How Query–Key Initialization Shapes Gradients and Training in Transformers

Anonymous authors
Paper under double-blind review

Abstract

We study how query and key (QK) initialization shapes gradient propagation and training in Transformers. We derive an explicit gradient-norm recurrence for softmax attention and Pre-Norm residual blocks, revealing an asymmetry between large and small initialization scales. Large initialization can cause gradient explosion: the expected input-gradient gain diverges in the long-sequence limit. For normalized, weakly correlated inputs, this yields $(2\ln T)^{1/4}$ as a practical upper reference for σ_{qk}, where QK weight entries have variance σ_{qk}^2/d, d is the input dimension, and T is the sequence length. Small, nonzero initialization preserves a nonvanishing gradient-to-weight ratio, while gradient normalization in Adam and Muon permits substantial updates despite nearly uniform attention. Experiments with 100-layer attention stacks and 124M-parameter language models support the analysis, with AdamW and Muon exhibiting a broad low-loss plateau down to $\sigma_{qk} = 10^{-5}$. These findings motivate a practical recipe: favor small, nonzero QK initialization; and for broader searches, stay below the upper reference $(2\ln T)^{1/4}$.

1 Introduction

Query and key (QK) scale plays a distinctive role in Transformer optimization (Vaswani et al., 2017). It shapes attention mixing and token geometry, with consequences for gradient propagation and training. Its importance is evident in practice: excessively large QK scales can produce highly concentrated attention and accompany unstable training, motivating techniques such as QK-Norm (Henry et al., 2020), σReparam (Zhai et al., 2023), logit soft-capping (Gemma Team, 2024), and QK-Clip (Kimi Team, 2025). At the other extreme, small logits produce diffuse attention, raising concerns about excessive token averaging and weak QK gradients (Dong et al., 2021; Noci et al., 2022). These observations motivate a basic question: *how large should QK be?*

A standard theoretical starting point for addressing this question is to study forward signal propagation at initialization (Dong et al., 2021; Cowsik et al., 2025; Giorlandino & Goldt, 2026). In this setting, forward analyses have characterized how QK scale affects token geometry and attention concentration. In particular, Giorlandino & Goldt (2026)[1] identified a sequence-length-dependent transition from diffuse to concentrated attention. Understanding the optimization consequences of the QK scale, however, requires explicitly connecting it to backward gradient propagation.

To this end, we study QK weights with entries drawn from $\mathcal{N}(0, \sigma_{qk}^2/d)$, where σ_{qk} is the initialization scale and d is the input dimension. As a basis, we derive an attention gradient norm recurrence and compose it with Pre-Norm residual attention blocks. The recurrence identifies the attention statistics that govern backward gradient amplification, allowing us to analyze how the concentration regimes induced by QK scale affect gradient propagation through depth.

Our central finding is an asymmetry between two extremes. For large QK initialization, we identify a mechanism for gradient explosion: on the concentrated side of the attention transition, the expected input-gradient gain diverges in the long-sequence limit. For normalized, weakly correlated inputs, as typically encountered with randomly initialized embeddings in Pre-Norm Transformers, this yields $(2\ln T)^{1/4}$ as a practical upper reference for QK initialization, where T is the sequence length.

[1] See additional related work in Appendix A.1.

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.

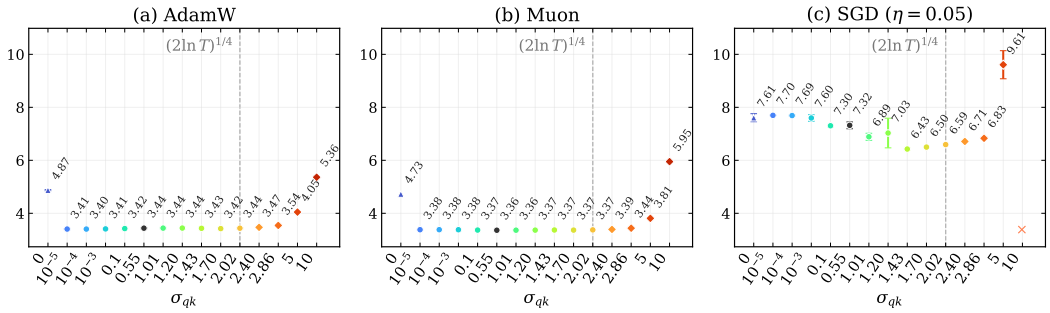

Figure 1: **Asymmetric effects of QK initialization on training.** Final validation loss of 124M-parameter LLaMA-style models with $T = 4096$. AdamW and Muon exhibit a broad low-loss plateau for small, nonzero scales, extending down to $\sigma_{qk} = 10^{-5}$, whereas sufficiently large scales incur substantial loss penalties. SGD incurs penalties at the small extreme, consistent with small gradients. Triangles mark exactly zero QK initialization, for which QK gradients vanish; red crosses indicate aborted runs. Results are averaged over three runs; most error bars are too small to be visible.

At the opposite extreme, very small, nonzero QK initialization need not impede learning, even when attention is initially nearly uniform. We show that uniform attention does not, by itself, eliminate the QK gradient signal. As $\sigma_{qk} \to 0_+$, the QK weight-gradient RMS scales linearly with σ_{qk} because of the bilinear structure of the logits. Weights and gradients therefore shrink together, leaving a nonvanishing gradient-to-weight ratio. Moreover, when its numerical stabilization term is negligible, Adam (Kingma & Ba, 2015) and Muon (Jordan et al., 2024) can produce an approximately scale-independent first-step update. Thus, small initial weights and raw gradients need not imply small parameter updates, providing a mechanism for rapid adaptation from nearly uniform attention.

Experiments with 100-layer Pre-Norm residual attention stacks support the predicted dependence of backward gradient amplification on QK scale and attention concentration across sequence lengths. We further train 124M-parameter LLaMA-style models with SGD, AdamW, and Muon over a wide range of σ_{qk}, obtaining results consistent with our initialization guidance (Figure 1). Beyond final losses, attention and weight trajectories further corroborate the proposed mechanism (Figure 5).

In summary, our explicit characterization of backward gradients in terms of QK scale provides practical guidance for Transformer pretraining. These results support small, nonzero QK initialization as a viable choice under Adam and Muon. When exploring larger initialization scales, $(2\ln T)^{1/4}$ serves as a practical upper reference under the normalized, weakly correlated input assumptions.

2 Preliminaries

In this section, we provide the necessary background to understand the main results.

Attention and initialization. In this paper, we mainly analyze the full attention, and the causal attention is discussed separately in Section 4.1. Let $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{T\times d}$ contain T input tokens as rows. With query/key dimension d_h and projection matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d\times d_h}$ and $\boldsymbol{W}_V \in \mathbb{R}^{d\times d}$, the attention map is

$$\begin{aligned} \boldsymbol{Q} = \boldsymbol{X}_{\text{in}}\boldsymbol{W}_Q, \quad \boldsymbol{K} = \boldsymbol{X}_{\text{in}}\boldsymbol{W}_K, \quad \boldsymbol{V} = \boldsymbol{X}_{\text{in}}\boldsymbol{W}_V, \\ \boldsymbol{G} = \boldsymbol{Q}\boldsymbol{K}^\top/\sqrt{d_h}, \quad \boldsymbol{A} = \text{softmax}(\boldsymbol{G}), \quad \boldsymbol{X}_{\text{out}} = \boldsymbol{A}\boldsymbol{V}, \end{aligned} \tag{1}$$

where softmax acts row-wise. Besides, all projection entries are independently initialized as

$$(W_P)_{ij} \sim \mathcal{N}(0, \sigma_P^2/d), \qquad P \in \{Q, K, V\}, \tag{2}$$

where σ_P is the standard-deviation multiplier (Glorot & Bengio, 2010; He et al., 2015; Yang et al., 2021). Our analysis allows distinct query and key scales, while we usually set $\sigma_Q = \sigma_K = \sigma_{qk}$ in experiments and practical discussions.

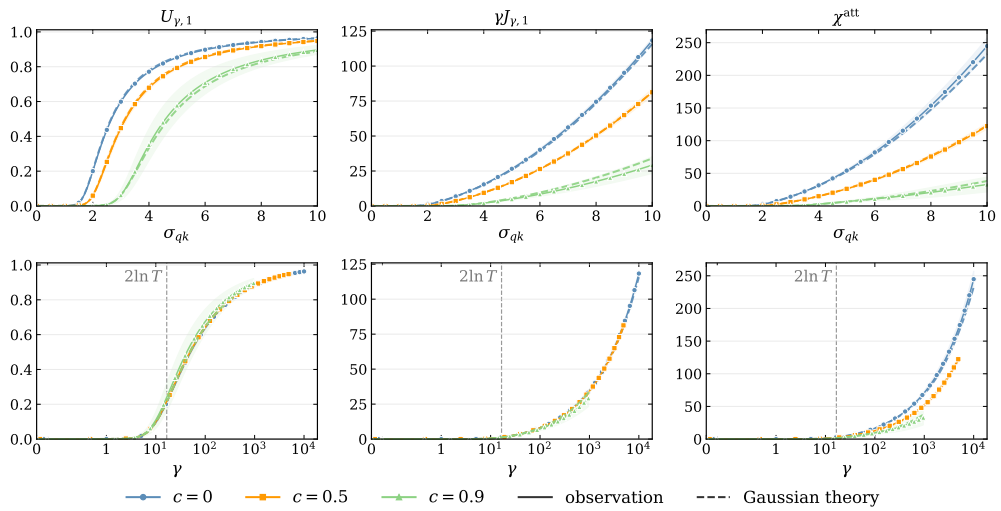

Figure 2: **Effective logit variance controls backward gain.** Columns show U, γJ, and the attention gradient multiplier $\dot{\chi}^{\mathrm{att}}$ of a self-attention layer in Eq. (10). The second row plots against γ reveals aligned transition points of $\gamma_+ = 2\ln T$, above which, backward gain rise sharply.

Writing the attention map as $\mathrm{Att}(\cdot)$, we also consider the pre-norm residual attention block

$$\boldsymbol{X} = \underline{\boldsymbol{X}} + \alpha\,\mathrm{Att}\big(\mathrm{Norm}(\underline{\boldsymbol{X}})\big), \tag{3}$$

where $\underline{\boldsymbol{X}}$ and $\boldsymbol{X}$ are the block input and output, and α is the residual-branch scale in $\{1, L^{-1/2}, L^{-1}\}$ (He et al., 2016; Yang et al., 2024; Dey et al., 2026; Zheng et al., 2026) given the depth L. Here $\mathrm{Norm}(\cdot)$ denotes token-wise RMSNorm with unit gain: $\mathrm{Norm}(\boldsymbol{x}) = \boldsymbol{x}/\sqrt{\|\boldsymbol{x}\|_2^2/d}$.

Token and gradient geometry. For $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, define $\|\boldsymbol{M}\|_{\mathrm{rms}} := \|\boldsymbol{M}\|_F/\sqrt{mn}$ and $\dot{\boldsymbol{M}} := \partial\mathcal{L}/\partial\boldsymbol{M}$, where $\mathcal{L}$ is the loss. We describe token and gradient geometry through squared RMS norms and pairwise cosine similarities. For theoretical analysis, we adopt a symmetric geometry below.

Assumption 2.1 (Token and gradient symmetry)**.** All token rows have the same squared RMS norm r, and all distinct pairs have the same correlation c. Their gradient rows also satisfy the analogous conditions with parameters $\dot{r}$ and $\dot{c}$.

Related symmetry assumptions are commonly used in mean-field analyses of signal propagation (Yang & Schoenholz, 2017; Noci et al., 2022; Cowsik et al., 2025; Giorlandino & Goldt, 2026). Independent, identically distributed Gaussian token embeddings at initialization motivate the forward approximation: for fixed T, their squared RMS norms concentrate around a common value and their pairwise cosine similarities approach zero as $d \to \infty$. We treat the corresponding symmetry of gradients and deeper representations as an additional modeling approximation.

Effective logit variance. Under Assumption 2.1 and the initialization in Eq. (2), logit rows become Gaussian as $d_h \to \infty$ at fixed T, with entry variance $\sigma_Q^2\sigma_K^2 r_{\mathrm{in}}^2$ and pairwise correlation c_{in}. Since only differences between logits affect softmax, subtracting a row's mean leaves softmax unchanged. Therefore, the attention row at limit has the same distribution as that induced by centered i.i.d. Gaussian logits with variance

$$\gamma := \sigma_Q^2\sigma_K^2 r_{\mathrm{in}}^2(1 - c_{\mathrm{in}}), \tag{4}$$

which we call the *effective logit variance*. In addition, each attention row has the same distribution as $\mathrm{softmax}(\sqrt{\gamma}\,\boldsymbol{z})$, where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_T)$. The derivation is provided in Appendix D.1. Thus, γ captures the joint effect of QK initialization, input magnitude, and token correlation on the attention.

Attention overlap statistics. To express the gradient scales, we introduce expected overlaps of attention rows and their softmax Jacobians. For $\rho \in [-1, 1]$, let $\boldsymbol{z}, \boldsymbol{z}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_T)$ be jointly Gaussian

with $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}'^{\top}] = \rho \boldsymbol{I}_T$, and write $\boldsymbol{a} = \mathrm{softmax}(\sqrt{\gamma}\,\boldsymbol{z})$ and $\boldsymbol{a}' = \mathrm{softmax}(\sqrt{\gamma}\,\boldsymbol{z}')$. We define

$$\mathrm{U}_{\gamma,\rho} \coloneqq \mathbb{E}[\langle \boldsymbol{a}, \boldsymbol{a}' \rangle], \quad \mathrm{J}_{\gamma,\rho} \coloneqq \mathbb{E}\big[\langle \boldsymbol{J}(\sqrt{\gamma}\,\boldsymbol{z}), \boldsymbol{J}(\sqrt{\gamma}\,\boldsymbol{z}') \rangle_F\big], \tag{5}$$

where $\boldsymbol{J}(\boldsymbol{u}) \coloneqq \partial \mathrm{softmax}(\boldsymbol{u})/\partial \boldsymbol{u}$ is the softmax Jacobian with respect to its logits. Intuitively, U describes expected attention concentration, and J describes its sensitivity to logit perturbations. For the attention model above, the joint Gaussian row law gives empirical counterparts correspond to their empirical average on the attention row: $\rho = 1$ for self-overlaps and $\rho = c_{\mathrm{in}}$ for overlaps between distinct query rows. Figure 2 compares these Gaussian-model predictions with empirical observations across QK scales γ and input correlations c_{in}.

Attention regimes in forward signal propagation. The self-overlap of an attention row $\boldsymbol{a}$ defines its effective support $N_{\mathrm{eff}}(\boldsymbol{a}) \coloneqq 1/\|\boldsymbol{a}\|_2^2$, which equals T for the uniform attention distribution over T keys and 1 for the attention concentrates on one particular key. At the distribution level, we summarize this concentration by the inverse expected self-overlap $1/\mathrm{U}_{\gamma,1} = 1/\mathbb{E}\|\boldsymbol{a}\|_2^2$.

