar readout logit at the current state, is

$$t_\varepsilon = \sigma(c\,[\cos 1 + \cos\varepsilon - 2])\,. \tag{B.19}$$

Use the positive-label binary cross-entropy $\ell(z) = \log(1 + e^{-z})$, equivalently two output logits $(z, 0)$. Then $\ell''(z) = \sigma(z)(1 - \sigma(z)) > 0$. The score-to-logit derivative is $c\,t_\varepsilon(1 - t_\varepsilon)(0, 1, -1)$, so

$$\mathbf{B}_a = \chi_\varepsilon \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}, \qquad \chi_\varepsilon := c^2\ell''(t_\varepsilon)t_\varepsilon^2(1 - t_\varepsilon)^2. \tag{B.20}$$

This weight is not prescribed to contain S_2: it is obtained by differentiating the softmax and the fixed cross-entropy loss. Its nonzero cross term couples the distance--1 and distance-0 scores.

Since $\ell'' \le 1/4$ and $t^2(1-t)^2 \le 1/16$, its largest eigenvalue satisfies $2\chi_\varepsilon \le c^2/32 = 1/128$. Moreover,

$$t_\varepsilon \longrightarrow t_0 := \sigma(c(\cos 1 - 1)) \in (0,1), \qquad \chi_\varepsilon \longrightarrow \chi_0 := c^2 \ell''(t_0) t_0^2 (1-t_0)^2 > 0. \tag{B.21}$$

Uniform control of the remaining geometry. The displacement groups have $n_{-1} = 1$ and $n_0 = 2$. Direct substitution in Eqs. (B.9)–(B.10) gives

$$\begin{aligned} \mathbf{G}_{2,-1} &= \cos(\varepsilon/2)(0,1,0,0,-1,0,0,0), \\ \mathbf{G}_{2,0} &= \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \end{bmatrix}. \end{aligned} \tag{B.22}$$

Consequently, $\|\mathbf{G}_{2,-1}\|_{\mathrm{op}}^2/n_{-1} = 2\cos^2(\varepsilon/2) \le 2$ and $\|\mathbf{G}_{2,0}\|_{\mathrm{op}}^2/n_0 = 1$. Thus $\Gamma_2 = 2$ is uniform, including the zero-rotary-weight self group. The first pair has the same bound, with ε replaced by 1. All ordinary turning coefficients vanish at this aligned state, and each row of the Cartesian remainder matrix has at most four scale coefficients of magnitude at most one. Hence $\|\mathbf{J}_{\mathrm{rem}}\|_{\mathrm{op}} \le \|\mathbf{J}_{\mathrm{rem}}\|_{\mathrm{F}} \le \sqrt{12}$ uniformly as well. No model or bound constant grows as S_2 decreases.

The attainable S_2 scale. Using exactly the interaction-count distribution from the main text, $\pi(-1) = 1/3$ and $\pi(0) = 2/3$, gives

$$S_2 = \frac{4}{3}\sin^2(\varepsilon/2). \tag{B.23}$$

Equations (B.18) and (B.20) imply

$$\begin{aligned} \mathbf{v}_2^\top \mathbf{H}_{\mathrm{R},2} \mathbf{v}_2 &= \boldsymbol{\xi}_*^\top \mathbf{H}_{\mathrm{GN}} \boldsymbol{\xi}_* \\ &= \chi_\varepsilon \sin^2 \varepsilon = n_{\mathrm{att}} \chi_\varepsilon \cos^2(\varepsilon/2) S_2. \end{aligned} \tag{B.24}$$

The normalized ratio therefore tends to $c_* := \chi_0 > 0$, proving Eq. (B.16) and the sharpness part of Proposition 3.2. In particular, positive finite constants c_-, c_+ and $\varepsilon_0 > 0$ exist such that

$$c_- S_2 \le \frac{\mathbf{v}_2^\top \mathbf{H}_{\mathrm{R},2} \mathbf{v}_2}{n_{\mathrm{att}}} \le c_+ S_2, \qquad 0 < \varepsilon \le \varepsilon_0.$$

If a uniform bound with $C\, g(S_2)$ instead of a constant times S_2 held for this class, with C independent of ε and $g(s) = o(s)$, division by S_2 would force the left-hand ratio to zero, contradicting its positive limit. □

The conditional lower-bound assumptions are also realizable in this example. On the one-dimensional turn subspace $\mathbf{U}_2 = \mathbf{v}_2$, the remaining normalized response in Eq. (B.36) is $\cos^2(\varepsilon/2) \ge \cos^2(1/2) > 0$. Its score response is along $(0,1,0)^\top$, where the downstream Rayleigh quotient is χ_ε. Continuity and positivity on $[0,1]$ give a common lower bound $\mu = \min_{0 \le \varepsilon \le 1} \chi_\varepsilon > 0$. Thus the attention and loss nondegeneracy conditions hold on this direction, although they do not hold on the entire turn space.

The construction shows that the linear S_i order is attainable, including along a direction of the full Q/K-block GGN where the remainder does not cancel it. It does not make the numerical constants in every bound optimal, assert a positive lower bound in all directions, or characterize the exact nonlinear Hessian of a pretrained Transformer. Frequencies are fixed within each member; varying the base only establishes uniform sharpness across the admissible class. In particular, a fixed finite frequency grid in one trained model need not exhibit the limiting regime. A coarser upper bound independent of S_i remains possible using $S_i \le 4$; what cannot be uniformly improved is the stated small-S_i order under bounded remaining geometry.

B.4 All Q/K perturbations and the full local GGN

To include scale changes, all pairs, and their interactions, collect the Cartesian cell perturbations in a fixed order:

$$\begin{aligned} \boldsymbol{\xi}_{p,i} &= \left(\boldsymbol{\xi}_{p,i,1}^\top, \ldots, \boldsymbol{\xi}_{p,i,d_{\mathrm{model}}}^\top\right)^\top \in \mathbb{R}^{2d_{\mathrm{model}}}, \\ \boldsymbol{\xi}_p &= \left(\boldsymbol{\xi}_{p,1}^\top, \ldots, \boldsymbol{\xi}_{p,N}^\top\right)^\top, \qquad p \in \{Q, K\}, \\ \boldsymbol{\xi} &= \left(\boldsymbol{\xi}_Q^\top, \boldsymbol{\xi}_K^\top\right)^\top \in \mathbb{R}^{4Nd_{\mathrm{model}}}. \end{aligned} \tag{B.25}$$

Each $\boldsymbol{\xi}_{p,i,j} \in \mathbb{R}^2$ has scale and turn coordinates $s_{p,i,j} = \langle \boldsymbol{\xi}_{p,i,j}, \mathbf{e}_{r,p,i,j} \rangle / r_{p,i,j}$ and $\phi_{p,i,j} = \langle \boldsymbol{\xi}_{p,i,j}, \mathbf{e}_{\phi,p,i,j} \rangle / r_{p,i,j}$. Thus $\boldsymbol{\phi}_i$ is a linear function of $\boldsymbol{\xi}$ at the fixed state, not an independent perturbation.

Order the interactions as (m_k, n_k), $k = 1, \dots, n_{\mathrm{att}}$, and set

$$\mathbf{a} = (a_{m_1 n_1}, \dots, a_{m_{n_{\mathrm{att}}} n_{n_{\mathrm{att}}}})^\top, \qquad a_{mn} = \sum_{i=1}^{N} a_{mn}^{(i)}. \tag{B.26}$$

Let $\mathbf{J}_a = \partial \mathbf{a} / \partial (\mathbf{W}_Q, \mathbf{W}_K)$ use this interaction order and the parameter order in Eq. (B.25). Its product with $\boldsymbol{\xi}$ gives the total-score changes. Define the components in the same order:

$$\mathbf{J}_{\mathrm{R},i} \boldsymbol{\phi}_i = \mathrm{col}_{(m,n)\in\mathcal{I}} (\delta a_{\mathrm{R},mn}^{(i)}), \tag{B.27}$$

$$\mathbf{J}_{\mathrm{rem}} \boldsymbol{\xi} = \mathrm{col}_{(m,n)\in\mathcal{I}} \left(\sum_i \delta a_{\mathrm{rem},mn}^{(i)} \right). \tag{B.28}$$

Here col stacks scalar entries. The first identity restates Eq. (9); the second defines the remainder map using Eq. (B.5). Its scale terms can still depend on RoPE.