For the Gaussian row model, Giorlandino & Goldt (2026) characterize a concentration transition in the long-sequence limit. Fix $\lambda = \gamma/\ln T > 0$ and let $T \to \infty$. Their result gives

$$\mathrm{U}_{\gamma,1} \longrightarrow \begin{cases} 0, & 0 < \lambda < 2, \\ 1 - \sqrt{2/\lambda}, & \lambda > 2. \end{cases} \tag{6}$$

Accordingly, $1/\mathrm{U}_{\gamma,1}$ diverges for $0 < \lambda < 2$ and remains finite for $\lambda > 2$, distinguishing the *diffuse* and *concentrated* regimes. The scale $\gamma = 2\ln T$ therefore marks the asymptotic concentration boundary. Besides, Giorlandino & Goldt (2026) states that a lower scale $\gamma = \frac{1}{2}\ln T$ marks the breakdown of Gaussian fluctuations in the finite size. Figure 2 illustrates this transition.

3 A Basis: Attention Gradients at Initialization

Forward analysis can describe the token geometry and attention state at initialization, but to understand the impact of initialization on training, it is essential to analyze the backward gradient. In this section, we analyze how QK initialization affects the backward pass.

We begin with the backward expressions for the attention map in Eq. (1). Given the attention row $\boldsymbol{A}_t$ and logits $\boldsymbol{G}_t$, we let $\boldsymbol{J}_t \coloneqq \partial \boldsymbol{A}_t/\partial \boldsymbol{G}_t = \partial \mathrm{softmax}(\boldsymbol{G}_t)/\partial \boldsymbol{G}_t$ denote the softmax Jacobian for $\boldsymbol{A}_t$, then exact differentiation gives

$$\textit{input gradient}: \quad \dot{\boldsymbol{X}}_{\mathrm{in}} = \dot{\boldsymbol{Q}}\boldsymbol{W}_Q^{\top} + \dot{\boldsymbol{K}}\boldsymbol{W}_K^{\top} + \dot{\boldsymbol{V}}\boldsymbol{W}_V^{\top}, \tag{7}$$

$$\textit{weight gradients}: \quad \dot{\boldsymbol{W}}_Q = \boldsymbol{X}_{\mathrm{in}}^{\top}\dot{\boldsymbol{Q}}, \qquad \dot{\boldsymbol{W}}_K = \boldsymbol{X}_{\mathrm{in}}^{\top}\dot{\boldsymbol{K}}, \qquad \dot{\boldsymbol{W}}_V = \boldsymbol{X}_{\mathrm{in}}^{\top}\dot{\boldsymbol{V}}, \tag{8}$$

where

$$\dot{\boldsymbol{Q}} = \frac{\dot{\boldsymbol{G}}\boldsymbol{X}_{\mathrm{in}}\boldsymbol{W}_K}{\sqrt{d_h}}, \; \dot{\boldsymbol{K}} = \frac{\dot{\boldsymbol{G}}^{\top}\boldsymbol{X}_{\mathrm{in}}\boldsymbol{W}_Q}{\sqrt{d_h}} \text{ with } \dot{\boldsymbol{G}}_t^{\top} = \boldsymbol{J}_t\boldsymbol{X}_{\mathrm{in}}\boldsymbol{W}_V\dot{\boldsymbol{X}}_{\mathrm{out},t}^{\top}, \text{ and } \dot{\boldsymbol{V}} = \boldsymbol{A}^{\top}\dot{\boldsymbol{X}}_{\mathrm{out}}. \tag{9}$$

We note that input and weight gradients play distinct roles in attention backward analysis. Input gradients propagate the learning signal backward through each layer's token representations, while weight gradients are used directly to update model parameters.

Evaluating their second moments requires accounting for dependencies between the forward and backward variables, as well as cross terms between input-gradient paths. To obtain tractable expressions, we supplement Assumption 2.1 with the following two assumptions.

Assumption 3.1 (Fresh-weight decoupling)**.** The upstream gradient is independent of the current layer's initialization. When computing local gradient second moments, reused forward weights are replaced by independent copies sampled from their original initialization distributions.

Assumption 3.2 (Path orthogonality)**.** We assume zero expected pairwise inner products among the QK/V input-gradient paths in Eq. (7), and between the skip and residual gradient paths.

The first is a usual mean-field decoupling approximation (Schoenholz et al., 2017; Yang & Schoenholz, 2017), which lets us ignore the complex dependence of reused forward weights on backward

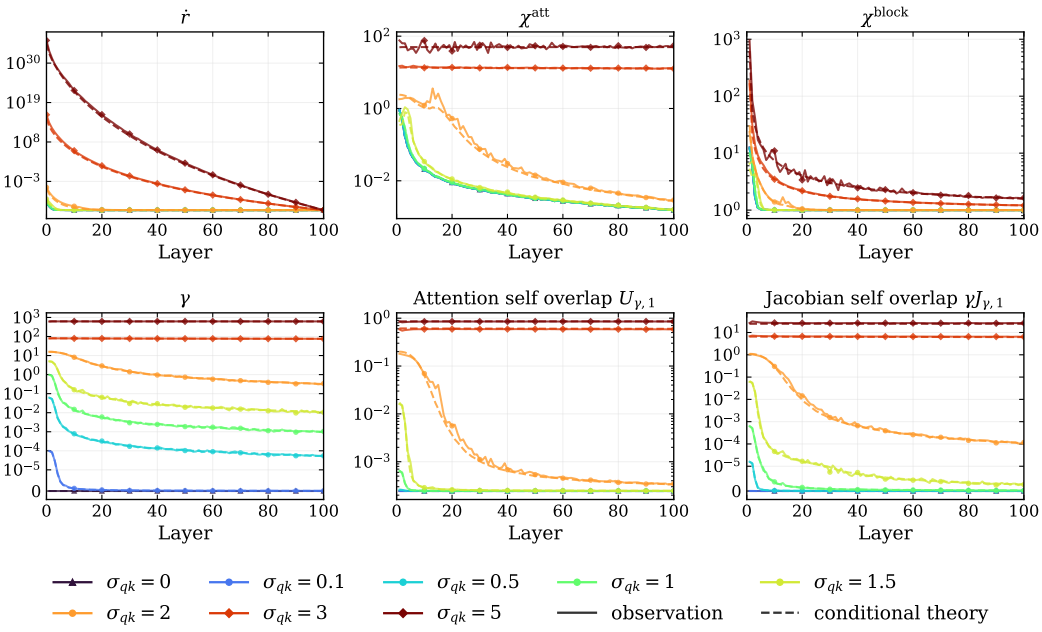

Figure 3: **Propagation in a** 100**-layer Pre-RMSNorm residual attention stack with** $\alpha = L^{-1/2}$**.** The upper row examines theoretical gradient recurrences, supporting predictions in Eqs. (10) and (13). The lower row measures forward quantities involved in backward gradient formulas.

signals. The second eliminates expected cross terms when combining input-gradient paths and when composing the residual block. We validate it empirically and present the results in Figure 8 of Appendix F.3, where the observed correlations are indeed negligible.

Let $(r_{\text{in}}, c_{\text{in}})$ describe the attention input and $(\dot{r}_{\text{out}}, \dot{c}_{\text{out}})$ its upstream gradient. We write the expected input gradient $\dot{r}_{\text{in}} := \mathbb{E}\|\dot{\boldsymbol{X}}_{\text{in}}\|^2_{\text{rms}}$ and the expected weight gradient $\dot{r}_{\boldsymbol{W}_P} := \mathbb{E}\|\dot{\boldsymbol{W}}_P\|^2_{\text{rms}}$ for $P \in \{Q, K, V\}$. The following result gives their backward recurrence.

Theorem 3.3 (Attention gradient-norm recurrence, proof in Appendix C.2)**.** *Under Assumptions 2.1, 3.1, and 3.2, let* $\gamma = \sigma_Q^2 \sigma_K^2 r_{\text{in}}^2 (1 - c_{\text{in}})$*. For the attention operation in Eq.* (1)*, take* $d, d_h \to \infty$ *while holding* d/d_h *fixed, the limiting expected squared-RMS moments satisfy*

$$\dot{r}_{\text{in}} = \sigma_V^2 \Big[\underbrace{\mathrm{U}_{\gamma,1} + (T-1)\dot{c}_{\text{out}} \mathrm{U}_{\gamma,c_{\text{in}}}}_{V \text{ path}} + \underbrace{(2 - c_{\text{in}})\gamma \mathrm{J}_{\gamma,1} + (T-1)\dot{c}_{\text{out}} \big(c_{\text{in}} \gamma \mathrm{J}_{\gamma,c_{\text{in}}}\big)}_{QK \text{ path}} \Big] \dot{r}_{\text{out}}, \tag{10}$$

The Q/K/V weight-gradients satisfy

$$\begin{pmatrix} \dot{r}_{\boldsymbol{W}_Q} \\ \dot{r}_{\boldsymbol{W}_K} \end{pmatrix} = \frac{d}{d_h} \sigma_V^2 \begin{pmatrix} \sigma_K^2 \\ \sigma_Q^2 \end{pmatrix} r_{\text{in}}^3 (1 - c_{\text{in}})^2 T \left(\mathrm{J}_{\gamma,1} + (T-1)\dot{c}_{\text{out}} c_{\text{in}} \mathrm{J}_{\gamma,c_{\text{in}}} \right) \dot{r}_{\text{out}}, \tag{11}$$

$$\dot{r}_{\boldsymbol{W}_V} = r_{\text{in}} T \Big[(1 - c_{\text{in}}) \left(\mathrm{U}_{\gamma,1} + (T-1)\dot{c}_{\text{out}} \mathrm{U}_{\gamma,c_{\text{in}}} \right) + c_{\text{in}} \big(1 + (T-1)\dot{c}_{\text{out}} \big) \Big] \dot{r}_{\text{out}}. \tag{12}$$

We denote the coefficient of $\dot{r}_{\text{out}}$ in Eq. (10) by $\dot{\chi}^{\text{att}}(r_{\text{in}}, c_{\text{in}}, \dot{c}_{\text{out}})$, the input-gradient squared-RMS multiplier. Though this paper focuses on QK initialization analysis, we preserve the V weight result here for completeness and for independent interest.

The formula connects the backward pass to the attention statistics introduced in Section 2. The V path differentiates the mixing operation $\boldsymbol{AV}$, so its second moment involves attention-row overlaps U. The QK paths pass through the softmax Jacobian and therefore involve Jacobian overlaps J. Terms proportional to $(T-1)\dot{c}_{\text{out}}$ account for correlations between distinct upstream-gradient rows.

Input and weight-gradient moments have different scale factors. The QK contribution to the input multiplier contains $\gamma \mathrm{J}$, whose asymptotic behavior will be used to show how large QK initialization

leads to gradient explosion in Section 4. Whereas the QK weight moments contain $T\mathrm{J}$. The factor T reflects token aggregation in shared parameter gradients; the input-gradient squared RMS additionally averages over T tokens. The following sections use these distinctions to analyze input-gradient amplification at large QK scales and parameter learning near zero initialization.

Composition with a Pre-Norm residual block. For the block in Eq. (3), let $(\underline{r}, \underline{c})$ denote the block-input geometry, with $\underline{r} > 0$, and $(\dot{r}, \dot{c})$ the block-output gradient geometry. RMSNorm maps the input geometry to $(1, \underline{c})$, while the residual scale makes the attention-output gradient $\dot{\boldsymbol{X}}_{\mathrm{out}} = \alpha \dot{\boldsymbol{X}}$. Combining Theorem 3.3 with the RMSNorm Jacobian and skip–branch orthogonality gives

$$\dot{\underline{r}} = \dot{\chi}^{\mathrm{block}} \dot{r}, \qquad \dot{\chi}^{\mathrm{block}} = 1 + \frac{\alpha^2}{\underline{r}} \dot{\chi}^{\mathrm{att}}(1, \underline{c}, \dot{c}). \tag{13}$$

The additive 1 comes from the skip path, α^2 from the residual scale, and $1/\underline{r}$ from the leading squared-norm multiplier of the RMSNorm Jacobian. Thus, QK initialization determines the attention gain, while $\alpha^2/\underline{r}$ scales its contribution to the block multiplier. We defer the derivation to Appendix C.3. Figure 3 evaluates these conditional recurrences in a 100-layer Pre-RMSNorm residual attention stack. The predictions capture the dependence of both local multipliers and accumulated gradient amplification on the QK scale.

4 Transition in Backward: Explosion at Large Initialization

The QK self contribution to the input-gradient multiplier in Eq. (10) is governed by $\gamma \mathrm{J}_{\gamma,1}$. We show that this factor undergoes a transition at the same boundary $\gamma = 2 \ln T$ as forward attention concentration (Giorlandino & Goldt, 2026), connecting the forward regimes to backward amplification. We first take the Gaussian limit and then study the long-sequence asymptotics.

Theorem 4.1 (Attention statistics in two regimes)**.** *Let $\lambda := \gamma / \ln T > 0$ be fixed as $T \to \infty$. Then*

$$0 < \lambda < 2: \quad \gamma \mathrm{J}_{\gamma,1} \to 0; \qquad \lambda > 2: \quad \frac{\gamma \mathrm{J}_{\gamma,1}}{\ln T} \to \frac{\sqrt{2\lambda} - 2}{3}. \tag{14}$$

To translate this result into a statement about the full attention multiplier, retain the assumptions of Theorem 3.3 and suppose $c_{\mathrm{in}} \in [0, 1)$, $\dot{c}_{\mathrm{out}} \geq 0$, and $\sigma_V > 0$ is fixed. All contributions to Eq. (10) are then nonnegative. Keeping only the QK self term gives

$$\dot{\chi}^{\mathrm{att}} \geq \sigma_V^2 (2 - c_{\mathrm{in}}) \gamma \mathrm{J}_{\gamma,1}. \tag{15}$$

Consequently, for fixed $\lambda > 2$, $\dot{\chi}^{\mathrm{att}} = \Omega(\ln T)$: the input-gradient second-moment multiplier diverges in the concentrated regime. For a Pre-Norm residual block, its contribution to the block multiplier is further scaled by $\alpha^2/\underline{r}$, as given by Eq. (13). We leave the derivation to Appendix D.2.

4.1 Predicting a warning upper reference

For Pre-RMSNorm attention with equal QK scales, Eq. (4) gives $\gamma = \sigma_{qk}^4 (1 - \underline{c})$. The asymptotic boundary $\gamma_+ = 2 \ln T$ therefore corresponds to $\sigma_{qk,+}(\underline{c}) = (2 \ln T / (1 - \underline{c}))^{1/4}$. This reference depends on token correlation, which varies across inputs and layers. The good news is that the indicator induced from the first layer does provide a useful diagnostic of attention behavior across depth. Under the forward moment closure in Proposition E.1, a common QK scale across layers yields non-decreasing token correlation $\underline{c}$ and non-increasing effective logits γ, thus the corresponding requirement $\sigma_{qk,+}(\underline{c})$ is looser than $\sigma_{qk,+}$. This persistent non-increasing γ over depth is validated by experiments in Figure 3. When random embedding initialization gives approximately uncorrelated first-layer inputs as in Transformers, it simplifies to

$$\sigma_{qk,+} \approx (2 \ln T)^{1/4}. \tag{16}$$

We use this scale as a practical upper warning reference for QK initialization. Figure 6 illustrates its weak logarithmic dependence on sequence length. Table 1 compares it with common initialization settings in our standard-deviation multiplier convention. Both listed settings lie below their corresponding upper references.

Extension to causal attention. For causal attention, query t has t visible keys. The row-wise statistics therefore use t in place of T, and the concentration boundary becomes $\gamma_{+,t} = 2\ln t$. These row-dependent references can be summarized by their quantiles. For fixed $p \in (0, 1]$, let $\gamma_+^{(p)}$ denote the p-quantile of the boundaries over rows $t = 2, \ldots, T$. With a common effective variance across rows and the same normalized, weakly correlated input approximation, corresponding references are

$$\gamma_+^{(p)} \approx 2\ln(pT), \qquad \sigma_{qk,+}^{(p)} \approx [2\ln(pT)]^{1/4}. \tag{17}$$

For fixed p, $\ln(pT) = \ln T + O(1)$, so these references have the same leading-order scaling as full attention. At $T = 4096$, for example, the median QK reference ($p = 0.5$) is approximately 1.98, compared with 2.02 for full attention, a decrease of about 2.2%. We therefore retain $(2\ln T)^{1/4}$ as a practical warning reference for both full and causal attention. This extension concerns row-wise concentration; a full causal backward recurrence additionally requires joint statistics of queries with different visible prefixes.

5 Rapid Learning in Training at Small Initialization

The preceding section connects large QK initialization to input-gradient amplification. We now examine QK parameter learning at small, nonzero initialization. As attention approaches uniformity, there are two distinct concerns: (1) token rank collapse, where $\underline{c} = 1$; and (2) diffuse attention, whose input-gradient terms can become small at long sequence lengths, such as $\mathrm{U}_{\gamma,1}, \gamma\mathrm{J}_{\gamma,1} \to 0$ in Theorem 4.1.

For the first, prior works have clearly illustrated that rank collapse indeed induces zero QK weight gradients, and adding skip connections can fix it by preserving token differences(Noci et al., 2022; Giorlandino & Goldt, 2026). For the second, we will show that uniform attention does not, by itself, eliminate QK weight gradients, even in the uniform-attention limit. The weight-gradient result further reveals a linear dependence on σ_{qk}, yielding a nonvanishing gradient-to-weight ratio. Under common optimizers like AdamW and Muon, small yet nonzero QK weights can even adapt rapidly.

5.1 Uniform attention does not vanish gradients

We use the conditional gradient second moments of Theorem 3.3. At uniform attention ($\gamma = 0$), every row equals $\mathbf{1}^\top/T$, so $\boldsymbol{J}_t = \frac{1}{T}\left(\boldsymbol{I}_T - \frac{\mathbf{1}\mathbf{1}^\top}{T}\right)$, $\mathrm{J}_{0,\rho} = \frac{T-1}{T^2}$ for any ρ. Substituting this into the QK weight-gradient formulas gives

$$\dot{r}_{\boldsymbol{W}_Q}\big|_{\gamma=0} = C^2\sigma_K^2, \qquad \dot{r}_{\boldsymbol{W}_K}\big|_{\gamma=0} = C^2\sigma_Q^2, \tag{18}$$

where the common coefficient $C \geq 0$ is defined by

$$C^2 \coloneqq \frac{d}{d_h}\sigma_V^2 r_{\mathrm{in}}^3(1 - c_{\mathrm{in}})^2\frac{T-1}{T}\big[1 + (T-1)c_{\mathrm{in}}\dot{c}_{\mathrm{out}}\big]\dot{r}_{\mathrm{out}}. \tag{19}$$

For $T \geq 2$, $0 \leq c_{\mathrm{in}} < 1$, and $\dot{c}_{\mathrm{out}} \geq 0$, we have $C > 0$. We thus see that uniformity alone does NOT eliminate the gradient signal. By contrast, token collapse is a geometric degeneracy: $c_{\mathrm{in}} = 1$ forces $C = 0$. Although the attention statistic $\mathrm{J}_{0,\rho}$ decays as $1/T$, the QK weight gradient aggregates contributions across tokens and contains an additional factor T. Uniformity therefore leaves a positive gradient coefficient; the remaining QK scale comes from the opposite projection in the bilinear logits. For example, $\sigma_Q = 0$ and $\sigma_K > 0$ give exactly uniform attention with a positive Q-gradient second moment. If both scales are zero, both QK gradients are exactly zero.

5.2 Training from Small QK Initialization

The uniform-attention coefficient determines the leading gradient scale at small, nonzero initialization.

Proposition 5.1 (QK weight-gradient norms at small initialization)**.** *Under Assumptions 2.1 and 3.1, work in the Gaussian width limit with fixed T, d/d_h, σ_V, and input and upstream-gradient geometry satisfying $C > 0$ in Eq.* (19)*. As $\sigma_Q = \sigma_K = \sigma_{qk} \to 0^+$,*

$$\sqrt{\dot{r}_{\boldsymbol{W}_Q}} = \sqrt{\dot{r}_{\boldsymbol{W}_K}} = C\sigma_{qk} + o(\sigma_{qk}). \tag{20}$$

This follows from $\gamma \to 0$ and continuity of $\mathrm{J}_{\gamma,\rho}$ at zero. The complete proof is in Appendix C.2.3.

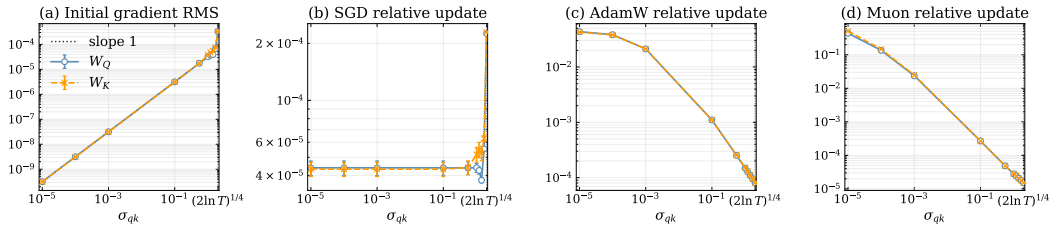

Figure 4: **Initial QK weight gradients and updates at small initialization.** The left panel compares gradient RMS to the predicted slope-one scaling. The remaining panels show relative updates: SGD is predicted to be scale-independent, whereas AdamW and Muon should scale as σ_{qk}^{-1} between the numerical floor and upper reference.

SGD. Let $\boldsymbol{W}$ denote either QK matrix and $\eta > 0$ a fixed learning rate. Its initialization RMS is $\|\boldsymbol{W}\|_{\mathrm{rms}} = \sigma_{qk}/\sqrt{d}$. Since weights and gradients shrink at the same rate, their RMS ratio does not vanish. For the first plain-SGD step,

$$\begin{aligned} \text{absolute update:} \quad & \|\Delta \boldsymbol{W}\|_{\mathrm{rms}} = \eta C \sigma_{qk} + o(\sigma_{qk}), \\ \text{relative RMS scale:} \quad & \frac{\|\Delta \boldsymbol{W}\|_{\mathrm{rms}}}{\|\boldsymbol{W}\|_{\mathrm{rms}}} = \eta C \sqrt{d} + o(1). \end{aligned} \tag{21}$$

Thus SGD preserves relative motion, while its absolute first step decreases linearly with the initialization scale.

Adam. With zero initial moments and bias correction, Adam's first gradient update (Kingma & Ba, 2015), before any decoupled weight decay, is $\Delta W_{ij} = -\eta \frac{\dot{W}_{ij}}{|\dot{W}_{ij}|+\epsilon}$, where $\epsilon > 0$ is the numerical stabilizer. When most coordinates satisfy $|\dot{W}_{ij}| \gg \epsilon$,

$$\text{absolute update:} \quad \|\Delta \boldsymbol{W}\|_{\mathrm{rms}} \approx \eta, \qquad \text{relative update:} \quad \frac{\|\Delta \boldsymbol{W}\|_{\mathrm{rms}}}{\|\boldsymbol{W}\|_{\mathrm{rms}}} \approx \frac{\eta\sqrt{d}}{\sigma_{qk}}. \tag{22}$$

Small gradients can therefore produce substantial absolute updates and larger relative steps for smaller initial weights. Once ϵ dominates, the predicted update RMS instead becomes $\eta C \sigma_{qk}/\epsilon$, and the relative RMS scale saturates at $\eta C\sqrt{d}/\epsilon$.

Muon. Muon (Jordan et al., 2024) removes gradient magnitude by orthogonalization. For the idealized first Muon-Kimi update (Liu et al., 2025), write $\dot{\boldsymbol{W}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ for the compact SVD retaining its nonzero singular values. Then

$$\Delta \boldsymbol{W} = -0.2\eta\sqrt{\max(d, d_h)}\,\boldsymbol{U}\boldsymbol{V}^\top, \qquad \|\Delta \boldsymbol{W}\|_{\mathrm{rms}} = 0.2\eta\sqrt{\frac{\mathrm{rank}(\dot{\boldsymbol{W}})}{\min(d, d_h)}}. \tag{23}$$

At fixed rank, the absolute step is independent of gradient magnitude. The full-rank case gives update RMS 0.2η and relative scale $0.2\eta\sqrt{d}/\sigma_{qk}$; the rank factor remains for lower-rank gradients.

Figure 4 compares the Q and K weight gradient RMS with the predicted linear scaling, and tests the relative updates under SGD, AdamW, and Muon. Over the interior of the small-scale range, the nearly constant SGD relative update and the decreasing AdamW and Muon curves agree with the predicted optimizer-dependent scaling. At the left edge, AdamW's numerical ϵ and Muon's finite-step numerical orthogonalization cap the inverse-scale growth of the relative update. At the right edge, σ_{qk} approaches the upper reference, so attention is no longer in the asymptotically small, near-uniform regime on which these predictions rely.

Practical implication. These first-step mechanisms support choosing *small, nonzero QK initialization*: it retains a gradient signal, while Adam and Muon can produce substantial parameter motion under the conditions above. The training trajectories in Figure 9 and 10 of Appendix F.4 validate that this motion leads to escape from near-uniform attention and improved loss over training.

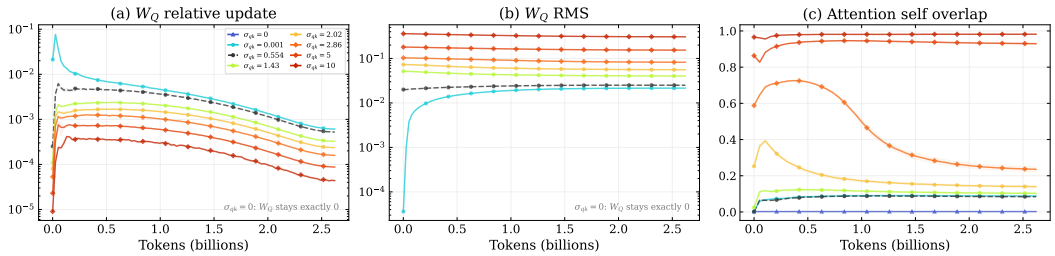

Figure 5: **Small nonzero QK initialization adapts rapidly under AdamW.** Small nonzero weights receive large relative updates and move attention rapidly away from the initial near-uniform state. Large weights move slowly relative to their scale and retain concentrated attention. $\sigma_{qk} = 0$ serves as a controlled setting where QK has zero gradients. Curves and shaded bands show the mean and one standard deviation across three seeds.

6 Training Experiments

We train a 124M-parameter LLaMA-style Transformer on FineWeb, varying only the QK initialization scale across paired runs. The context length is set to $T = 4096$. AdamW and Muon follow the standard training recipe Karpathy (2022) with warmup, gradient clipping, and cosine learning rate schedule, while plain SGD serves as a raw-gradient control. We compare QK scales under a fixed training configuration. $\sigma_{qk} = 0.554$ is regarded as a baseline by default, as it corresponds to the practice in GPT-2 or nanoGPT, where $\sigma_{qk}/d = 0.02$, as summarized in Table 1. Figure 1 shows the final validation loss; Figure 5 in the main text and Figures 9–11 in the appendix show training details; Appendix F.4 gives the complete configurations.

Large QK initialization. At $\sigma_{qk} = 5$ or 10, the initial pre-clipping QK gradients are much larger than at moderate scales. Plain SGD exposes the raw-gradient mechanism: without normalization or clipping, the large gradients become large absolute steps, causing aborted runs (Figure 1 and Figure 12). Under AdamW and Muon, however, the actual QK updates are small relative to the large initial weights, and therefore learn slowly, with attention remaining highly concentrated for a long time (Figure 5). As a result, large initialization retains substantial loss penalties (Figure 1). The degradation occurs consistently on the right side of the predicted upper warning reference.

Small, nonzero QK initialization. At near-zero initialization, the raw QK gradients are small, but AdamW and Muon produce much larger relative updates (Figure 1 and Figure 12). The QK weights rapidly leave their near-zero state, and the attention overlap moves away from their near-uniform initial values (Figure 5). Consequently, both optimizers exhibit a broad low-loss plateau down to the smallest tested positive scale, $\sigma_{qk} = 10^{-5}$. Plain SGD's absolute update shrinks with the raw gradient, so near-zero QK weights and attention change little in early training.

The zero point and practical implication. Exactly $\sigma_{qk} = 0$ is problematic (Figure 1): both QK gradients are zero, so the QK weights remain frozen under all three optimizers. It should not be identified with the limit $\sigma_{qk} \to 0^+$. Within the tested AdamW and Muon recipes, small but nonzero QK initialization is therefore a sound practical choice: Their scale-normalizing updates enable rapid QK learning despite the small initial weight and gradient magnitudes.

7 Conclusion and Discussion

By explicitly characterizing the dependence of backward gradients on QK scale and validating our analysis through a range of experiments, we establish a practical recipe for QK initialization in Transformers: favor small, nonzero scales under Adam and Muon and, when exploring larger scales, use $(2 \ln T)^{1/4}$ as a practical upper reference under normalized, weakly correlated inputs.

An important direction is to understand how QK initialization interacts with methods that control query and key norms during training Henry et al. (2020); Kimi Team (2025). From a practical perspective, extending this investigation to sparse and linear attention is also valuable: their different attention operators may alter initialization sensitivity and call for different scale-selection principles.

AI USE STATEMENT

During project exploration and development, we used AI tools to discuss research ideas, identify relevant works, assist in checking the main proofs, and support coding. We take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work studies Transformer initialization through theoretical analysis and experiments using publicly available code and data. We do not anticipate ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

We provide explicit assumptions and detailed proofs in the appendix for our theoretical results. Our experiments use the standard open-source codebase and the dataset. We will make our code publicly available after publication.

REFERENCES

Sergey Alekseev. Subcritical signal propagation at initialization in normalization-free transformers. *arXiv preprint arXiv:2604.11890*, 2026.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *International Conference on Learning Representations*, volume 2026, pp. 86454–86478, 2026.

Aditya Cowsik, Tamra Nebabu, Xiaoliang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *Physical Review E*, 112(5):055301, 2025.

Nolan Dey, Bin Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers. *Advances in Neural Information Processing Systems*, 38:137707–137739, 2026.