Lemma B.3 (Aggregation of the RoPE-specific turning response)**.** *For the parameter and interaction orders above,*

$$\mathbf{J}_a \boldsymbol{\xi} = \mathbf{J}_{\mathrm{rem}} \boldsymbol{\xi} + \sum_i \mathbf{J}_{\mathrm{R},i} \boldsymbol{\phi}_i. \tag{B.29}$$

Proof. For each interaction, Eq. (B.4) gives $\delta a_{mn} = \sum_i \delta a_{\mathrm{rem},mn}^{(i)} + \sum_i \delta a_{\mathrm{R},mn}^{(i)}$. Stack these equalities using Eqs. (B.27) and (B.28). □

Keeping interactions in the loss geometry. The full GGN for this Q/K parameter block is

$$\mathbf{H}_{\mathrm{GN}} := \mathbf{J}_a^\top \mathbf{B}_a \mathbf{J}_a. \tag{B.30}$$

Writing $\|\mathbf{v}\|_{\mathbf{B}_a}^2 = \mathbf{v}^\top \mathbf{B}_a \mathbf{v}$, Lemma B.3 gives

$$\boldsymbol{\xi}^\top \mathbf{H}_{\mathrm{GN}} \boldsymbol{\xi} = \left\| \mathbf{J}_{\mathrm{rem}} \boldsymbol{\xi} + \sum_i \mathbf{J}_{\mathrm{R},i} \boldsymbol{\phi}_i \right\|_{\mathbf{B}_a}^2. \tag{B.31}$$

The loss weights the sum of the score changes, not their squared sizes separately. Expanding the square produces cross terms between pairs and with the remainder; these can have either sign. Thus $\mathbf{H}_{\mathrm{R},i}$ is neither an independent additive Hessian block nor the difference between the GGN with and without RoPE. The next bound controls the cross terms instead of assuming that they vanish.

Proposition B.4 (Upper bound for the full local GGN)**.** *Suppose* $0 \preceq \mathbf{B}_a \preceq \beta \mathbf{I}$, $\|\mathbf{J}_{\mathrm{rem}}\|_{\mathrm{op}} \le M$, *and* $\mathbf{G}_{i,h}^\top \mathbf{G}_{i,h} / n_h \preceq \Gamma_i \mathbf{I}$ *for all pairs and nonempty distance groups, with* $\beta, M, \Gamma_i \ge 0$*. Then every Q/K perturbation satisfies*

$$\tfrac{1}{2} \boldsymbol{\xi}^\top \mathbf{H}_{\mathrm{GN}} \boldsymbol{\xi} \le \beta M^2 \|\boldsymbol{\xi}\|_2^2 + \beta n_{\mathrm{att}} N \sum_i \Gamma_i S_i \|\boldsymbol{\phi}_i\|_2^2. \tag{B.32}$$

Proof. Use $\frac{1}{2}\|u+v\|_{\mathbf{B}_a}^2 \le \|u\|_{\mathbf{B}_a}^2 + \|v\|_{\mathbf{B}_a}^2$ in Eq. (B.31). The remainder term is at most $\beta M^2 \|\boldsymbol{\xi}\|_2^2$. For the sum of turning responses, Cauchy–Schwarz and Eq. (B.14) give

$$\left\| \sum_i \mathbf{J}_{\mathrm{R},i} \boldsymbol{\phi}_i \right\|_{\mathbf{B}_a}^2 \le \beta N \sum_i \|\mathbf{J}_{\mathrm{R},i} \boldsymbol{\phi}_i\|_2^2 \le \beta n_{\mathrm{att}} N \sum_i \Gamma_i S_i \|\boldsymbol{\phi}_i\|_2^2.$$

Adding the two bounds proves the result. □

The first term controls the remainder; the second exposes the sensitivity weights even when all pairs are considered together. However,

$$\|\boldsymbol{\phi}_i\|_2^2 = \sum_{p,j} \frac{\|\boldsymbol{\Pi}_{\phi,p,i,j} \boldsymbol{\xi}_{p,i,j}\|_2^2}{r_{p,i,j}^2}, \qquad \boldsymbol{\Pi}_{\phi,p,i,j} := \mathbf{e}_{\phi,p,i,j} \mathbf{e}_{\phi,p,i,j}^\top.$$

Cell norms and data-dependent factors therefore remain in the bound. It does not identify the full nonlinear loss Hessian, order total curvature by S_i, or imply Assumption A.4 without further conditions.

B.5 JOINT CANCELLATIONS AND CONDITIONAL LOWER BOUNDS

An upper response bound does not imply a positive response in every direction. Joint turns can cancel, and the upper-bound assumptions also permit a downstream weight that vanishes on the response. We first identify a structural cancellation and then state precisely what a lower bound needs.

Lemma B.5 (Common-turn cancellation)**.** *Let* $\mathbf{c}_i = (1, \dots, 1)^\top \in \mathbb{R}^{2d_{\mathrm{model}}}$ *represent a common turn of all query and key cells of pair* i*. Then*

$$\mathbf{J}_{\mathrm{R},i}\mathbf{c}_i = 0, \qquad \mathbf{G}_{i,h}\mathbf{c}_i = 0. \tag{B.33}$$

Consequently, the corresponding Gram and GGN matrices cannot have a strictly positive identity lower bound on the entire joint turn space.

Proof. A common cell rotation gives $\mathbf{q}' = \mathbf{R}(\alpha)\mathbf{q}$ and $\mathbf{k}' = \mathbf{R}(\alpha)\mathbf{k}$. Since planar rotations commute,

$$(\mathbf{q}')^\top \mathbf{R}(h\theta_i)\mathbf{k}' = \mathbf{q}^\top \mathbf{R}(-\alpha)\mathbf{R}(h\theta_i)\mathbf{R}(\alpha)\mathbf{k} = \mathbf{q}^\top \mathbf{R}(h\theta_i)\mathbf{k}.$$

The same invariance holds with $\mathbf{R}(h\theta_i) - \mathbf{I}$ instead of $\mathbf{R}(h\theta_i)$. Differentiating in α gives $\mathbf{J}_{\mathrm{R},i}\mathbf{c}_i = 0$. The sum of the entries of a row of $\mathbf{G}_{i,h}$ is

$$(\mathbf{J}\mathbf{q})^\top \mathbf{Q}\mathbf{k} + \mathbf{q}^\top \mathbf{Q}\mathbf{J}\mathbf{k} = 0, \qquad \mathbf{Q} = \mathbf{R}(h\theta_i/2)\mathbf{J},$$

because $\mathbf{J}^\top = -\mathbf{J}$ and $\mathbf{Q}\mathbf{J} = \mathbf{J}\mathbf{Q}$. This also proves the second identity when the scalar rotary weight is zero. □

Inputs or parameters can create additional null directions, and $\mathbf{B}_a$ can annihilate a nonzero score response. In particular, $\mathbf{B}_a = 0$ satisfies the upper-bound assumptions but gives $\mathbf{H}_{\mathrm{R},i} = 0$ even if $S_i > 0$. A positive lower bound therefore requires additional nondegeneracy conditions.