Emily Dinan, Sho Yaida, and Susan Zhang. Effective theory of transformers at initialization. *arXiv preprint arXiv:2304.02034*, 2023.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803, 2021.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: A unified theory of signal propagation at initialisation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=utSqpxQHXq.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020. doi: 10.18653/v1/2020.findings-emnlp.379. URL https://aclanthology.org/2020.findings-emnlp.379/.

Hugging Face. LLaMA configuration and initialization in Transformers v4.44.2. Source code, configuration_llama.py and modeling_llama.py, 2024. URL https://github.com/huggingface/transformers/tree/v4.44.2/src/transformers/models/llama.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Andrej Karpathy. nanoGPT. Source code, 2022. URL https://github.com/karpathy/nanoGPT/blob/master/model.py.

Akhil Kedia, Mohd Abbas Zaidi, Sushil Khyalia, Jungho Jung, Harshith Goka, and Haejun Lee. Transformers get stable: An end-to-end signal propagation theory for language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23449–23531, 2024.

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. URL https://arxiv.org/abs/2507.20534.

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

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*, 2025.

Cécile Monthus and Thomas Garel. Critical weight statistics of the random energy model and of the directed polymer on the cayley tree. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 75(5):051119, 2007.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36:54250–54281, 2023.

OpenAI. GPT-2: Released model source code. Source code, model.py, 2019. URL https://github.com/openai/gpt-2/blob/master/src/model.py.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Jim Pitman and Marc Yor. The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997. doi: 10.1214/aop/1024404422. URL https://www.stat.berkeley.edu/~pitman/433.pdf.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding, 2021. URL https://arxiv.org/abs/2104.09864.

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.

Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers. *Advances in Neural Information Processing Systems*, 37: 14774–14809, 2024.

Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. *Advances in neural information processing systems*, 30, 2017.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite depth neural networks. In *International Conference on Learning Representations*, volume 2024, pp. 55099–55150, 2024.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803, 2023. URL `https://proceedings.mlr.press/v202/zhai23a.html`.

Chenyu Zheng, Rongzhen Wang, Xinyu Zhang, and Chongxuan Li. Spectral condition for μp under width-depth scaling. *CoRR*, abs/2603.00541, 2026.

A Additional Discussion

A.1 Additional Related Works

Forward propagation and token geometry. Repeated attention mixing can reduce token diversity, with the outcome depending on the architecture. Dong et al. (2021) establish rank-collapse bounds for pure attention networks, while Wu et al. (2024) characterize how masking and normalization alter collapse and permit non-collapsing configurations. Noci et al. (2023) modify the attention operator and derive a forward covariance SDE in a joint depth-and-width limit. These works primarily characterize representation geometry and its evolution with depth.

Sequence length and attention concentration. Sequence length changes the logit scale required to maintain concentrated attention. Giorlandino & Goldt (2026) identify a diffuse-to-concentrated attention transition in forward under random QK initialization in the long-sequence limit and utilize the forward token correlation as an indicator for trainability. We utilize this important result and identify the transition in backward to connect based on our derived attention gradient formulas. Chen et al. (2026) establish length-dependent transitions in token geometry and input-Jacobian norms for identity Q/K/V projections. We build on the Gaussian concentration transition to show that expected input-gradient gain diverges in the concentrated regime. Our near-zero initialization analysis addresses a separate limit.

Backward propagation. Noci et al. (2022) impose uniform attention but retain its softmax derivative. Their Theorem 3.2 identifies identical tokens as a cause of vanishing QK loss gradients. Cowsik et al. (2025) and Alekseev (2026) also conduct backward analysis on uniform attention. These resulting backward results therefore do not describe its dependence on QK scale away from collapse. Chen et al. (2026) study the Jacobian transition with identity QK/V projections.

Specifically, Dinan et al. (2023) derive general second-moment expressions for nonuniform attention and use them to determine width scalings of initialization and learning rates. In these expressions, the dependence on QK initialization remains implicit in softmax derivatives. Our technical contribution is a structural reduction of these contractions: by exploiting softmax invariances and symmetric token geometry, we express the input and parameter gradient moments through two shared attention statistics, $\mathrm{U}_{\gamma,\rho}$ and $\mathrm{J}_{\gamma,\rho}$, with explicit geometric coefficients. This representation separates the mechanisms of backward amplification and geometric degeneration, and makes their dependence on initialization scale analytically accessible. Our result is compatible with their general framework.

Kedia et al. (2024) derive signal propagation through more complete Transformer architectures. While their attention backward calculation in Appendix A.8 nevertheless treats the softmax output as the constant $\mathbf{1}\mathbf{1}^\top/T$ at small QK initialization. It consequently retains only the value path. After removing dropout and the output projection and translating the weight-variance convention, this gives $\dot{\chi}^{\mathrm{att}} = \sigma_V^2 \frac{1+(T-1)\dot{c}_{\mathrm{out}}}{T}$, which our recurrence recovers at $\gamma = 0$. Their separate softmax analysis in Appendix A.7 retains random attention weights, but approximates their moments by a lognormal calculation and replaces $\mathrm{Diag}(\boldsymbol{a}) - \boldsymbol{a}\boldsymbol{a}^\top$ with $\mathrm{Diag}(\boldsymbol{a})$. In our notation, its diffuse-softmax estimate gives $\mathbb{E}\|\boldsymbol{a}\|_2^2 \approx e^\gamma/T$. This motivates their large-QK instability discussion in Section 3.3. While, since every probability vector satisfies $\|\boldsymbol{a}\|_2^2 \leq 1$, this estimate can not apply to $\gamma/\ln T > 1$, and hence cannot resolve the concentrated regime $\gamma/\ln T > 2$ studied here.

Giorlandino & Goldt (2026) also retain softmax attention and analyze QK weight gradient norms with respect to token cosine similarity and attention concentration. Their result (Section 3.3, Result 2) retains an unspecified constant. Our conditional second-moment formula instead makes the uniform-attention coefficient explicit and separates it from token collapse.

B Anatomy of Our Attention Gradient Reductions

Our core analysis is how attention input and parameter gradients depend on QK initialization, conditional on the input-token and upstream-gradient geometry. The central technical step is to reduce the coupled second moments of attention backpropagation to the two statistics $\mathrm{U}_{\gamma,\rho}$ and $\mathrm{J}_{\gamma,\rho}$. This section explains our core reductions.

First, eliminate the common token direction. Let $\boldsymbol{g}_t = \boldsymbol{G}_{t,:}^\top$, $\boldsymbol{a}_t = \boldsymbol{A}_{t,:}^\top$, and $\boldsymbol{J}_t = \boldsymbol{J}(\boldsymbol{g}_t)$. The column vector corresponding to row t of the logit gradient satisfies the exact chain-rule identity

$$\dot{\boldsymbol{g}}_t = \boldsymbol{J}_t \boldsymbol{X}_{\mathrm{in}} \boldsymbol{W}_V \dot{\boldsymbol{X}}_{\mathrm{out},t}^\top. \tag{24}$$

The key simplification is obtained before expanding individual attention entries. Since $\boldsymbol{J}_t \mathbf{1} = 0$, token symmetry gives

$$\boldsymbol{X}_{\mathrm{in}} \boldsymbol{X}_{\mathrm{in}}^\top = d r_{\mathrm{in}} \left[(1 - c_{\mathrm{in}}) \boldsymbol{I}_T + c_{\mathrm{in}} \mathbf{1}\mathbf{1}^\top \right], \tag{25}$$

$$\boldsymbol{J}_t \boldsymbol{X}_{\mathrm{in}} \boldsymbol{X}_{\mathrm{in}}^\top \boldsymbol{J}_u = d r_{\mathrm{in}} (1 - c_{\mathrm{in}}) \boldsymbol{J}_t \boldsymbol{J}_u. \tag{26}$$

Averaging over $\boldsymbol{W}_V$, conditional on the input, upstream gradient, and softmax Jacobians, therefore yields

$$\mathbb{E}_{W_V}[\dot{\boldsymbol{g}}_t \dot{\boldsymbol{g}}_u^\top] = \sigma_V^2 r_{\mathrm{in}} (1 - c_{\mathrm{in}}) \left\langle \dot{\boldsymbol{X}}_{\mathrm{out},t}, \dot{\boldsymbol{X}}_{\mathrm{out},u} \right\rangle \boldsymbol{J}_t \boldsymbol{J}_u. \tag{27}$$

This identity connects the gradient second moment directly to softmax-Jacobian overlaps. It also exposes the geometric source of collapse: when $c_{\mathrm{in}} = 1$, there is no token contrast for the QK paths to act on.

Second, condense the QK initialization effects into one condensed quantity, the effective logit scale. The same common direction can be removed from the forward logits. Define

$$\boldsymbol{C}_T = \boldsymbol{I}_T - \frac{1}{T} \mathbf{1}\mathbf{1}^\top, \qquad \rho_{tu} = c_{\mathrm{in}} + (1 - c_{\mathrm{in}}) \delta_{tu}.$$

Softmax and its Jacobian are unchanged by adding a constant to all entries of a row. In the Gaussian logit limit,

$$\mathrm{Cov}(\boldsymbol{C}_T \boldsymbol{g}_t, \boldsymbol{C}_T \boldsymbol{g}_u) = \gamma \rho_{tu} \boldsymbol{C}_T, \qquad \gamma = \sigma_Q^2 \sigma_K^2 r_{\mathrm{in}}^2 (1 - c_{\mathrm{in}}). \tag{28}$$

The centered rows can thus be represented by projected standard Gaussian vectors with correlation ρ_{tu}. Applying shift invariance again removes the projection inside softmax, giving

$$\mathbb{E}\langle \boldsymbol{a}_t, \boldsymbol{a}_u \rangle = \mathrm{U}_{\gamma, \rho_{tu}}, \qquad \mathbb{E}\langle \boldsymbol{J}_t, \boldsymbol{J}_u \rangle_F = \mathrm{J}_{\gamma, \rho_{tu}}. \tag{29}$$

The common token component is therefore eliminated twice: through the zero-sum property of the backward Jacobian and through the shift invariance of the forward softmax. Together, these reductions isolate the effective scale γ without approximating attention as uniform.

Next, when encountering cross terms in the gradient second moment, we retain and evaluate the row-pair contractions. There are T same-row pairs and $T(T-1)$ distinct-row pairs. Their upstream inner products are respectively $d\dot{r}_{\mathrm{out}}$ and $d\dot{r}_{\mathrm{out}} \dot{c}_{\mathrm{out}}$. Consequently, the value path depends on

$$\mathrm{U}_{\gamma,1} + (T-1) \dot{c}_{\mathrm{out}} \mathrm{U}_{\gamma, c_{\mathrm{in}}},$$

whereas the key path and shared QK parameter gradients contain

$$\mathrm{J}_{\gamma,1} + (T-1) c_{\mathrm{in}} \dot{c}_{\mathrm{out}} \mathrm{J}_{\gamma, c_{\mathrm{in}}}.$$

The query input path is local to its query row and contributes the additional factor $(1 - c_{\mathrm{in}}) \gamma \mathrm{J}_{\gamma,1}$. Applying the stated fresh-weight contractions and adding the input paths under path orthogonality gives Eq. (10).

The same logit-gradient covariance also gives parameter moments at general γ. For example, defining $\dot{r}_{\boldsymbol{W}_Q} = \mathbb{E}\|\dot{\boldsymbol{W}}_Q\|_F^2 / (d d_h)$, we obtain

$$\begin{aligned} \dot{r}_{\boldsymbol{W}_Q} = \frac{d}{d_h} \sigma_K^2 \sigma_V^2 r_{\mathrm{in}}^3 (1 - c_{\mathrm{in}})^2 \dot{r}_{\mathrm{out}} \, T \\ \times \left[\mathrm{J}_{\gamma,1} + (T-1) c_{\mathrm{in}} \dot{c}_{\mathrm{out}} \mathrm{J}_{\gamma, c_{\mathrm{in}}} \right]. \end{aligned} \tag{30}$$

The key formula exchanges σ_K^2 and σ_Q^2. These are general-scale formulas; the near-zero laws follow by evaluating their attention statistics at $\gamma = 0$.

C Proofs for Gradients at Initialization

We prove Theorem 3.3 and then derive the Pre-Norm residual recurrence in Eq. (13). The argument proceeds from exact gradient expressions to Gaussian contractions, then identifies the joint logit law and its softmax expectations. We evaluate the V, Q, and K input paths in that order, extend the same contractions to the weight gradients, and finally compose attention with RMSNorm and the residual connection.

Throughout the attention calculation, abbreviate

$$\boldsymbol{X} \coloneqq \boldsymbol{X}_{\mathrm{in}}, \qquad \boldsymbol{Y} \coloneqq \dot{\boldsymbol{X}}_{\mathrm{out}}, \qquad r \coloneqq r_{\mathrm{in}}, \quad c \coloneqq c_{\mathrm{in}}, \quad \dot{r} \coloneqq \dot{r}_{\mathrm{out}}, \quad \dot{c} \coloneqq \dot{c}_{\mathrm{out}}. \tag{31}$$

We retain row-vector notation throughout. Assumption 2.1 is equivalently

$$\begin{aligned} \boldsymbol{X}\boldsymbol{X}^\top &= dr\big[(1-c)\boldsymbol{I}_T + c\mathbf{1}\mathbf{1}^\top\big], \\ \boldsymbol{Y}\boldsymbol{Y}^\top &= d\dot{r}\big[(1-\dot{c})\boldsymbol{I}_T + \dot{c}\mathbf{1}\mathbf{1}^\top\big]. \end{aligned} \tag{32}$$

Thus $\langle \boldsymbol{X}_t, \boldsymbol{X}_{t'} \rangle$ equals dr for $t = t'$ and drc otherwise; $\langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'} \rangle$ similarly equals $d\dot{r}$ or $d\dot{r}\dot{c}$. For $T \geq 2$, positive semidefiniteness requires $c, \dot{c} \in [-1/(T-1), 1]$.

All expectations below are conditional on the input and upstream gradient. The recurrence concerns expected squared RMS quantities, such as $\dot{r}_{\mathrm{in}} = \mathbb{E}\|\dot{\boldsymbol{X}}_{\mathrm{in}}\|_F^2/(Td)$, under the fresh-weight closure in Assumption 3.1. We take $d, d_h \to \infty$ while holding d/d_h, T, the geometry, and the initialization scales fixed. Second-moment formulas involving U and J are evaluated in the limiting Gaussian logit model. All subsequent moment recurrences are identities between these limiting quantities. These are not samplewise norm identities for finite-width attention.