A condition on the combined response, rather than every distance. For $S_i > 0$, define normalized rotary weights

$$\omega_i(h) := \frac{\pi(h)\, 4\sin^2(h\theta_i/2)}{S_i}, \qquad \sum_{h:n_h>0} \omega_i(h) = 1. \tag{B.34}$$

These are analytical weights, not attention probabilities. They are positive only at distances carrying nonzero rotary weight. The exact identity in Sec. 3.2 becomes

$$\frac{\mathbf{J}_{\mathrm{R},i}^\top \mathbf{J}_{\mathrm{R},i}}{n_{\mathrm{att}}} = S_i \sum_{h:n_h>0} \omega_i(h) \frac{\mathbf{G}_{i,h}^\top \mathbf{G}_{i,h}}{n_h}. \tag{B.35}$$

This separates a scalar scale from a *rotary-weighted* remaining geometry. That geometry still depends on frequency and can be singular; factoring out S_i alone proves no lower bound.

Proposition B.6 (Conditional two-sided bound on an active turn subspace)**.** *Suppose* $S_i > 0$*, and let* $\mathbf{U}_i$ *have orthonormal columns spanning a nonempty chosen subspace of the joint turn coordinates. Assume*

$$\gamma_i \mathbf{I} \preceq \sum_{h:n_h>0} \omega_i(h) \frac{\mathbf{U}_i^\top \mathbf{G}_{i,h}^\top \mathbf{G}_{i,h} \mathbf{U}_i}{n_h} \preceq \Gamma_i \mathbf{I}, \qquad 0 < \gamma_i \le \Gamma_i, \tag{B.36}$$

and, for every $\mathbf{z} \in \mathrm{range}(\mathbf{J}_{\mathrm{R},i}\mathbf{U}_i)$*,*

$$\mu\|\mathbf{z}\|_2^2 \le \mathbf{z}^\top \mathbf{B}_a \mathbf{z} \le \beta\|\mathbf{z}\|_2^2, \qquad 0 < \mu \le \beta. \tag{B.37}$$

Then

$$n_{\mathrm{att}}\mu\gamma_i S_i \mathbf{I} \preceq \mathbf{U}_i^\top \mathbf{H}_{\mathrm{R},i}\mathbf{U}_i \preceq n_{\mathrm{att}}\beta\Gamma_i S_i \mathbf{I}. \tag{B.38}$$

Proof. For any subspace-coordinate vector $\mathbf{t}$, set $\boldsymbol{\phi}_i = \mathbf{U}_i\mathbf{t}$. Orthonormality gives $\|\boldsymbol{\phi}_i\|_2 = \|\mathbf{t}\|_2$. Equations (B.35) and (B.36) imply

$$n_{\mathrm{att}}\gamma_i S_i \|\mathbf{t}\|_2^2 \le \|\mathbf{J}_{\mathrm{R},i}\mathbf{U}_i\mathbf{t}\|_2^2 \le n_{\mathrm{att}}\Gamma_i S_i \|\mathbf{t}\|_2^2.$$

Apply Eq. (B.37) to $\mathbf{z} = \mathbf{J}_{\mathrm{R},i}\mathbf{U}_i\mathbf{t}$ and use $\mathbf{t}^\top \mathbf{U}_i^\top \mathbf{H}_{\mathrm{R},i}\mathbf{U}_i\mathbf{t} = \mathbf{z}^\top \mathbf{B}_a \mathbf{z}$. The resulting scalar inequalities for every $\mathbf{t}$ prove the matrix statement. □

The first condition means that turns in the chosen subspace produce a nonvanishing combined score response after the explicit rotary scale is removed. The second means that the downstream loss still detects those responses. Together they make S_i a lower as well as an upper scale. The earlier per-distance condition

$$\gamma_i \mathbf{I} \preceq \frac{\mathbf{U}_i^\top \mathbf{G}_{i,h}^\top \mathbf{G}_{i,h} \mathbf{U}_i}{n_h} \preceq \Gamma_i \mathbf{I} \tag{B.39}$$

for every group with nonzero rotary weight is sufficient for Eq. (B.36), by taking its convex combination with weights $\omega_i(h)$. The combined condition is weaker: different distances may reveal different directions, even when no single group has full column rank on the subspace.

Removing the common-turn direction alone guarantees neither condition. Uniform same-order scaling across a family additionally requires μ, γ_i to stay bounded away from zero and β, Γ_i to stay bounded above. Otherwise the lower constants can hide another vanishing factor. When $S_i = 0$, the isolated response is zero and no positive lower bound is claimed. These results concern the isolated response: a lower bound for the full Q/K-block GGN would also need to prevent cancellation with the other responses in Eq. (B.31).

C Practical Profile and Implementation Details

C.1 Analytic PA, standard PA, and PA-Log profiles

These profiles use the same cellwise operator and finite update. They differ only in their fixed multipliers. To distinguish the strengths locally, write κ_{ana} for the analytic rule and κ_{LF} for the low-frequency surrogate. For a fixed finite displacement distribution π, Taylor expansion gives

$$\begin{aligned} 4\sin^2(h\theta_i/2) &= h^2\theta_i^2 + \mathcal{O}(h^4\theta_i^4), \\ S_i &= \mu_2\theta_i^2 + \mathcal{O}(\mu_4\theta_i^4), \qquad \mu_k := \mathbb{E}_\pi[h^k], \quad k \in \{2,4\}, \qquad \theta_i \to 0. \end{aligned} \tag{C.1}$$

Consequently, the leading-order substitution in $p_i^{\mathrm{ana}} = (1 + \kappa_{\mathrm{ana}} S_i)^{-1}$ gives

$$p_i^{\mathrm{LF}} = (1 + \kappa_{\mathrm{LF}}\theta_i^2)^{-1}, \qquad \kappa_{\mathrm{LF}} = \mu_2 \kappa_{\mathrm{ana}}. \tag{C.2}$$

Equation (17) denotes the rescaled strength again by κ. This is the current standard PA profile. The approximation is local to small relevant rotary angles, not uniform over the full frequency grid. Equal numerical strengths in the two formulas generally do not match their attenuation. If $\mu_2 = 0$, the distribution is supported at $h = 0$, so $S_i = 0$ and the matched strength is $\kappa_{\mathrm{LF}} = 0$. Both exact and surrogate multipliers are precomputed, so neither requires extra training-time forward or backward passes.

PA-Log, the separate bounded alternative in Sec. 5.2, uses the normalized log-frequency coordinate

$$\nu_i = \frac{\log\theta_i - \min_k \log\theta_k}{\max_k \log\theta_k - \min_k \log\theta_k}, \qquad \lambda_i = 1 + (\lambda_{\max} - 1)\nu_i, \qquad p_i^{\mathrm{Log}} = \lambda_i^{-2}, \tag{C.3}$$

where $\lambda_{\max} \geq 1$ and the grid contains at least two distinct positive frequencies. The reported bounded-profile comparison uses range $[1, 4]$ (Appendix E.5.2). For $\theta_i = b^{-2(i-1)/d_h}$, $b > 1$ and $N = d_h/2 > 1$, $\nu_i = 1 - (i-1)/(N-1)$. Thus the bounded profile depends on frequency rank, not on b or π at fixed N. It spaces the penalty parameters λ_i evenly, not the multipliers p_i^{Log}. This is an empirical alternative, not clipping of S_i, and it need not follow the sensitivity ordering for an arbitrary displacement distribution.

C.2 Finite map, storage layout, and numerical conventions

The ideal update remains Eq. (16); it is not an exponential scale map or the exact polar representation of an additive endpoint. Appendix A.3.2 quantifies its second-order difference from the local preconditioned increment. The base increment is formed before PA: the recorded implementation uses the bias-corrected Adam direction for PA-AdamW and the direction after momentum, Newton–Schulz orthogonalization, and shape scaling for PA-Muon.

The documented Q/K storage shape is $[n_{\mathrm{head}} d_h, d_{\mathrm{model}}]$, reshaped to $[n_{\mathrm{head}}, d_h, d_{\mathrm{model}}]$. Its RoPE-coupled coordinates are half-split pairs $(i, i + d_h/2)$ within each head and input column. The adjacent coordinates in the mathematical cell definition denote these pairs after permutation. The cached

multiplier is shared across cells with the same frequency index; local directions still depend on each cell.

For a cell $\mathbf{w} = (x, y)^\top$ and a base direction $\mathbf{d} = (d_x, d_y)^\top$, let $\mathbf{\Delta}^{\text{base}} = -\eta\mathbf{d}$ denote the increment in base optimization that excludes decoupled weight decay, where η is the learning rate. To avoid numerical instability for cells with very small norms, we use the squared-radius floor $r_\epsilon^2 = \max\{x^2 + y^2, \epsilon_r\}$ with $\epsilon_r = 10^{-8}$. The numerical PA update is

$$s_\epsilon = -\eta\frac{xd_x + yd_y}{r_\epsilon^2}, \qquad \phi_\epsilon = \eta p_i \frac{yd_x - xd_y}{r_\epsilon^2}, \qquad \widehat{\mathbf{w}} = (1 + s_\epsilon)\mathbf{R}(\phi_\epsilon)\mathbf{w}. \tag{C.4}$$

With an inactive floor, these are exactly s_A and $p_i\phi_A$ in the main text.

Let $\widehat{\mathbf{w}}$ denote the endpoint obtained after applying the PA scale–rotation update to the pre-update cell $\mathbf{w}$. Weight decay is then added separately using the same pre-update parameter:

$$\mathbf{w}^+ = \widehat{\mathbf{w}} - \eta\omega\mathbf{w}, \tag{C.5}$$

where ω is the weight-decay coefficient. The convergence analysis in Appendix D concerns the ideal PA update without the numerical radius floor or the separate weight-decay term. PA adds $\mathcal{O}(|\mathbf{W}_Q| + |\mathbf{W}_K|)$ arithmetic and a cached profile, without another per-parameter moment or curvature tensor.

D Optimization Guarantees and Proofs

These results concern the ideal finite map, not its first-order approximation. They do not assume the structured loss bound of Assumption A.4. We first prove one smoothness inequality along the finite path, then reuse it for deterministic gradients, unbiased stochastic gradients, and general descent-related directions.

D.1 One finite-map bound for all three settings

Use the cell index set $\mathcal{C}$, map $\mathrm{Retr}_{\mathbf{W}}$, and blockwise operator $\mathbf{P}_{\mathbf{W}}$ from Appendix A.4. All multipliers are fixed in $[p_{\min}, 1]$, with $p_{\min} > 0$, while their local bases vary with $\mathbf{W}$. Hence

$$p_{\min}\mathbf{I} \preceq \mathbf{P}_{\mathbf{W}} \preceq \mathbf{I}, \qquad \|\mathbf{P}_{\mathbf{W}}\mathbf{v}\|_2^2 \le \mathbf{v}^\top\mathbf{P}_{\mathbf{W}}\mathbf{v} =: \|\mathbf{v}\|_{\mathbf{P}_{\mathbf{W}}}^2. \tag{D.1}$$

Let $\mathbf{W}_0$ be deterministic and define $\Delta_f = f(\mathbf{W}_0) - f_{\inf}$, $\mathbf{g}_t = \nabla f(\mathbf{W}_t)$, and $\mathbf{P}_t = \mathbf{P}_{\mathbf{W}_t}$. The objective is continuously differentiable, bounded below by $f_{\inf}$, and has an L-Lipschitz gradient, $L > 0$, in the flattened parameter coordinates.

Fix a proof margin $c_r \in (0, 1)$. A tangent step is *admissible* when $|s_c| = |\langle\boldsymbol{\xi}_c, \mathbf{e}_{r,c}\rangle|/r_c \le c_r$ for every cell, so $1 + s_c \ge 1 - c_r > 0$. This is a condition on the step, not an added clipping rule. For a bound $M \ge 0$ on the current gradient, define

$$L_{\mathrm{PA}}(M) := L(1 + c_r)^2 + \frac{2 + c_r}{r_{\min}}M. \tag{D.2}$$

Lemma D.1 (Smoothness along the finite scale–turn map)**.** *At a state $\mathbf{W}$ with cell radii at least $r_{\min} > 0$ and $\|\nabla f(\mathbf{W})\|_2 \le M$, every admissible increment satisfies*

$$f(\mathrm{Retr}_{\mathbf{W}}(\boldsymbol{\xi})) \le f(\mathbf{W}) + \langle\nabla f(\mathbf{W}), \boldsymbol{\xi}\rangle + \frac{L_{\mathrm{PA}}(M)}{2}\|\boldsymbol{\xi}\|_2^2. \tag{D.3}$$

Proof. For one cell, write $\boldsymbol{\xi} = r(s\mathbf{e}_r + \phi\mathbf{e}_\phi)$ and follow the exact path $\boldsymbol{\gamma}(\tau) = (1 + \tau s)\mathbf{R}(\tau\phi)\mathbf{w}$, $0 \le \tau \le 1$. With $\mathbf{u}(\tau) = \mathbf{R}(\tau\phi)\mathbf{e}_r$,

$$\boldsymbol{\gamma}'(\tau) = r[s\mathbf{u}(\tau) + (1 + \tau s)\phi\mathbf{J}\mathbf{u}(\tau)],$$
$$\boldsymbol{\gamma}''(\tau) = r[2s\phi\mathbf{J}\mathbf{u}(\tau) - (1 + \tau s)\phi^2\mathbf{u}(\tau)].$$

Orthogonality, $|s| \le c_r$, and $r \ge r_{\min}$ imply

$$\|\boldsymbol{\gamma}'(\tau)\|_2 \le (1 + c_r)\|\boldsymbol{\xi}\|_2, \qquad \|\boldsymbol{\gamma}''(\tau)\|_2 \le \frac{2 + c_r}{r_{\min}}\|\boldsymbol{\xi}\|_2^2.$$

The same estimates hold for the full path, including additive non-PA coordinates. For the second bound, combine cells using $(\sum_c \|\boldsymbol{\xi}_c\|_2^4)^{1/2} \le \sum_c \|\boldsymbol{\xi}_c\|_2^2$. Since $\boldsymbol{\gamma}(0) = \mathbf{W}$ and $\boldsymbol{\gamma}'(0) = \boldsymbol{\xi}$, integration gives

$$\|\boldsymbol{\gamma}(\tau) - \mathbf{W}\|_2 \le \tau(1 + c_r)\|\boldsymbol{\xi}\|_2, \qquad \|\boldsymbol{\gamma}'(\tau) - \boldsymbol{\xi}\|_2 \le \tau \frac{2 + c_r}{r_{\min}} \|\boldsymbol{\xi}\|_2^2.$$

For $H(\tau) = f(\boldsymbol{\gamma}(\tau))$,

$$\begin{aligned} H'(\tau) - H'(0) &= \langle \nabla f(\boldsymbol{\gamma}(\tau)) - \nabla f(\mathbf{W}), \boldsymbol{\gamma}'(\tau) \rangle + \langle \nabla f(\mathbf{W}), \boldsymbol{\gamma}'(\tau) - \boldsymbol{\xi} \rangle \\ &\le \tau L_{\mathrm{PA}}(M) \|\boldsymbol{\xi}\|_2^2. \end{aligned}$$

Integrating once more proves the claim. Only the gradient bound at the current state is used; the curved path's second-order effect is explicitly retained. □

For any direction $\mathbf{d}$, applying this lemma to the admissible step $\boldsymbol{\xi} = -\eta \mathbf{P}_{\mathbf{W}} \mathbf{d}$ and using Eq. (D.1) yields the common bound

$$f(\mathrm{Retr}_{\mathbf{W}}(-\eta \mathbf{P}_{\mathbf{W}} \mathbf{d})) \le f(\mathbf{W}) - \eta \langle \nabla f(\mathbf{W}), \mathbf{P}_{\mathbf{W}} \mathbf{d} \rangle + \frac{\eta^2 L_{\mathrm{PA}}(M)}{2} \|\mathbf{d}\|_{\mathbf{P}_{\mathbf{W}}}^2. \tag{D.4}$$

The two update terms separate descent alignment from direction size, allowing the same inequality to handle all three specializations below.