C.1 Exact Gradient Expressions

The softmax Jacobian from the main text satisfies

$$\boldsymbol{J}_t = \mathrm{Diag}(\boldsymbol{A}_t) - \boldsymbol{A}_t^\top \boldsymbol{A}_t, \qquad \boldsymbol{J}_t^\top = \boldsymbol{J}_t, \quad \boldsymbol{J}_t\mathbf{1} = \mathbf{0}, \quad \mathbf{1}^\top \boldsymbol{J}_t = \mathbf{0}^\top. \tag{33}$$

The last two identities follow from $\boldsymbol{A}_t\mathbf{1} = 1$. Differentiating $\boldsymbol{X}_{\mathrm{out}} = \boldsymbol{A}\boldsymbol{V}$ gives $\dot{\boldsymbol{V}} = \boldsymbol{A}^\top \boldsymbol{Y}$ and $\dot{\boldsymbol{A}} = \boldsymbol{Y}\boldsymbol{V}^\top$. The row-wise softmax chain rule yields

$$\dot{\boldsymbol{G}}_t^\top = \boldsymbol{J}_t \boldsymbol{V}\boldsymbol{Y}_t^\top = \boldsymbol{J}_t \boldsymbol{X}\boldsymbol{W}_V\boldsymbol{Y}_t^\top, \qquad \dot{\boldsymbol{G}}\mathbf{1} = \mathbf{0}. \tag{34}$$

In coordinates, this is

$$\dot{G}_{ts} = A_{ts}\left\langle \boldsymbol{Y}_t, \left(\boldsymbol{X}_s - \sum_{u=1}^{T} A_{tu}\boldsymbol{X}_u\right)\boldsymbol{W}_V \right\rangle. \tag{35}$$

Next, differentiating $\boldsymbol{G} = \boldsymbol{Q}\boldsymbol{K}^\top/\sqrt{d_h}$ gives

$$\dot{\boldsymbol{Q}} = \frac{\dot{\boldsymbol{G}}\boldsymbol{K}}{\sqrt{d_h}}, \qquad \dot{\boldsymbol{K}} = \frac{\dot{\boldsymbol{G}}^\top \boldsymbol{Q}}{\sqrt{d_h}}, \qquad \dot{\boldsymbol{V}} = \boldsymbol{A}^\top \boldsymbol{Y}. \tag{36}$$

The input gradient is therefore

$$\dot{\boldsymbol{X}}_{\mathrm{in}} = \dot{\boldsymbol{X}}^{(Q)} + \dot{\boldsymbol{X}}^{(K)} + \dot{\boldsymbol{X}}^{(V)}, \tag{37}$$

where

$$\dot{\boldsymbol{X}}^{(Q)} = \frac{\dot{\boldsymbol{G}}\boldsymbol{K}\boldsymbol{W}_Q^\top}{\sqrt{d_h}}, \qquad \dot{\boldsymbol{X}}^{(K)} = \frac{\dot{\boldsymbol{G}}^\top \boldsymbol{Q}\boldsymbol{W}_K^\top}{\sqrt{d_h}}, \qquad \dot{\boldsymbol{X}}^{(V)} = \boldsymbol{A}^\top \boldsymbol{Y}\boldsymbol{W}_V^\top. \tag{38}$$

For individual input-token rows, these identities read

$$\dot{\boldsymbol{X}}_t^{(Q)} = \frac{\dot{\boldsymbol{G}}_t \boldsymbol{K}\boldsymbol{W}_Q^\top}{\sqrt{d_h}}, \qquad \dot{\boldsymbol{X}}_s^{(K)} = \frac{(\dot{\boldsymbol{G}}_{:s})^\top \boldsymbol{Q}\boldsymbol{W}_K^\top}{\sqrt{d_h}}, \qquad \dot{\boldsymbol{X}}_s^{(V)} = \sum_{t=1}^{T} A_{ts}\boldsymbol{Y}_t\boldsymbol{W}_V^\top. \tag{39}$$

Finally, differentiating the three linear projections gives

$$\begin{aligned} \dot{\boldsymbol{W}}_Q &= \boldsymbol{X}^\top \dot{\boldsymbol{Q}} = \frac{\boldsymbol{X}^\top \dot{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{W}_K}{\sqrt{d_h}}, \\ \dot{\boldsymbol{W}}_K &= \boldsymbol{X}^\top \dot{\boldsymbol{K}} = \frac{\boldsymbol{X}^\top \dot{\boldsymbol{G}}^\top \boldsymbol{X}\boldsymbol{W}_Q}{\sqrt{d_h}}, \qquad \dot{\boldsymbol{W}}_V = \boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{Y}. \end{aligned} \tag{40}$$

This establishes Eqs. (7)–(9). All identities in this subsection are exact at finite width and do not use the moment assumptions.

C.2 Attention Gradient Recurrence

Gaussian contraction identities. We first record the contractions used for both input and weight gradients. When a reused projection outside the logits is contracted with a quantity depending on those logits, we use an independent copy as in Assumption 3.1, keeping the original notation.

Lemma C.1 (Gaussian linear contraction)**.** *Let $\boldsymbol{W} \in \mathbb{R}^{n\times m}$ have independent entries with law $\mathcal{N}(0, \sigma^2/n)$. For deterministic vectors of compatible dimensions,*

$$\mathbb{E}\langle \boldsymbol{W}^\top \boldsymbol{u}, \boldsymbol{W}^\top \boldsymbol{v}\rangle = \frac{m\sigma^2}{n}\langle \boldsymbol{u}, \boldsymbol{v}\rangle, \tag{41}$$

$$\mathbb{E}\langle \boldsymbol{W}\boldsymbol{a}, \boldsymbol{W}\boldsymbol{b}\rangle = \sigma^2\langle \boldsymbol{a}, \boldsymbol{b}\rangle, \tag{42}$$

$$\mathbb{E}[(\boldsymbol{W}\boldsymbol{a})(\boldsymbol{W}\boldsymbol{b})^\top] = \frac{\sigma^2}{n}\langle \boldsymbol{a}, \boldsymbol{b}\rangle \boldsymbol{I}_n. \tag{43}$$

The same identities hold conditionally for random vectors independent of $\boldsymbol{W}$.

Proof. For example,

$$\mathbb{E}\langle \boldsymbol{W}\boldsymbol{a}, \boldsymbol{W}\boldsymbol{b}\rangle = \sum_{i=1}^{n}\sum_{j,j'=1}^{m} a_j b_{j'} \mathbb{E}[W_{ij}W_{ij'}] = \sum_{i=1}^{n}\sum_{j=1}^{m} a_j b_j \frac{\sigma^2}{n} = \sigma^2\langle \boldsymbol{a}, \boldsymbol{b}\rangle. \tag{44}$$

The other two identities follow from the same entrywise expansion, using $\mathbb{E}[W_{ij}W_{i'j'}] = (\sigma^2/n)\delta_{ii'}\delta_{jj'}$. □

Lemma C.2 (Second moment of the logit-gradient rows)**.** *Conditional on $\boldsymbol{X}$, $\boldsymbol{Y}$, and the softmax Jacobians, averaging over $\boldsymbol{W}_V$ gives*

$$\mathbb{E}_{W_V}[\dot{\boldsymbol{G}}_t^\top \dot{\boldsymbol{G}}_{t'}] = \sigma_V^2 r(1-c)\langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle \boldsymbol{J}_t \boldsymbol{J}_{t'}. \tag{45}$$

In particular,

$$\mathbb{E}_{W_V}\|\dot{\boldsymbol{G}}_t\|_2^2 = \sigma_V^2 d\dot{r}\, r(1-c)\|\boldsymbol{J}_t\|_F^2. \tag{46}$$

Proof. The value weights are independent of the logits and, by Assumption 3.1, of the upstream gradient. Equations (34) and (43) give

$$\begin{aligned}\mathbb{E}_{W_V}[\dot{\boldsymbol{G}}_t^\top \dot{\boldsymbol{G}}_{t'}] &= \boldsymbol{J}_t \boldsymbol{X}\, \mathbb{E}_{W_V}[\boldsymbol{W}_V \boldsymbol{Y}_t^\top \boldsymbol{Y}_{t'} \boldsymbol{W}_V^\top] \boldsymbol{X}^\top \boldsymbol{J}_{t'} \\ &= \frac{\sigma_V^2}{d}\langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle \boldsymbol{J}_t \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{J}_{t'}.\end{aligned} \tag{47}$$

By Eq. (32) and $\boldsymbol{J}_t \mathbf{1} = 0$,

$$\boldsymbol{J}_t \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{J}_{t'} = dr(1-c)\boldsymbol{J}_t \boldsymbol{J}_{t'}. \tag{48}$$

This proves Eq. (45). Setting $t = t'$ and taking the trace proves Eq. (46). □

QK weight variances determine the logit covariance. Since $G_{ts} = d_h^{-1/2}\sum_{j=1}^{d_h} Q_{tj}K_{sj}$, the summands indexed by j are independent and identically distributed. Independence of the query and key weights and Eq. (32) give

$$\mathbb{E}[Q_{tj}Q_{t'j}] = \frac{\sigma_Q^2}{d}\langle \boldsymbol{X}_t, \boldsymbol{X}_{t'}\rangle, \qquad \mathbb{E}[K_{sj}K_{s'j}] = \frac{\sigma_K^2}{d}\langle \boldsymbol{X}_s, \boldsymbol{X}_{s'}\rangle, \tag{49}$$

and hence

$$\mathbb{E}[G_{ts}G_{t's'}] = \frac{\sigma_Q^2\sigma_K^2}{d^2}\langle \boldsymbol{X}_t, \boldsymbol{X}_{t'}\rangle\langle \boldsymbol{X}_s, \boldsymbol{X}_{s'}\rangle. \tag{50}$$

The multivariate central limit theorem gives a jointly centered Gaussian limit for the finite collection of logit entries. For two query rows, use the correlation parameter ρ from Eq. (5), with $\rho = 1$ when $t = t'$ and $\rho = c$ otherwise. Their covariance is

$$\mathrm{Cov}(\boldsymbol{G}_t^\top, \boldsymbol{G}_{t'}^\top) = \sigma_Q^2\sigma_K^2 r^2 \rho\big[(1-c)\boldsymbol{I}_T + c\mathbf{1}\mathbf{1}^\top\big]. \tag{51}$$

Let $\boldsymbol{C}_T \coloneqq \boldsymbol{I}_T - \mathbf{1}\mathbf{1}^\top/T$. Softmax and its Jacobian are invariant to a common shift, so

$$\mathrm{softmax}(\boldsymbol{G}_t^\top) = \mathrm{softmax}(\boldsymbol{C}_T\boldsymbol{G}_t^\top), \qquad \boldsymbol{J}(\boldsymbol{G}_t^\top) = \boldsymbol{J}(\boldsymbol{C}_T\boldsymbol{G}_t^\top). \tag{52}$$

Centering the rows therefore removes the rank-one covariance term:

$$\mathrm{Cov}(\boldsymbol{C}_T\boldsymbol{G}_t^\top, \boldsymbol{C}_T\boldsymbol{G}_{t'}^\top) = \gamma\rho\boldsymbol{C}_T, \qquad \gamma = \sigma_Q^2\sigma_K^2 r^2(1-c). \tag{53}$$

Equivalence of Gaussian expectations. For the jointly Gaussian pair $(\boldsymbol{z}, \boldsymbol{z}')$ in Eq. (5), take the same ρ as above. The pair $(\sqrt{\gamma}\boldsymbol{C}_T\boldsymbol{z}, \sqrt{\gamma}\boldsymbol{C}_T\boldsymbol{z}')$ has marginal covariance $\gamma\boldsymbol{C}_T$ and cross-covariance $\gamma\rho\boldsymbol{C}_T$, exactly matching the centered logit rows. Applying Eq. (52) once more gives, in the Gaussian limit,

$$\mathbb{E}\langle \boldsymbol{A}_t, \boldsymbol{A}_{t'}\rangle = \mathrm{U}_{\gamma,\rho}, \qquad \mathbb{E}\langle \boldsymbol{J}_t, \boldsymbol{J}_{t'}\rangle_F = \mathrm{J}_{\gamma,\rho}. \tag{54}$$

The overlaps are bounded, so convergence in distribution also gives convergence of these expectations. Diagonal row pairs use $\rho = 1$, whereas distinct row pairs use $\rho = c$. This centering argument also applies to admissible negative c; it does not require a decomposition with a real factor $\sqrt{c}$.

C.2.1 Input Gradient

Proof of the input-gradient formula in Theorem 3.3. For $B \in \{V, Q, K\}$, write

$$\dot{r}_{\mathrm{in}}^{(B)} \coloneqq \frac{1}{Td}\mathbb{E}\|\dot{\boldsymbol{X}}^{(B)}\|_F^2. \tag{55}$$

We compute these three moments separately before using path orthogonality.

V path. Since $\boldsymbol{W}_V$ is independent of $\boldsymbol{A}$ and $\boldsymbol{Y}$, Eq. (39) and Lemma C.1 give

$$\mathbb{E}_{W_V}\frac{\|\dot{\boldsymbol{X}}_s^{(V)}\|_2^2}{d} = \frac{\sigma_V^2}{d}\left\|\sum_{t=1}^{T} A_{ts}\boldsymbol{Y}_t\right\|_2^2 = \frac{\sigma_V^2}{d}\sum_{t,t'=1}^{T} A_{ts}A_{t's}\langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle. \tag{56}$$

Averaging over input tokens s and exchanging the sums yields

$$\begin{aligned}\dot{r}_{\mathrm{in}}^{(V)} &= \frac{\sigma_V^2}{dT}\sum_{t,t'=1}^{T}\langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle\mathbb{E}\langle \boldsymbol{A}_t, \boldsymbol{A}_{t'}\rangle \\ &= \sigma_V^2\dot{r}\left[\mathrm{U}_{\gamma,1} + (T-1)\dot{c}\,\mathrm{U}_{\gamma,c}\right].\end{aligned} \tag{57}$$

The last equality uses Eq. (54) and the T diagonal and $T(T-1)$ off-diagonal row pairs.