D.2 Deterministic gradients and the explicit step-size threshold

Proof of Theorem 4.2. If $\Delta_f = 0$, smoothness and the lower bound imply $\mathbf{g}_0 = 0$, so the update stays fixed and any $\eta_0 > 0$ suffices. Otherwise define $M_0 = \sqrt{2L\Delta_f}$ and

$$\eta_0 := \min\left\{ \frac{1}{L_{\mathrm{PA}}(M_0)}, \frac{c_r r_{\min}}{M_0} \right\}. \tag{D.5}$$

The ordinary smoothness inequality evaluated at $\mathbf{W} - L^{-1}\nabla f(\mathbf{W})$ gives

$$\|\nabla f(\mathbf{W})\|_2^2 \le 2L[f(\mathbf{W}) - f_{\inf}].$$

We prove nonincrease by induction. If $f(\mathbf{W}_t) \le f(\mathbf{W}_0)$, then $\|\mathbf{g}_t\|_2 \le M_0$ and the scale coordinates of $-\eta \mathbf{P}_t \mathbf{g}_t$ satisfy $|s_{c,t}| \le \eta \|\mathbf{g}_t\|_2 / r_{\min} \le c_r$. Thus the step is admissible. Taking $\mathbf{d} = \mathbf{g}_t$ in Eq. (D.4) gives

$$f(\mathbf{W}_{t+1}) \le f(\mathbf{W}_t) - \eta(1 - \eta L_{\mathrm{PA}}(M_0)/2)\|\mathbf{g}_t\|_{\mathbf{P}_t}^2 \le f(\mathbf{W}_t) - \frac{\eta}{2}\|\mathbf{g}_t\|_{\mathbf{P}_t}^2. \tag{D.6}$$

This closes the induction. Summing gives $\frac{\eta}{2} \sum_{t<T} \|\mathbf{g}_t\|_{\mathbf{P}_t}^2 \le \Delta_f$. Since $\|\mathbf{g}_t\|_{\mathbf{P}_t}^2 \ge p_{\min} \|\mathbf{g}_t\|_2^2$, Eq. (21) follows. The radius bound is assumed along the iterates; the proof ensures the relative scale margin, not a uniform radius lower bound by itself. □

D.3 Unbiased stochastic gradients

Let $\mathcal{F}_t$ be the history before sampling iteration t's direction, with $\mathbf{W}_t$, $\mathbf{g}_t$, and $\mathbf{P}_t$ measurable, and write $\mathbb{E}_t = \mathbb{E}[\,\cdot \mid \mathcal{F}_t]$. For $\widetilde{\mathbf{g}}_t = \mathbf{g}_t + \boldsymbol{\zeta}_t$, assume $\mathbb{E}_t \boldsymbol{\zeta}_t = 0$ and $\mathbb{E}_t \|\boldsymbol{\zeta}_t\|_{\mathbf{P}_t}^2 \le \sigma^2$. In addition to the smoothness and lower-bound assumptions above, assume almost surely that all visited cell radii are at least $r_{\min}$, $\|\mathbf{g}_t\|_2 \le M < \infty$, and every realized step $-\eta \mathbf{P}_t \widetilde{\mathbf{g}}_t$ is admissible. Set $\overline{L}_{\mathrm{PA}} = L_{\mathrm{PA}}(M)$.

Theorem D.2 (Stochastic first-order rate)**.** *For $\mathbf{W}_{t+1} = \mathrm{Retr}_{\mathbf{W}_t}(-\eta \mathbf{P}_t \widetilde{\mathbf{g}}_t)$ under these assumptions, every constant $0 < \eta \le 1/\overline{L}_{\mathrm{PA}}$ satisfies*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{g}_t\|_2^2 \le \frac{1}{p_{\min}} \left(\frac{2\Delta_f}{\eta T} + \overline{L}_{\mathrm{PA}} \eta \sigma^2 \right). \tag{D.7}$$

For $\Delta_f > 0$, $\sigma > 0$, and $T \ge 2\overline{L}_{\mathrm{PA}} \Delta_f / \sigma^2$, the horizon-dependent choice $\eta = \sqrt{2\Delta_f / (\overline{L}_{\mathrm{PA}} \sigma^2 T)}$ gives an $\mathcal{O}(T^{-1/2})$ rate, provided the same admissibility assumption holds for that choice. More generally, for $0 < \eta < 2/\overline{L}_{\mathrm{PA}}$ and $c_\eta = \eta(1 - \eta \overline{L}_{\mathrm{PA}}/2) > 0$,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{g}_t\|_{\mathbf{P}_t}^2 \le \frac{\Delta_f / T + \overline{L}_{\mathrm{PA}} \eta^2 \sigma^2 / 2}{c_\eta}. \tag{D.8}$$

Proof. Conditional unbiasedness and measurability imply

$$\mathbb{E}_t\langle \mathbf{g}_t, \mathbf{P}_t\widetilde{\mathbf{g}}_t\rangle = \|\mathbf{g}_t\|^2_{\mathbf{P}_t}, \qquad \mathbb{E}_t\|\widetilde{\mathbf{g}}_t\|^2_{\mathbf{P}_t} \le \|\mathbf{g}_t\|^2_{\mathbf{P}_t} + \sigma^2.$$

Apply Eq. (D.4) pathwise, then condition:

$$\mathbb{E}_t f(\mathbf{W}_{t+1}) \le f(\mathbf{W}_t) - c_\eta\|\mathbf{g}_t\|^2_{\mathbf{P}_t} + \tfrac{1}{2}\overline{L}_{\mathrm{PA}}\eta^2\sigma^2.$$

Taking total expectations and summing proves Eq. (D.8). For $\eta \le 1/\overline{L}_{\mathrm{PA}}$, use $c_\eta \ge \eta/2$ and Eq. (D.1) to obtain Eq. (D.7). The stated step size gives $(2/p_{\min})\sqrt{2\Delta_f\overline{L}_{\mathrm{PA}}\sigma^2/T}$. The explicit admissibility assumption is necessary: bounded variance alone does not bound every realized step. □

D.4 General base-optimizer directions

The same finite-map inequality applies to biased or adaptive directions if the *corrected* direction remains aligned with descent. Keep the smoothness, lower-bound, fixed-multiplier, filtration, radius, and uniform true-gradient-bound assumptions above, but replace unbiasedness by the conditions below.

Theorem D.3 (Conditional extension to a base direction)**.** *Let $\mathbf{d}_t$ be a possibly random base direction. Suppose $-\eta\mathbf{P}_t\mathbf{d}_t$ is admissible almost surely and, for $c, C > 0$ and $\sigma_d^2 \ge 0$,*

$$\mathbb{E}_t\langle \mathbf{g}_t, \mathbf{P}_t\mathbf{d}_t\rangle \ge c\|\mathbf{g}_t\|_2^2, \qquad \mathbb{E}_t\|\mathbf{d}_t\|^2_{\mathbf{P}_t} \le C\|\mathbf{g}_t\|_2^2 + \sigma_d^2. \tag{D.9}$$

For $\mathbf{W}_{t+1} = \mathrm{Retr}_{\mathbf{W}_t}(-\eta\mathbf{P}_t\mathbf{d}_t)$, every $0 < \eta \le c/(\overline{L}_{\mathrm{PA}}C)$ satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{g}_t\|_2^2 \le \frac{2\Delta_f}{c\eta T} + \frac{\overline{L}_{\mathrm{PA}}\eta\sigma_d^2}{c}. \tag{D.10}$$

For $0 < \eta < 2c/(\overline{L}_{\mathrm{PA}}C)$, the sharper form is

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{g}_t\|_2^2 \le \frac{\Delta_f/T + \overline{L}_{\mathrm{PA}}\eta^2\sigma_d^2/2}{\eta c - \overline{L}_{\mathrm{PA}}\eta^2 C/2}. \tag{D.11}$$

Proof. Conditioning Eq. (D.4) gives

$$\mathbb{E}_t f(\mathbf{W}_{t+1}) \le f(\mathbf{W}_t) - (\eta c - \overline{L}_{\mathrm{PA}}\eta^2 C/2)\|\mathbf{g}_t\|_2^2 + \tfrac{1}{2}\overline{L}_{\mathrm{PA}}\eta^2\sigma_d^2.$$

Take total expectations, sum over iterations, and use $f \ge f_{\inf}$ to obtain Eq. (D.11). For $\eta \le c/(\overline{L}_{\mathrm{PA}}C)$, its denominator is at least $c\eta/2$, giving Eq. (D.10). □

The alignment condition is on $\mathbf{P}_t\mathbf{d}_t$, not merely on $\mathbf{d}_t$. A descent direction before correction need not satisfy it afterward. These are conditional frameworks for momentum or adaptive proposals, not unconditional theorems for the reported AdamW/Muon runs. The bounds concern stationarity, not improved generalization, validation loss, or wall-clock time.

E Additional Experimental Details

E.1 Experimental Setup

Models and data. We evaluate PA on two LLaMA-style models and Qwen2.5-0.5B. Their main architectural configurations are summarized in Table 2. The two LLaMA-style models use standard multi-head attention, whereas Qwen2.5-0.5B uses grouped-query attention with 14 query heads and 2 key/value heads. Qwen2.5-0.5B also differs substantially in vocabulary size, native context length, and RoPE base, providing an additional architectural setting for evaluating PA. All models are pre-trained on the C4 training split using the Hugging Face datasets library.

Optimizers. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay 0.1. Muon uses momentum 0.95 with Nesterov acceleration and five bfloat16 Newton–Schulz iterations, with

Table 2: Model configurations used in the pretraining experiments. Q and KV denote the numbers of query and key/value attention heads, respectively.

Model	Layers	d_{model}	d_{ff}	Q	KV	d_{head}	Vocab.	Max pos.	RoPE base	Params. (M)
LLaMA-130M	12	768	2,048	12	12	64	32,000	1,024	10^4	134.1
LLaMA-350M	24	1,024	2,736	16	16	64	32,000	1,024	10^4	368.0
Qwen2.5-0.5B	24	896	4,864	14	2	64	151,936	32,768	10^6	494.0

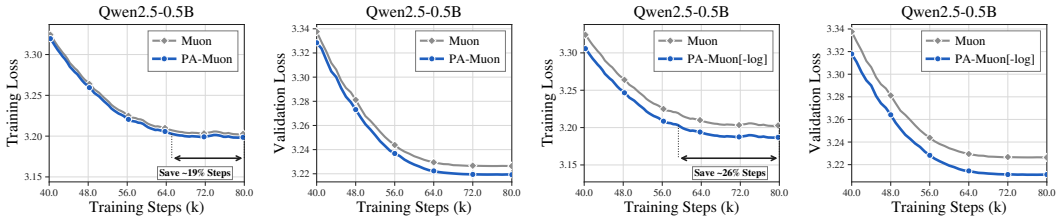


Figure 4: Training and evaluation curves for Qwen2.5-0.5B with Muon. The left two panels compare Muon with the standard PA correction, while the right two panels compare Muon with PA-Log.

$\epsilon = 10^{-7}$. Parameters outside the matrix groups handled by Muon use AdamW as the fallback optimizer. For both base optimizers, PA modifies only the RoPE-aligned query/key projection cells and leaves all remaining parameter updates unchanged.

Training protocol. All experiments use a global batch size of 256. The LLaMA-130M, LLaMA-350M, and Qwen2.5-0.5B experiments are trained for 20,000, 60,000, and 80,000 optimizer updates using 2, 4, and 4 GPUs, respectively, correspond to approximately 1.31B, 3.93B, and 5.24B training tokens. All runs use a 10% linear warmup followed by cosine learning-rate decay and bfloat16 training. Evaluation is performed periodically on the C4 validation set.

E.2 Additional Qwen2.5-0.5B Results

We further evaluate PA on Qwen2.5-0.5B using Muon as the base optimizer. We consider both the standard PA correction based on the θ_i^2 frequency profile and the PA-Log variant, and compare each of them against the same Muon baseline under the Qwen training configuration described in Appendix E.1. This experiment examines whether the RoPE-aware correction remains effective on a different model family and how the two frequency profiles behave in this setting.

Figure 4 further evaluates PA on Qwen2.5-0.5B with Muon. Both frequency profiles accelerate pretraining relative to the base optimizer. The standard PA correction reaches the corresponding Muon training loss with approximately 19% fewer optimizer updates, while PA-Log increases the reduction to approximately 26%. The validation curves exhibit the same overall trend, indicating that the faster training progress is not confined to the training objective. These results provide additional evidence that the RoPE-aware correction transfers beyond the LLaMA-style models considered in the main experiments, with the log-smoothed profile providing a stronger gain in this Qwen setting.

E.3 Learning-Rate and Hyperparameter Selection

We select learning rates and PA profile parameters using 10,000-update runs on the 130M and 350M LLaMA-style models. All sweep runs use training and evaluation seed 42 and the same learning-rate schedule as the main experiments. Candidate configurations are compared using validation loss at the final update.

Standard PA. We first tune the base optimizers and the standard PA variant with $p_i^{\text{LF}} = (1+\kappa\theta_i^2)^{-1}$. For AdamW-family methods, we sweep a scalar learning rate. For Muon-family methods, the reported learning rates are ordered pairs $(\eta_{\text{mat}}, \eta_{\text{fb}})$, where the query/key projections belong to the matrix group. At 130M, we use $\kappa = 35$ for PA and sweep the learning rate. At 350M, we additionally compare $\kappa \in \{15, 35, 70\}$. Table 3 summarizes the candidate sets and the configurations selected for the long-run experiments.

Table 3: Hyperparameter selection for the base optimizers and standard PA. Learning rates are reported in units of 10^{-3}. Muon-family learning rates are ordered pairs $(\eta_{\mathrm{mat}}, \eta_{\mathrm{fb}})$. Boldface marks the selected configuration, which attains the lowest final validation loss among the reported candidates.

Scale	Method	Candidate LR(s)	Candidate κ	Val. loss
130M	AdamW	0.6, 0.8, **1**, 1.5, 1.8	–	3.4413
130M	PA-AdamW	1, **1.5**, 1.8	**35**	3.4332
130M	Muon	$(6, 1.2)$, $(8, 1.6)$, $(\mathbf{12}, \mathbf{2.4})$, $(16, 3.2)$, $(20, 4)$	–	3.3740
130M	PA-Muon	$(20, 4)$, $(24, 4.8)$, $(\mathbf{28}, \mathbf{5.6})$, $(32, 6.4)$	**35**	3.3524
350M	AdamW	0.6, **0.8**, 1, 1.5, 1.8	–	3.3081
350M	PA-AdamW	0.6, **0.8**, 1	15, 35, **70**	3.3013
350M	Muon	$(4, 1.2)$, $(\mathbf{8}, \mathbf{2.4})$, $(10, 3)$, $(12, 3.6)$	–	3.2313
350M	PA-Muon	$(4, 1.2)$, $(\mathbf{8}, \mathbf{2.4})$, $(10, 3)$	15, 35, **70**	3.2247

PA-Log. We separately tune the learning rate for PA-Log using the same 10,000-update protocol. Since the corresponding base-optimizer sweeps are already reported in Table 3, Table 4 reports only the PA-Log configurations.

Table 4: Learning-rate selection for PA-Log. Learning rates are reported in units of 10^{-3}, with Muon-family entries written as $(\eta_{\mathrm{mat}}, \eta_{\mathrm{fb}})$. Boldface marks the selected learning rate.