Q path. Use the fresh-weight contraction first on the outer $\boldsymbol{W}_Q$ in Eq. (39):

$$\mathbb{E}\frac{\|\dot{\boldsymbol{X}}_t^{(Q)}\|_2^2}{d} = \frac{\sigma_Q^2}{dd_h}\mathbb{E}\|\dot{\boldsymbol{G}}_t\boldsymbol{K}\|_2^2. \tag{58}$$

Next contract $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K$ independently of $\dot{\boldsymbol{G}}_t$, as prescribed by Assumption 3.1:

$$\mathbb{E}\|\dot{\boldsymbol{G}}_t\boldsymbol{K}\|_2^2 = \frac{d_h\sigma_K^2}{d}\mathbb{E}[\dot{\boldsymbol{G}}_t\boldsymbol{X}\boldsymbol{X}^\top\dot{\boldsymbol{G}}_t^\top]. \tag{59}$$

Since $\dot{\boldsymbol{G}}_t\mathbf{1} = 0$, Eq. (32) implies

$$\dot{\boldsymbol{G}}_t\boldsymbol{X}\boldsymbol{X}^\top\dot{\boldsymbol{G}}_t^\top = dr(1-c)\|\dot{\boldsymbol{G}}_t\|_2^2. \tag{60}$$

Combining these contractions with Eq. (46),

$$\begin{aligned}\mathbb{E}\frac{\|\dot{\boldsymbol{X}}_t^{(Q)}\|_2^2}{d} &= \frac{\sigma_Q^2\sigma_K^2 r(1-c)}{d}\mathbb{E}\|\dot{\boldsymbol{G}}_t\|_2^2 \\ &= \sigma_V^2\dot{r}\,\sigma_Q^2\sigma_K^2 r^2(1-c)^2\mathbb{E}\|\boldsymbol{J}_t\|_F^2 \\ &= \sigma_V^2\dot{r}\,\gamma(1-c)\mathrm{J}_{\gamma,1}.\end{aligned} \tag{61}$$

All query rows have the same marginal law, so averaging over t gives

$$\dot{r}_{\mathrm{in}}^{(Q)} = \sigma_V^2\dot{r}\,\gamma(1-c)\mathrm{J}_{\gamma,1}. \tag{62}$$

K path. For an input key token s, contract the outer $\boldsymbol{W}_K$ to obtain

$$\mathbb{E}\frac{\|\dot{\boldsymbol{X}}_s^{(K)}\|_2^2}{d} = \frac{\sigma_K^2}{dd_h}\mathbb{E}\|(\dot{\boldsymbol{G}}_{:s})^\top \boldsymbol{Q}\|_2^2$$
$$= \frac{\sigma_K^2}{dd_h}\sum_{t,t'=1}^{T}\mathbb{E}\left[\dot{G}_{ts}\dot{G}_{t's}\langle \boldsymbol{Q}_t, \boldsymbol{Q}_{t'}\rangle\right]. \tag{63}$$

The fresh query-weight contraction gives $\mathbb{E}\langle \boldsymbol{Q}_t, \boldsymbol{Q}_{t'}\rangle = (d_h\sigma_Q^2/d)\langle \boldsymbol{X}_t, \boldsymbol{X}_{t'}\rangle$, independently of the logit-gradient factors. Therefore

$$\mathbb{E}\frac{\|\dot{\boldsymbol{X}}_s^{(K)}\|_2^2}{d} = \frac{\sigma_Q^2\sigma_K^2}{d^2}\sum_{t,t'=1}^{T}\langle \boldsymbol{X}_t, \boldsymbol{X}_{t'}\rangle\mathbb{E}[\dot{G}_{ts}\dot{G}_{t's}]. \tag{64}$$

Averaging over s and using $\sum_s \dot{G}_{ts}\dot{G}_{t's} = \langle \dot{\boldsymbol{G}}_t, \dot{\boldsymbol{G}}_{t'}\rangle$ gives

$$\dot{r}_{\text{in}}^{(K)} = \frac{\sigma_Q^2\sigma_K^2}{d^2T}\sum_{t,t'=1}^{T}\langle \boldsymbol{X}_t, \boldsymbol{X}_{t'}\rangle\mathbb{E}\langle \dot{\boldsymbol{G}}_t, \dot{\boldsymbol{G}}_{t'}\rangle. \tag{65}$$

Taking the trace of Eq. (45) and then averaging over the logits yields

$$\mathbb{E}\langle \dot{\boldsymbol{G}}_t, \dot{\boldsymbol{G}}_{t'}\rangle = \sigma_V^2 r(1-c)\langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle\mathbb{E}\langle \boldsymbol{J}_t, \boldsymbol{J}_{t'}\rangle_F. \tag{66}$$

Substituting into Eq. (65),

$$\dot{r}_{\text{in}}^{(K)} = \frac{\sigma_Q^2\sigma_K^2\sigma_V^2 r(1-c)}{d^2T}\sum_{t,t'=1}^{T}\langle \boldsymbol{X}_t, \boldsymbol{X}_{t'}\rangle\langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle\mathbb{E}\langle \boldsymbol{J}_t, \boldsymbol{J}_{t'}\rangle_F$$
$$= \sigma_V^2\dot{r}\,\gamma\left[\mathrm{J}_{\gamma,1} + (T-1)c\dot{c}\,\mathrm{J}_{\gamma,c}\right]. \tag{67}$$

Here the product of the two row inner products is $d^2r\dot{r}$ on the diagonal and $d^2r\dot{r}\,c\dot{c}$ off the diagonal.

Combining the paths. Assumption 3.2 removes the expected cross inner products, giving

$$\frac{1}{Td}\mathbb{E}\left\|\dot{\boldsymbol{X}}^{(V)} + \dot{\boldsymbol{X}}^{(Q)} + \dot{\boldsymbol{X}}^{(K)}\right\|_F^2 = \dot{r}_{\text{in}}^{(V)} + \dot{r}_{\text{in}}^{(Q)} + \dot{r}_{\text{in}}^{(K)}. \tag{68}$$

Adding Eqs. (57), (62), and (67), and passing to the width limit gives

$$\dot{r}_{\text{in}} = \sigma_V^2\dot{r}\Big[\mathrm{U}_{\gamma,1} + (T-1)\dot{c}\,\mathrm{U}_{\gamma,c} + \gamma\big((2-c)\mathrm{J}_{\gamma,1} + (T-1)c\dot{c}\,\mathrm{J}_{\gamma,c}\big)\Big]. \tag{69}$$

Replacing $(r, c, \dot{r}, \dot{c})$ by $(r_{\text{in}}, c_{\text{in}}, \dot{r}_{\text{out}}, \dot{c}_{\text{out}})$ proves Eq. (10). □

C.2.2 Weight Gradients

Proof of the weight-gradient formulas in Theorem 3.3. The weight-gradient squared RMS uses the number of parameters as its denominator:

$$\dot{r}_{\boldsymbol{W}_Q} \coloneqq \frac{\mathbb{E}\|\dot{\boldsymbol{W}}_Q\|_F^2}{dd_h}, \qquad \dot{r}_{\boldsymbol{W}_K} \coloneqq \frac{\mathbb{E}\|\dot{\boldsymbol{W}}_K\|_F^2}{dd_h}, \qquad \dot{r}_{\boldsymbol{W}_V} \coloneqq \frac{\mathbb{E}\|\dot{\boldsymbol{W}}_V\|_F^2}{d^2}. \tag{70}$$

The following calculation uses the same logit-gradient second moment as the input recurrence. Path orthogonality is not needed for the individual weight moments.

Q and K weights. Contracting the final weight factor in Eq. (40) independently of $\boldsymbol{X}^\top\dot{\boldsymbol{G}}\boldsymbol{X}$ under Assumption 3.1 gives

$$\mathbb{E}\|\dot{\boldsymbol{W}}_Q\|_F^2 = \frac{\sigma_K^2}{d}\mathbb{E}\|\boldsymbol{X}^\top\dot{\boldsymbol{G}}\boldsymbol{X}\|_F^2,$$
$$\mathbb{E}\|\dot{\boldsymbol{W}}_K\|_F^2 = \frac{\sigma_Q^2}{d}\mathbb{E}\|\boldsymbol{X}^\top\dot{\boldsymbol{G}}\boldsymbol{X}\|_F^2. \tag{71}$$

Indeed, $\mathbb{E}[\boldsymbol{W}_K \boldsymbol{W}_K^\top] = (d_h \sigma_K^2/d)\boldsymbol{I}_d$, and the factor d_h cancels the $1/d_h$ from logit differentiation. Transposition leaves the Frobenius norm of $\boldsymbol{X}^\top \dot{\boldsymbol{G}} \boldsymbol{X}$ unchanged, explaining the symmetric QK formulas even when $\sigma_Q \neq \sigma_K$.

Since $\dot{\boldsymbol{G}}\mathbf{1} = 0$, token symmetry gives $\dot{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{X}^\top = dr(1-c)\dot{\boldsymbol{G}}$. Consequently,

$$\begin{aligned}\|\boldsymbol{X}^\top \dot{\boldsymbol{G}} \boldsymbol{X}\|_F^2 &= \operatorname{tr}(\dot{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{X}^\top \dot{\boldsymbol{G}}^\top \boldsymbol{X}\boldsymbol{X}^\top) \\ &= dr(1-c)\operatorname{tr}(\dot{\boldsymbol{G}}\dot{\boldsymbol{G}}^\top \boldsymbol{X}\boldsymbol{X}^\top) \\ &= dr(1-c)\sum_{t,t'=1}^{T} \langle \boldsymbol{X}_t, \boldsymbol{X}_{t'}\rangle \langle \dot{\boldsymbol{G}}_t, \dot{\boldsymbol{G}}_{t'}\rangle. \end{aligned} \tag{72}$$

Apply Eq. (66) and then Eq. (54):

$$\begin{aligned}\mathbb{E}\|\boldsymbol{X}^\top \dot{\boldsymbol{G}} \boldsymbol{X}\|_F^2 &= \sigma_V^2 dr^2(1-c)^2 \sum_{t,t'=1}^{T} \langle \boldsymbol{X}_t, \boldsymbol{X}_{t'}\rangle \langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle \mathbb{E}\langle \boldsymbol{J}_t, \boldsymbol{J}_{t'}\rangle_F \\ &= \sigma_V^2 d^3 \dot{r}\, r^3 (1-c)^2 T \left[\mathrm{J}_{\gamma,1} + (T-1)c\dot{c}\,\mathrm{J}_{\gamma,c}\right]. \end{aligned} \tag{73}$$

Combining with Eqs. (70) and (71) yields

$$\dot{r}_{\boldsymbol{W}_Q} = \frac{d}{d_h}\sigma_K^2 \sigma_V^2 r^3 (1-c)^2 T \left[\mathrm{J}_{\gamma,1} + (T-1)c\dot{c}\,\mathrm{J}_{\gamma,c}\right]\dot{r}, \tag{74}$$

$$\dot{r}_{\boldsymbol{W}_K} = \frac{d}{d_h}\sigma_Q^2 \sigma_V^2 r^3 (1-c)^2 T \left[\mathrm{J}_{\gamma,1} + (T-1)c\dot{c}\,\mathrm{J}_{\gamma,c}\right]\dot{r}. \tag{75}$$

One factor $1-c$ comes from $\dot{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{X}^\top = dr(1-c)\dot{\boldsymbol{G}}$; the other comes from the logit-gradient covariance. The factor T remains because a shared parameter gradient sums token contributions without the $1/T$ used to average input-gradient energy.

V weights. The exact value-weight gradient is $\dot{\boldsymbol{W}}_V = \boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{Y} = \sum_t (\boldsymbol{A}_t \boldsymbol{X})^\top \boldsymbol{Y}_t$, so

$$\|\dot{\boldsymbol{W}}_V\|_F^2 = \sum_{t,t'=1}^{T} \langle \boldsymbol{A}_t \boldsymbol{X}, \boldsymbol{A}_{t'} \boldsymbol{X}\rangle \langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle. \tag{76}$$

Using $\boldsymbol{A}_t \mathbf{1} = 1$ and the input Gram matrix,

$$\langle \boldsymbol{A}_t \boldsymbol{X}, \boldsymbol{A}_{t'} \boldsymbol{X}\rangle = \boldsymbol{A}_t \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{A}_{t'}^\top = dr\left[(1-c)\langle \boldsymbol{A}_t, \boldsymbol{A}_{t'}\rangle + c\right]. \tag{77}$$

Thus

$$\begin{aligned}\dot{r}_{\boldsymbol{W}_V} &= \frac{r}{d}\sum_{t,t'=1}^{T} \langle \boldsymbol{Y}_t, \boldsymbol{Y}_{t'}\rangle \left[(1-c)\mathbb{E}\langle \boldsymbol{A}_t, \boldsymbol{A}_{t'}\rangle + c\right] \\ &= rT\Big[(1-c)\big(\mathrm{U}_{\gamma,1} + (T-1)\dot{c}\,\mathrm{U}_{\gamma,c}\big) + c\big(1 + (T-1)\dot{c}\big)\Big]\dot{r} + o(1). \end{aligned} \tag{78}$$

This moment has no explicit σ_V factor because differentiation removes $\boldsymbol{W}_V$. Restoring the original geometry subscripts proves all the weight-gradient statements and completes the proof of Theorem 3.3. □

C.2.3 Small-Initialization QK Weight Gradients

Proof of Proposition 5.1. Take the Gaussian-width limit first and write $s = \sigma_Q = \sigma_K$. Then $\gamma = s^4 r_{\mathrm{in}}^2 (1 - c_{\mathrm{in}}) \to 0$. For fixed T, boundedness and continuity of the softmax Jacobian imply, by dominated convergence, $\mathrm{J}_{\gamma,\rho} \to \mathrm{J}_{0,\rho}$. At zero logits,

$$\mathrm{J}_{0,\rho} = \left\| \frac{1}{T}\left(\boldsymbol{I}_T - \frac{\mathbf{1}\mathbf{1}^\top}{T}\right)\right\|_F^2 = \frac{T-1}{T^2}, \tag{79}$$

independently of ρ. Dividing Eqs. (74) and (75) by s^2 and taking $s \to 0^+$ gives

$$\lim_{s\to 0^+} \frac{\dot{r}_{\boldsymbol{W}_Q}}{s^2} = \lim_{s\to 0^+} \frac{\dot{r}_{\boldsymbol{W}_K}}{s^2} = C^2, \tag{80}$$

with C from Eq. (19). Since $C > 0$, taking square roots proves Eq. (20). □

C.3 Pre-Norm Residual Block Gradient Recurrence

We now return to the block notation and prove Eq. (13), with $(\underline{r}, \underline{c})$ denoting block-input geometry and $(\dot{r}, \dot{c})$ denoting block-output gradient geometry.

RMSNorm Jacobian. For a token $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\|_2^2/d = \underline{r} > 0$, differentiation gives

$$\frac{\partial \operatorname{Norm}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{1}{\sqrt{\underline{r}}} \left(\boldsymbol{I}_d - \frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|_2^2} \right). \tag{81}$$

Let the star denote the adjoint Jacobian acting token-wise. For the attention-input gradient $\dot{\boldsymbol{X}}_{\text{in}}$, its exact squared norm after backpropagation through RMSNorm is

$$\begin{aligned} &\frac{1}{d} \left\| \left(\mathcal{D} \operatorname{Norm}(\underline{\boldsymbol{X}})^* [\dot{\boldsymbol{X}}_{\text{in}}] \right)_t \right\|_2^2 \\ &\qquad = \frac{1}{\underline{r}} \left[\frac{\|\dot{\boldsymbol{X}}_{\text{in},t}\|_2^2}{d} - \frac{\langle \underline{\boldsymbol{X}}_t, \dot{\boldsymbol{X}}_{\text{in},t} \rangle^2}{d \|\underline{\boldsymbol{X}}_t\|_2^2} \right]. \end{aligned} \tag{82}$$

In the fresh-weight calculation, each attention input-gradient path has an independent outer Gaussian projection, as in Eq. (39). Equation (43) then shows that its expected energy along a fixed unit direction is $1/d$ of its total expected energy. Applying Cauchy–Schwarz to the sum of the three paths bounds its radial energy by three times the sum of their radial energies. Together with path orthogonality, the term removed in Eq. (82) is therefore $O(d^{-1})$ of the expected branch energy. Averaging over tokens gives the leading RMSNorm multiplier $1/\underline{r}$.

Residual branch and the factor α^2. Write the forward block as

$$\boldsymbol{X}_{\text{in}} = \operatorname{Norm}(\underline{\boldsymbol{X}}), \qquad \boldsymbol{X}_{\text{out}} = \operatorname{Att}(\boldsymbol{X}_{\text{in}}), \qquad \boldsymbol{X} = \underline{\boldsymbol{X}} + \alpha \boldsymbol{X}_{\text{out}}. \tag{83}$$

The upstream gradient entering attention already contains the residual scale:

$$\dot{\boldsymbol{X}}_{\text{out}} = \alpha \dot{\boldsymbol{X}}, \qquad \dot{r}_{\text{out}} = \alpha^2 \dot{r}. \tag{84}$$

Exact differentiation of the block gives

$$\dot{\underline{\boldsymbol{X}}} = \dot{\boldsymbol{X}} + \mathcal{D} \operatorname{Norm}(\underline{\boldsymbol{X}})^* [\dot{\boldsymbol{X}}_{\text{in}}]. \tag{85}$$

The skip–branch orthogonality in Assumption 3.2 removes the expected cross term. Equation (82) then gives

$$\dot{\underline{r}} = \dot{r} + \frac{1}{\underline{r}} \dot{r}_{\text{in}}. \tag{86}$$

RMSNorm maps $(\underline{r}, \underline{c})$ to $(1, \underline{c})$, so Theorem 3.3 applied with upstream gradient $\alpha \dot{\boldsymbol{X}}$ yields

$$\dot{r}_{\text{in}} = \alpha^2 \dot{\chi}^{\text{att}}(1, \underline{c}, \dot{c}) \dot{r}. \tag{87}$$

Substituting into Eq. (86),

$$\dot{\underline{r}} = \left[1 + \frac{\alpha^2}{\underline{r}} \dot{\chi}^{\text{att}}(1, \underline{c}, \dot{c}) \right] \dot{r}, \tag{88}$$

which proves Eq. (13). The factor α^2 comes from squaring the residual amplitude, and $1/\underline{r}$ from the RMSNorm Jacobian. In particular, $\alpha = L^{-1/2}$ and $\alpha = L^{-1}$ give branch coefficients $1/(L\underline{r})$ and $1/(L^2\underline{r})$, respectively.

For the projection weights in this block, the upstream gradient at $\boldsymbol{X}_{\text{out}}$ is $\alpha \dot{\boldsymbol{X}}$. Their second moments are therefore Eqs. (74)–(78) evaluated at $(r, c) = (1, \underline{c})$ and multiplied by α^2.

D Proofs for Attention Statistics

D.1 Effective Logit Variance

Softmax depends only on differences between logits. We show that these differences have the same limiting distribution as those of independent Gaussian logits with variance γ.

Condition on $\boldsymbol{X}_{\text{in}}$ satisfying Assumption 2.1. Each logit is $G_{ts} = d_h^{-1/2} \sum_{j=1}^{d_h} Q_{tj} K_{sj}$. The summands are centered and independent across j, so the multivariate central limit theorem makes the logits jointly Gaussian as $d_h \to \infty$. Independence of the query and key weights gives the exact covariance

$$\begin{aligned}\mathbb{E}[G_{ts} G_{t'u}] &= \mathbb{E}[Q_{t1} Q_{t'1}]\, \mathbb{E}[K_{s1} K_{u1}] \\ &= \frac{\sigma_Q^2 \sigma_K^2}{d^2} \langle \boldsymbol{X}_{\text{in},t}, \boldsymbol{X}_{\text{in},t'} \rangle \langle \boldsymbol{X}_{\text{in},s}, \boldsymbol{X}_{\text{in},u} \rangle. \end{aligned} \tag{89}$$

Thus, within one row, the logit variance is $\sigma_Q^2 \sigma_K^2 r_{\text{in}}^2$ and the covariance between distinct logits is $\sigma_Q^2 \sigma_K^2 r_{\text{in}}^2 c_{\text{in}}$.

Subtract the last logit from every entry of each row; this leaves softmax unchanged. Expanding the covariance above and using the input geometry gives, for $s, u < T$,

$$\begin{aligned} \mathbb{E}[(G_{ts} - G_{tT})(G_{t'u} - G_{t'T})] &= \gamma \rho \begin{cases} 2, & s = u, \\ 1, & s \neq u, \end{cases} \\ \gamma = \sigma_Q^2 \sigma_K^2 r_{\text{in}}^2 (1 - c_{\text{in}}), \qquad \rho &= \begin{cases} 1, & t = t', \\ c_{\text{in}}, & t \neq t'. \end{cases} \end{aligned} \tag{90}$$

These are exactly the covariances of $\sqrt{\gamma}(z_s - z_T)$ and $\sqrt{\gamma}(z'_u - z'_T)$, where $\boldsymbol{z}, \boldsymbol{z}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_T)$ are jointly Gaussian with $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}'^{\top}] = \rho \boldsymbol{I}_T$. Since the limiting differences are jointly centered Gaussian, matching their covariances matches their joint distribution. Applying softmax to each row of differences, whose last entry is zero, therefore gives, in the Gaussian limit,

$$(\boldsymbol{A}_t^{\top}, \boldsymbol{A}_{t'}^{\top}) \overset{\text{d}}{=} \big(\text{softmax}(\sqrt{\gamma}\, \boldsymbol{z}), \text{softmax}(\sqrt{\gamma}\, \boldsymbol{z}')\big). \tag{91}$$

The softmax Jacobian has the same shift invariance, giving the $\text{U}_{\gamma,\rho}$ and $\text{J}_{\gamma,\rho}$ representations in the main text. In short, γ is the logit variance minus the within-row covariance: shared variation does not affect attention. The argument also covers admissible negative c_{in}; for $T = 1$, attention is identically one.

D.2 Long-Sequence Asymptotics

Proof of Theorem 4.1. We now work in the limiting Gaussian model and let $T \to \infty$; the width limit is taken first. As in Eq. (5), write $\boldsymbol{a} = \text{softmax}(\sqrt{\gamma}\, \boldsymbol{z})$, with $\gamma = \lambda \ln T$ and fixed $\lambda > 0$.

Softmax identities. Since $\boldsymbol{J}(\sqrt{\gamma}\, \boldsymbol{z}) = \text{Diag}(\boldsymbol{a}) - \boldsymbol{a}\boldsymbol{a}^{\top}$,

$$\text{U}_{\gamma,1} = \mathbb{E} \sum_i a_i^2, \qquad \text{J}_{\gamma,1} = \mathbb{E}\left[\sum_i a_i^2 - 2 \sum_i a_i^3 + \left(\sum_i a_i^2 \right)^2 \right]. \tag{92}$$

Cauchy–Schwarz and $\sum_i a_i = 1$ imply $(\sum_i a_i^2)^2 \leq \sum_i a_i^3$, hence $0 \leq \text{J}_{\gamma,1} \leq \text{U}_{\gamma,1}$. Attention weights are positive, and softmax Jacobians are positive semidefinite, so both cross-overlaps are nonnegative. Cauchy–Schwarz and identical row marginals then give

$$0 \leq \text{U}_{\gamma,\rho} \leq \text{U}_{\gamma,1}, \qquad 0 \leq \text{J}_{\gamma,\rho} \leq \text{J}_{\gamma,1} \quad \text{for every } \rho \in [-1, 1]. \tag{93}$$

Diffuse regime: $0 < \lambda < 2$. Choose $\varepsilon > 0$ such that $\sqrt{2\lambda} + 2\varepsilon < 1 + \lambda/2$ and $\varepsilon < 1 - \lambda/2$. The number of coordinates with $z_i \geq \sqrt{\lambda \ln T}$ is binomial with mean

$$T \bar{\Phi}(\sqrt{\lambda \ln T}) = \Theta\Big(T^{1-\lambda/2} / \sqrt{\ln T} \Big). \tag{94}$$

Each such coordinate contributes at least T^{λ} to the softmax denominator. For all sufficiently large T, a Chernoff bound therefore gives

$$\mathbb{P}\left\{ \sum_i e^{\sqrt{\gamma} z_i} < T^{1+\lambda/2-\varepsilon} \right\} \leq \exp[-T^{1-\lambda/2-\varepsilon}]. \tag{95}$$

A Gaussian tail bound and a union bound also give

$$\mathbb{P}\left\{\max_i \sqrt{\gamma} z_i > (\sqrt{2\lambda}+\varepsilon)\ln T\right\} \le T^{1-(\sqrt{2\lambda}+\varepsilon)^2/(2\lambda)}. \tag{96}$$

Outside these two exceptional events,

$$\sum_i a_i^2 \le \max_i a_i \le T^{\sqrt{2\lambda}-1-\lambda/2+2\varepsilon}. \tag{97}$$

The exponent is negative, both exceptional probabilities are $o(1/\ln T)$, and $\max_i a_i \le 1$ always. Consequently $\mathrm{U}_{\gamma,1} = o(1/\ln T)$, so

$$0 \le \gamma \mathrm{J}_{\gamma,1} \le \lambda \ln T\, \mathrm{U}_{\gamma,1} \longrightarrow 0. \tag{98}$$

This also proves the diffuse limit in Eq. (6).

Concentrated regime: $\lambda > 2$. Set $\mu = \sqrt{2/\lambda} \in (0,1)$. The low-temperature Gaussian random-energy-model limit gives convergence in distribution of the ranked attention weights (padded with zeros) in ℓ^1 to the Poisson–Dirichlet law $\mathrm{PD}(\mu, 0)$ (Monthus & Garel, 2007; Pitman & Yor, 1997). The functionals in Eq. (92) are bounded and continuous in ℓ^1, so their expectations converge as well. The one- and two-block moment formula (Pitman & Yor, 1997, Proposition 51) yields

$$\begin{aligned}
\mathbb{E}\sum_i a_i^2 &\longrightarrow 1-\mu, \\
\mathbb{E}\sum_i a_i^3 &\longrightarrow \frac{(1-\mu)(2-\mu)}{2}, \\
\mathbb{E}\left(\sum_i a_i^2\right)^2 &\longrightarrow \frac{(1-\mu)(2-\mu)(3-\mu)+\mu(1-\mu)^2}{6}.
\end{aligned} \tag{99}$$

In the last line, the two terms come from $\sum_i a_i^4$ and $\sum_{i\neq j} a_i^2 a_j^2$, respectively. Substitution into Eq. (92) gives

$$\mathrm{J}_{\gamma,1} \longrightarrow \frac{\mu(1-\mu)}{3}, \qquad \frac{\gamma \mathrm{J}_{\gamma,1}}{\ln T} \longrightarrow \frac{\lambda\mu(1-\mu)}{3} = \frac{\sqrt{2\lambda}}{3}\left(1-\sqrt{\frac{2}{\lambda}}\right). \tag{100}$$

The first line of Eq. (99) also proves the concentrated limit in Eq. (6). □

Consequences for the gradient multiplier. Assume $c_{\mathrm{in}} \in [0,1)$ and $\dot{c}_{\mathrm{out}} \ge 0$, and keep $\sigma_V > 0$ fixed. Every term in Eq. (10) is nonnegative by Eq. (93). Keeping only the self QK term proves Eq. (15); the concentrated limit above then forces $\dot{\chi}^{\mathrm{att}} \to \infty$.

For $0 < \lambda < 2$, the same bounds give

$$0 \le \dot{\chi}^{\mathrm{att}} \le \sigma_V^2 \big[1 + (T-1)\dot{c}_{\mathrm{out}}\big]\big[\mathrm{U}_{\gamma,1} + 2\gamma \mathrm{J}_{\gamma,1}\big]. \tag{101}$$

Thus the full multiplier vanishes if $(T-1)\dot{c}_{\mathrm{out}}$ remains bounded. It need not vanish otherwise: when $c_{\mathrm{in}} = 0$, the Gaussian rows are independent and permutation symmetry gives $\mathbb{E}a_i = 1/T$, hence $\mathrm{U}_{\gamma,0} = 1/T$ exactly. For fixed $\dot{c}_{\mathrm{out}} > 0$, the recurrence reduces to

$$\dot{\chi}^{\mathrm{att}} = \sigma_V^2\left[\mathrm{U}_{\gamma,1} + \frac{T-1}{T}\dot{c}_{\mathrm{out}} + 2\gamma \mathrm{J}_{\gamma,1}\right] \longrightarrow \sigma_V^2 \dot{c}_{\mathrm{out}}. \tag{102}$$

Finally, RMSNorm gives $r_{\mathrm{in}} = 1$ and $c_{\mathrm{in}} = \underline{c}$. For equal QK scales, substituting $\gamma = \sigma_{qk}^4(1-\underline{c})$ into the boundary $\gamma = 2\ln T$ yields $\sigma_{qk,+}(\underline{c}) = [2\ln T/(1-\underline{c})]^{1/4}$, and Eq. (16) follows when $\underline{c} \approx 0$.

E Further Theoretical Results

We next study how token geometry feeds back into attention statistics across depth and which conclusions extend to causal attention. The depth result uses an explicit forward moment closure; the causal extension separates rowwise concentration from the full backward operator.

E.1 Depth Feedback Through Token Correlation

Proposition E.1 (Correlation growth under the moment closure)**.** *Work under the wide-limit normalization convention of Section 2. Consider a Pre-Norm residual attention block with input squared RMS $r > 0$, token correlation $c \in [0, 1)$, residual multiplier α, and independent centered value weights. Write $s = \sigma_Q = \sigma_K$. For an equal-norm exchangeable input Gram matrix, set*

$$A_c = c + (1-c)\mathrm{U}_{\gamma,1}, \qquad B_c = c + (1-c)\mathrm{U}_{\gamma,c}, \qquad \gamma = s^4(1-c). \tag{103}$$

In the Gaussian row limit, the expected diagonal and off-diagonal coordinate moments after the block are $r^+ = r + \alpha^2\sigma_V^2 A_c$ and $r^+c^+ = rc + \alpha^2\sigma_V^2 B_c$. Their ratio satisfies

$$c^+ - c = \frac{\alpha^2\sigma_V^2(1-c)[c(1-\mathrm{U}_{\gamma,1}) + \mathrm{U}_{\gamma,c}]}{r + \alpha^2\sigma_V^2 A_c} \geq 0. \tag{104}$$

If each layer is closed by replacing its input Gram matrix by the equal-norm exchangeable matrix with these moments, and s is the same at every layer, then c_ℓ is nondecreasing and $\gamma_{\ell+1} = s^4(1-c_\ell)$ is nonincreasing.

Proof of Proposition E.1. Write $C = (1-c)I + c\mathbf{1}\mathbf{1}^\top$, and let the input satisfy $XX^\top/d = rC$. Equal row norms imply that RMSNorm gives $Y = X/\sqrt{r}$ and hence $YY^\top/d = C$. Conditional on Y and the attention matrix A, independence of the centered value weights implies

$$\mathbb{E}_{W_V}\left[\frac{(AYW_V)(AYW_V)^\top}{d}\right] = \sigma_V^2 ACA^\top. \tag{105}$$

Because $A\mathbf{1} = \mathbf{1}$, this is $\sigma_V^2[(1-c)AA^\top + c\mathbf{1}\mathbf{1}^\top]$. The diagonal entries of $AA^\top$ have expectation $\mathrm{U}_{\gamma,1}$; distinct off-diagonal entries have expectation $\mathrm{U}_{\gamma,c}$ under the correlated Gaussian row law. This proves the expressions $\sigma_V^2 A_c$ and $\sigma_V^2 B_c$ for the attention output moments.