Scale	Method	Candidate LR(s)	Val. loss
130M	PA-AdamW-Log	0.6, 0.8, 1, **1.5**, 1.8	3.4294
130M	PA-Muon-Log	$(8, 1.6)$, $(12, 2.4)$, $(16, 3.2)$, $(20, 4)$, $(\mathbf{24}, \mathbf{4.8})$, $(28, 5.6)$	3.3467
350M	PA-AdamW-Log	0.8, **1**, 1.5, 1.8	3.2857
350M	PA-Muon-Log	$(4, 1.2)$, $(\mathbf{8}, \mathbf{2.4})$, $(10, 3)$, $(12, 3.6)$	3.2179

For each model, optimizer, and PA profile, the long-run experiment uses the configuration selected by the corresponding sweep above.

Qwen2.5-0.5B. Because the Qwen2.5-0.5B experiments are substantially more expensive than the LLaMA-style runs, we perform a compact learning-rate check rather than an exhaustive sweep. For all Muon-family methods, we parameterize the learning rate as an ordered pair $(\eta_{\mathrm{mat}}, \eta_{\mathrm{fb}})$. Muon and standard PA-Muon are evaluated at $\{(6.5, 1.105), (8, 1.36), (10, 1.7)\} \times 10^{-3}$, while PA-Muon-Log is evaluated at $\{(6.5, 1.105), (8, 1.36)\} \times 10^{-3}$. We use $(8, 1.36) \times 10^{-3}$ as the common learning-rate pair for the reported Qwen experiments. At this setting, the final validation losses are 3.22636, 3.21921, and 3.21125 for Muon, PA-Muon, and PA-Muon-Log, respectively. This limited check is intended to identify a reasonable common operating point rather than a fully tuned optimum for each method.

Table 5: Compact learning-rate check on Qwen2.5-0.5B. Muon-family learning rates are ordered pairs $(\eta_{\mathrm{mat}}, \eta_{\mathrm{fb}})$.

Method	Checked LR pairs ($\times 10^{-3}$)	Val. loss
Muon	$(6.5, 1.105)$, $(\mathbf{8}, \mathbf{1.36})$, $(10, 1.7)$	3.22636
PA-Muon	$(6.5, 1.105)$, $(\mathbf{8}, \mathbf{1.36})$, $(10, 1.7)$	3.21921
PA-Muon-Log	$(6.5, 1.105)$, $(\mathbf{8}, \mathbf{1.36})$	3.21125

E.4 Multi-Seed Pretraining Results

We next evaluate whether the PA-Log improvements are consistent across training seeds. For each model–optimizer setting, we run the baseline and PA-Log with training seeds 42, 43, and 44, pairing runs by seed and keeping the hyperparameters selected in Appendix E.3 fixed across all three runs. The evaluation protocol is identical within each paired comparison.

For seed s, we define the paired difference as $\Delta_s = \mathcal{L}_s^{\text{PA-Log}} - \mathcal{L}_s^{\text{baseline}}$. Table 6 reports the final validation losses, the mean paired difference, and the number of seeds for which PA-Log achieves a lower loss than the corresponding baseline.

Table 6: Multi-seed pretraining results for PA-Log. Validation losses are reported as mean $\pm$ sample standard deviation over three training seeds. The paired difference is $\Delta_s = \mathcal{L}_s^{\text{PA-Log}} - \mathcal{L}_s^{\text{baseline}}$; negative values favor PA-Log. "Wins" counts the number of seeds for which PA-Log obtains a lower final validation loss.

Base optimizer	Scale	Baseline	PA-Log	Paired Δ	Wins
AdamW	130M	3.3003 ± 0.0032	$\mathbf{3.2901 \pm 0.0019}$	-0.0101 ± 0.0044	$3/3$
AdamW	350M	3.0240 ± 0.0049	$\mathbf{3.0047 \pm 0.0025}$	-0.0193 ± 0.0063	$3/3$
Muon	130M	3.2596 ± 0.0015	$\mathbf{3.2385 \pm 0.0037}$	-0.0211 ± 0.0027	$3/3$
Muon	350M	2.9970 ± 0.0021	$\mathbf{2.9811 \pm 0.0028}$	-0.0159 ± 0.0025	$3/3$

PA-Log achieves a lower final validation loss for all three seeds in each of the four model–optimizer settings. The paired improvements are therefore consistent in sign across the tested seeds, although their magnitude varies with model scale and base optimizer. Given the small number of seeds, these results are intended as a consistency check rather than a formal statistical significance analysis.

E.5 Ablation Studies

Unless otherwise specified, all ablations use the 130M LLaMA-style model on C4 with training seed 42 and evaluation seed 12,345, and are trained for 10,000 optimizer updates under a common training configuration. We keep the learning-rate schedule and the remaining training settings fixed across compared variants and report the final validation loss. Each ablation is designed to isolate a specific component of PA.

E.5.1 Finite update map and positional sensitivity

We first separate the effect of PA's finite scale–rotation update from that of its RoPE-dependent angular correction. The polar-only control uses the same finite update map as PA but sets $p_i = 1$ for every RoPE pair, thereby removing all frequency-dependent angular attenuation.

We then compare the practical low-frequency profile used by standard PA with the exact displacement-dependent positional sensitivity

$$S_i(\pi) = \mathbb{E}_{h\sim\pi}\left[4\sin^2\left(\frac{h\theta_i}{2}\right)\right], \qquad p_i^{\text{ana}} = (1+\kappa S_i(\pi))^{-1}. \tag{E.1}$$

We consider two choices of the displacement distribution π. The first uniformly weights eight representative offsets $\{1, 2, 4, 8, 16, 32, 64, 128\}$, while the second weights causal token-pair displacements according to their frequency within a sequence of length 256. These exact profiles are compared with the practical multiplier $p_i^{\text{LF}} = (1+\kappa\theta_i^2)^{-1}$ used by standard PA. As discussed in Section 4.1, the latter follows from the low-frequency approximation $S_i(\pi) \propto \theta_i^2$, with the common displacement-dependent factor absorbed into κ.

Table 7 separates the effect of the finite update map from that of the RoPE-dependent angular weighting. Replacing Cartesian addition with the finite polar update alone yields only a modest improvement, reducing the validation loss from 3.42035 to 3.41352. In contrast, incorporating the exact positional sensitivity produces substantially larger gains under both tested displacement distributions, reaching 3.38649 with eight representative offsets and 3.38472 with causal-pair weighting.

The practical low-frequency profile used by standard PA achieves a validation loss of 3.38857, retaining most of the improvement obtained by the exact profiles while avoiding the need to specify an explicit displacement distribution. These results support the use of the θ_i^2 profile as a simple architecture-only surrogate for the exact positional sensitivity.

E.5.2 Sensitivity to PA Strength and PA-Log Range

We next examine how sensitive PA is to the strength of its angular correction. For standard PA, we vary κ in the low-frequency profile while keeping the learning rate and all other settings fixed.

Table 7: **Finite update map and positional sensitivity.** The polar-only control retains the finite PA update map but removes frequency-dependent angular attenuation. The exact variants use displacement-dependent sensitivities, while standard PA uses the low-frequency θ_i^2 profile. Negative Δ_{AdamW} indicates improvement over AdamW, with $\kappa = 35$.

Variant	Angular multiplier p_i	Val. loss ↓	Δ_{AdamW}
AdamW (Cartesian)	—	3.42035	—
Polar update only	1	3.41352	-0.00683
Exact, eight offsets	$(1+\kappa S_i^{\mathrm{E}})^{-1}$	3.38649	-0.03386
Exact, causal pairs	$(1+\kappa S_i^{\mathrm{C}})^{-1}$	**3.38472**	**-0.03563**
Standard PA (θ_i^2)	$(1+\kappa\theta_i^2)^{-1}$	3.38857	-0.03178

As shown in Table 8 (a), increasing κ initially improves the validation endpoint, after which performance remains nearly flat. The lowest observed loss occurs at $\kappa = 35$, while the results for $\kappa \in \{35, 55, 75, 100\}$ differ by at most 0.00036. Thus, the benefit of standard PA persists over a reasonably broad range of correction strengths rather than depending on a sharply tuned value.

For PA-Log, we fix $\lambda_{\min} = 1$ and vary $\lambda_{\max} \in \{2, 4, 8\}$, which controls the range of angular attenuation. The corresponding minimum multipliers are $1/4$, $1/16$, and $1/64$. Table 8 (b) shows that $\lambda_{\max} = 4$ gives the lowest validation loss, while increasing the range further to $\lambda_{\max} = 8$ degrades performance. Hence, stronger attenuation is not uniformly beneficial, and the selected PA-Log profile lies within an intermediate effective range.

Table 8: **Sensitivity to PA strength and PA-Log range.** Panel (a) varies the standard PA strength κ while keeping the learning rate fixed; $0^\dagger$ reuses the polar-only result from Table 7. Panel (b) varies the PA-Log upper bound $\lambda_{\max}$ with $\lambda_{\min} = 1$. "Gap to best" is measured relative to the lowest validation loss within each panel.

(a) Standard PA strength

κ	Val. loss ↓	Gap to best
$0^\dagger$	3.41352	+0.02495
5	3.39528	+0.00671
15	3.39271	+0.00414
35	**3.38857**	0
55	3.38869	+0.00012
75	3.38893	+0.00036
100	3.38881	+0.00024

(b) PA-Log attenuation range

$\lambda_{\max}$	$\min_i p_i^{\mathrm{Log}}$	Val. loss ↓	Gap to best
2	$1/4$	3.38491	+0.00232
4	$1/16$	**3.38259**	0
8	$1/64$	3.38979	+0.00720

E.5.3 Frequency and Cell Alignment

We finally test whether PA benefits merely from attenuating angular updates, or whether the attenuation must be aligned with the underlying RoPE structure. Using the standard profile $p_i^{\mathrm{LF}} = (1 + 35\theta_i^2)^{-1}$, we construct several controls that deliberately break this alignment. Reversed and shuffled assignments apply the same set of multipliers to incorrect RoPE pairs, while a mean-matched control replaces the pair-dependent multipliers by their arithmetic mean. We also consider a random row-pairing control that applies PA to randomly paired query/key projection rows within each attention head rather than to the native RoPE-aligned cells.

Correctly aligned PA achieves the lowest validation loss among all tested controls. Reversing or shuffling the multiplier assignment degrades performance despite preserving the same set of mul-

Table 9: **Frequency and cell alignment controls.** The RoPE-aligned row uses standard PA with $p_i^{\mathrm{LF}} = (1+35\theta_i^2)^{-1}$. The remaining variants preserve or simplify the same attenuation profile while disrupting its assignment to RoPE frequencies or its underlying cell structure. Positive $\Delta_{\mathrm{aligned}}$ indicates higher validation loss than correctly aligned PA.

Variant	Cell pairing	Multiplier	Val. loss ↓	$\Delta_{\mathrm{aligned}}$
RoPE-aligned (reference)	RoPE	p_i^{LF}	**3.38857**	0
Reversed assignment	RoPE	p_{N+1-i}^{LF}	3.40498	+0.01641
Shuffled assignment	RoPE	$p_{\sigma(i)}^{\mathrm{LF}}$	3.39997	+0.01140
Mean-matched constant	RoPE	$\bar{p}^{\mathrm{LF}}$	3.39698	+0.00841
Random row pairing	Random, within head	p_i^{LF}	3.39906	+0.01049

tiplier values, while replacing the profile by its mean also performs worse than the frequency-dependent assignment. Randomizing the row pairing produces a similar degradation. Together, these results indicate that the benefit of PA depends not only on generic angular attenuation, but also on aligning that attenuation with the RoPE frequencies and their native parameter cells. These controls do not explicitly match the realized update norm across variants, and should therefore be interpreted as evidence for the importance of RoPE alignment rather than as an identification of a unique optimization mechanism.

E.6 Downstream Evaluation

We evaluate the final checkpoints after 20,000 updates for the 130M models and 60,000 updates for the 350M models. All evaluations use the zero-shot setting. ARC-E (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2020) use normalized accuracy, while SciQ (Welbl et al., 2017), WinoGrande (Sakaguchi et al., 2021), SocialIQA (Sap et al., 2019), and WSC273 (Levesque et al., 2012) use accuracy. The reported average is the unweighted mean across these seven tasks. Downstream results are not used for learning-rate or hyperparameter selection.

Table 10 reports the downstream performance of the baseline, standard PA, and PA-Log checkpoints. Across both model scales and base optimizers, the PA variants improve the average downstream score over the corresponding baselines. Standard PA achieves the highest average in three of the four model-optimizer settings, while PA-Log performs best for 350M AdamW. The gains are not uniform across individual tasks, suggesting that PA preserves and, in aggregate, modestly improves downstream transfer quality while accelerating pretraining.

Table 10: Zero-shot downstream accuracy (%) of the final pretraining checkpoints. Boldface marks the best score within each base-optimizer family, with ties at the reported precision boldfaced jointly. Avg. is the unweighted mean across the seven tasks.

Model	Optimizer	ARC-E	Hella.	PIQA	SciQ	WG	SocialIQA	WSC273	Avg.
130M	AdamW	36.32	28.82	60.88	**65.70**	50.75	35.06	54.95	47.50
	PA-AdamW	35.82	**29.24**	**61.37**	64.50	**51.62**	**35.77**	**58.61**	**48.13**
	PA-AdamW-Log	**36.36**	28.94	61.15	64.70	50.59	35.21	56.04	47.57
	Muon	37.37	28.76	61.64	65.90	51.07	33.67	54.21	47.52
	PA-Muon	**38.17**	29.51	61.97	**67.50**	51.46	34.24	**57.51**	**48.62**
	PA-Muon-Log	36.95	**29.79**	**62.02**	65.30	**52.72**	**34.85**	**57.51**	48.45
350M	AdamW	38.64	32.68	**65.13**	**70.40**	50.75	34.60	57.51	49.96
	PA-AdamW	**40.03**	33.39	64.09	68.80	51.14	32.66	**60.07**	50.03
	PA-AdamW-Log	38.47	**33.87**	64.04	69.90	**51.62**	**35.11**	57.88	**50.13**
	Muon	39.31	33.31	64.91	69.30	50.59	34.95	57.51	49.98
	PA-Muon	**40.11**	33.82	**65.51**	**70.60**	**51.70**	32.25	60.44	**50.63**
	PA-Muon-Log	39.10	**34.02**	64.80	68.70	51.07	**35.01**	**61.17**	50.55

We additionally evaluate the Qwen2.5-0.5B checkpoints trained with Muon, standard PA, and PA-Log on ten zero-shot downstream tasks. This evaluation suite extends the seven-task LLaMA

evaluation with ARC-C (Clark et al., 2018), MMLU (Hendrycks et al., 2020), and OpenBookQA (OBQA) (Mihaylov et al., 2018). As shown in Table 11, both PA variants maintain comparable performance across individual tasks while modestly improving the aggregate average over the Muon baseline. The ten-task average increases from 42.25 for Muon to 42.36 with standard PA and 42.48 with PA-Log. Together with the pretraining results in Appendix E.2, these results indicate that the faster Qwen pretraining obtained by PA does not come at an evident loss of downstream quality.

Table 11: Zero-shot downstream accuracy (%) of Qwen2.5-0.5B checkpoints trained with Muon, standard PA, and PA-Log. Boldface marks the best score in each column. Avg. is the unweighted mean across the ten tasks.

Optimizer	ARC-C	ARC-E	Hella.	MMLU	OBQA	PIQA	SciQ	WG	SocialIQA	WSC273	Avg.
Muon	24.06	38.68	34.71	22.96	26.00	66.00	**69.20**	51.62	33.54	55.68	42.25
PA-Muon	24.23	**39.98**	34.48	**23.00**	**27.60**	65.94	68.50	**51.70**	33.23	54.95	42.36
PA-Muon-Log	**24.57**	39.06	**34.83**	**23.00**	26.40	**66.16**	68.20	51.54	**33.90**	**57.14**	**42.48**