For $X^+ = X + \alpha AYW_V$, the mixed terms have zero expectation: conditional on X, A, each contains a single factor of the centered matrix W_V. Thus its expected Gram matrix has diagonal $r^+ = r + \alpha^2\sigma_V^2 A_c$ and off-diagonal $rc + \alpha^2\sigma_V^2 B_c$. Subtracting c from their ratio gives

$$c^+ - c = \frac{\alpha^2\sigma_V^2(B_c - cA_c)}{r + \alpha^2\sigma_V^2 A_c}, \tag{106}$$

$$B_c - cA_c = c + (1-c)\mathrm{U}_{\gamma,c} - c^2 - c(1-c)\mathrm{U}_{\gamma,1} \tag{107}$$

$$= (1-c)[c(1-\mathrm{U}_{\gamma,1}) + \mathrm{U}_{\gamma,c}]. \tag{108}$$

The denominator is positive. The numerator is nonnegative because $0 \leq c < 1$, $\mathrm{U}_{\gamma,1} \leq 1$, and $\mathrm{U}_{\gamma,c} \geq 0$. Moreover, $\mathrm{U}_{\gamma,c} \leq \mathrm{U}_{\gamma,1}$ by Cauchy–Schwarz and identical row marginals, implying $B_c \leq A_c$. In fact,

$$1 - c^+ = \frac{r(1-c) + \alpha^2\sigma_V^2(A_c - B_c)}{r^+} > 0, \tag{109}$$

because $r > 0$ and $c < 1$, so the updated correlation remains in the same admissible interval. At finite variance and $T \geq 2$, strictly positive softmax weights give $\mathrm{U}_{\gamma,c} > 0$; therefore the inequality is strict when $\alpha\sigma_V \neq 0$ and $c < 1$.

To iterate, the additional moment closure replaces the random output Gram matrix by $r^+[(1-c^+)I + c^+\mathbf{1}\mathbf{1}^\top]$ before the next block. Repeating the proved one-layer inequality gives $c_\ell \geq c_{\ell-1}$ by induction. The next ideal normalization sets the attention-input squared RMS to one, so fixed s gives $\gamma_{\ell+1} = s^4(1-c_\ell) \leq s^4(1-c_{\ell-1}) = \gamma_\ell$. The proof uses this explicit centered moment closure, nonnegative correlations, and fixed QK scale; it is not an induction about the random sample Gram matrix of an individual network. □

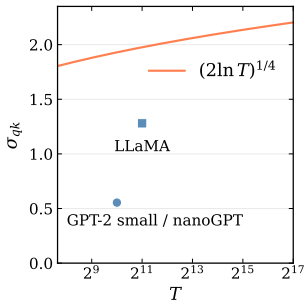

Figure 6: QK initialization upper reference. Curves show $\sigma_{qk,+} = (2\ln T)^{1/4}$. Markers show common initialization conventions; numerical values are given in Table 1.

Implementation	d	T	σ_{qk}	$\sigma_{qk,+}$
GPT-2 small / nanoGPT default	768	1024	0.554	1.930
Transformers LLaMA default	4096	2048	1.280	1.976

Table 1: Common QK initialization conventions and their corresponding upper reference scales. Both implementations use entrywise QK standard deviation 0.02, giving $\sigma_{qk} = 0.02\sqrt{d}$. The proposed warning scale is $\sigma_{qk,+} = (2\ln T)^{1/4}$, assuming $r = 1$ and $c = 0$. Sources are the released GPT-2 implementation and nanoGPT (OpenAI, 2019; Karpathy, 2022), and the default LLaMA configuration and initializer in Transformers v4.44.2 (Hugging Face, 2024). The LLaMA entry describes a software default rather than all Meta pretraining recipes.

F Experimental Details and Additional Results

This appendix defines the measurement conventions, describes the controlled initialization experiments and their validation results, and records the language-model training protocol. All QK sweeps use paired initial random matrices and fixed data streams within each seed; only the QK standard-deviation multiplier changes.

F.1 Measurement and Aggregation

The following conventions apply to the initialization and training measurements reported in this paper.

For one attention row define $u = \sum_j a_j^2$, $D_1 = \max_j a_j$, and $D_2 = \max_j a_j^2/u$. The empirical mean overlap is $\widehat{U} = \mathrm{mean}(u)$. The following participation summaries are distinct:

$$\frac{1}{\mathrm{mean}(u)}, \qquad \mathrm{mean}\left(\frac{1}{u}\right), \qquad \exp[-\mathrm{mean}(\ln u)]\,. \tag{110}$$

The concentration theorem concerns a row's logarithmic scaling. Arithmetic mean U and J enter the gradient law. Experimental support curves labeled $1/\widehat{U}$ use the first summary; they are not asserted to obey the typical exponent without separately examining the geometric-mean statistic. Maximum-weight diagnostics use row medians as specified in their captions.

Each seed's geometry prediction, product $\widehat{\gamma}\widehat{J}$, and inverse overlap are computed before seed aggregation. Bands show one standard deviation across seeds, not confidence intervals. Layer averages include attention layers 1 through L and exclude the entrance state. The geometric mean of local block multipliers and total observed propagation $\dot{r}_0/\dot{r}_L$ are reported separately. Failed runs and nonfinite layers are retained as failures rather than discarded from averages.

For $P \in \{Q, K, V\}$, parameter-gradient RMS is computed before clipping. The actual relative step and relative displacement from initialization are

$$R_{\Delta,P}(k) = \frac{\|\boldsymbol{W}_{P,k+1} - \boldsymbol{W}_{P,k}\|_F}{\|\boldsymbol{W}_{P,k}\|_F}, \qquad R_{0,P}(k) = \frac{\|\boldsymbol{W}_{P,k} - \boldsymbol{W}_{P,0}\|_F}{\|\boldsymbol{W}_{P,0}\|_F}. \tag{111}$$

For aggregate QK/V metrics, numerator and denominator concatenate the corresponding matrices across layers before taking norms. Exact-zero denominators are undefined and displayed separately. Gradient RMS and squared RMS are explicitly distinguished when comparing power laws.

F.2 Controlled Initialization Experiments

Forward and backward calculations use FP32; statistics are reduced in float64. Gaussian-softmax reference curves use 4096 Monte Carlo rows and seed 2026, with common Gaussian draws across scales. Local recurrence predictions condition on measured layer geometry. A separate comparison of γ and $\widehat{\gamma}$ tests the Gaussian geometry reduction; measured logit variance is not silently substituted for the predicted variance.

Concentration and backward gain. Single-layer controlled experiments have $T = 4096$, $d = d_h = 768$, four sequences per seed, target input energy $r = 1$, and token correlations $c \in \{0, 0.5, 0.9\}$. An independent Gaussian probe produces target upstream geometry $\dot{r} = 1, \dot{c} = 0$ through $\mathcal{L} = \langle \dot{\boldsymbol{X}}_{\mathrm{out}}, \boldsymbol{X}_{\mathrm{out}} \rangle_F$. Input rows are normalized before the forward pass, and measured geometry is retained because finite-width normalization perturbs the target correlations. The scale grid $s = 0, 0.1, \ldots, 10$ gives 909 configurations over seeds $1, 22, 333$.

This sweep tests the predicted relation between γ, attention concentration, and backward gain (Figure 2). The two analytic reference scales organize smooth finite-T crossovers; their locations are not fitted to the observations.

Conditional gradient recurrence. The attention-only stacks have $d = d_h = 768$, one head, depth $L = 100$, and value standard-deviation multiplier $\sigma_V = 1$. They have no output projection, MLP, dropout, RoPE, or final normalization. Token and absolute position embeddings have entrywise standard deviation 0.02. Pure attention applies Norm once to their sum at the entrance. Residual architectures instead start directly from the embedding sum, whose expected per-coordinate energy is approximately 0.0008, and apply Norm before each attention branch. The three residual amplitudes are $\alpha \in \{1, L^{-1/2}, L^{-1}\}$. These experiments implement Norm with RMSNorm, gain initialized to one, and $\epsilon_{\mathrm{norm}} = 10^{-6}$.

The depth grid uses

$$\begin{aligned} T &\in \{512, 1024, 2048, 3072, 4096, 6144, 8192, 12000\}, \\ s &\in \{0, 0.001, 0.01, 0.1, 0.5, 1, 1.1, 1.3, 1.5, 1.7, 1.9, \\ &\qquad 2, 2.1, 2.3, 2.5, 2.7, 2.9, 3, 5, 10\}, \end{aligned}$$

with seeds $1, 22, 333$. The number of input sequences per seed is respectively $16, 8, 4, 4, 4, 2, 1, 1$ for the eight lengths. This gives 1920 stack configurations. The representative recurrence comparison reuses the $T = 4096$ subset with $s \in \{0, 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 5, 10\}$, rather than running a separate depth sweep. The upstream task gradient comes from masked-token cross entropy on FineWeb sequences with a 15% corruption mask. The attention itself is full and unmasked. The language-model head shares the token embedding matrix, with vocabulary size 50304 and mask token ID 50303.

The principal comparison uses Pre-Norm residual attention with $\alpha = L^{-1/2}$; pure attention and residual scales 1 and L^{-1} are controls. For the representative $T = 4096$ comparison, each local prediction uses the measured token and upstream-gradient geometry. The same terminal gradient then initializes the backward reconstruction (Figure 3). The measurements support the predicted scale-dependent trends, with finite-width and accumulated discrepancies: local agreement is stronger than agreement after one hundred layers.

First-layer indicator and depth feedback. The depth grid above covers eight sequence lengths from 512 to 12000. Small $\gamma_1 / \ln T$ tends to accompany increasing token correlation and decreasing

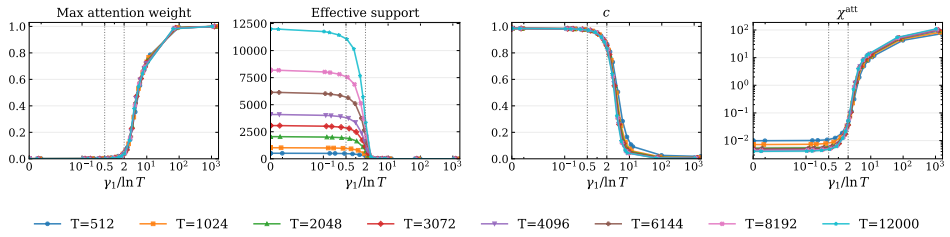

Figure 7: Effective logit variance in the first layer indicates deep layers' regimes.

later variance, while sufficiently large initial values can sustain concentrated attention over the observed depth (Figure 7). This persistence is an empirical finding: the moment-closure theorem gives a one-sided upper reference on later variance, rather than a guarantee of remaining above the boundary.

Small-scale gradients and first-step motion. The dense single-layer grid checks the attention-regime mechanism; resolving the strict near-zero power law requires the separate logarithmic small-scale panel in Figure 4.

The near-zero sweep tests Q and K parameter-gradient RMS separately against the predicted slope of one, then compares their relative first steps under SGD, AdamW, and Muon (Figure 4). Their paired curves also check that the symmetric QK scaling is not an artifact of averaging the two matrices. Subsequent attention measurements determine whether this parameter motion actually changes the initial attention state.

F.3 Gradient-Path Orthogonality

We test Assumption 3.2 using the controlled stacks in Appendix F.2. QK/V paths are compared at the attention input; skip and residual paths are compared at the block input, after backpropagation through normalization. The relative norm-additivity error is the absolute difference between the squared norm of the summed gradient and the sum of the path squared norms, divided by the former. Zero denominators are left undefined.

F.4 Language-Model Training

Model, data, and initialization. The model has approximately 124M parameters, with 12 blocks, width 768, 12 attention heads, and context length 4096. It uses bias-free projections, Pre-RMSNorm, RoPE (Su et al., 2021), SwiGLU, a final RMSNorm, and a tied embedding/output head. RMSNorm has gain initialized to one and $\epsilon_{\text{norm}} = 10^{-5}$; RoPE has base 10000; the SwiGLU hidden width is 2048; and dropout is zero.

We train with BF16 causal next-token cross entropy on the FineWeb `sample-10BT` subset (Penedo et al., 2024), using GPT-2 tokenization and vocabulary size 50304. A micro-batch of 16 sequences and eight accumulation steps give $524{,}288$ raw tokens per update. Ordinary linear and embedding weights have entrywise standard deviation 0.02; attention-output and MLP-down projections use $0.02/\sqrt{2L}$. The sweep overrides only Q and K, whose entries have standard deviation $\sigma_{qk}/\sqrt{768}$.

The 15 initialization settings are

$$\sigma_{qk} \in \{0.02\sqrt{768}\} \cup \{(\kappa \ln 4096)^{1/4} : \kappa \in \{0, 0.125, 0.25, 0.5, 1, 2, 4, 8\}\} \cup \{10^{-5}, 10^{-4}, 10^{-3}, 0.1, 5, 10\}.$$

Seeds are $1, 22, 333$. Within a seed, all scales reuse the same base QK matrices and training-token order; model initialization, training, validation, and diagnostic probes use separate random streams.

Optimizer recipes. Table 2 gives the three recipes. AdamW and Muon share the same learning-rate schedule, token budget, weight decay, and global gradient clipping. Muon updates every hidden

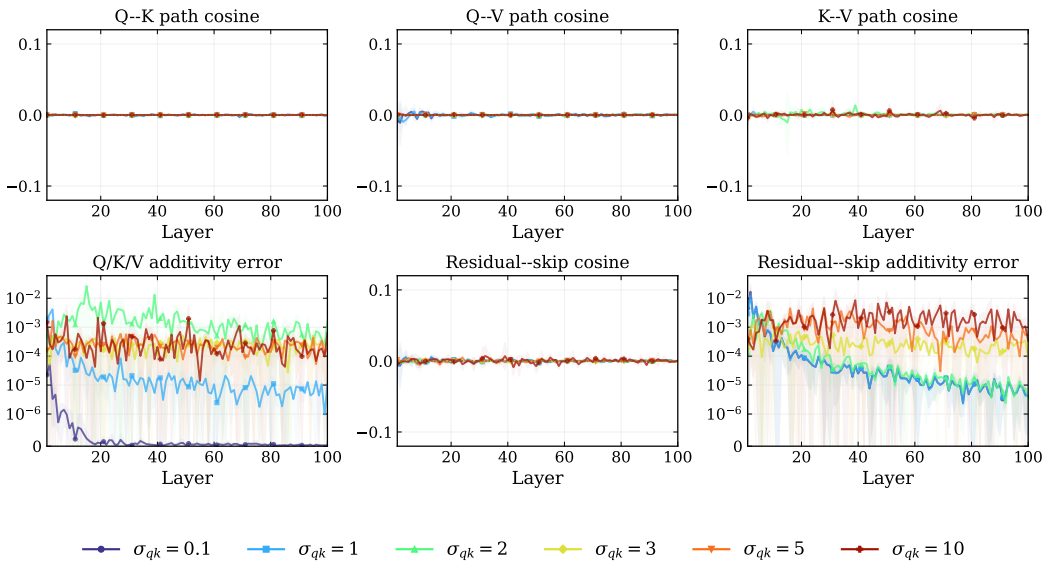

Figure 8: **Gradient paths are approximately orthogonal, supporting the norm-additivity assumption.** Across 100 layers and representative QK scales, pairwise Frobenius cosines among the Q/K/V input-gradient paths and between the residual and skip paths remain near zero, while the corresponding squared-norm additivity errors remain small. The setup matches Figure 3 ($T = 4096$ and $\alpha = L^{-1/2}$); curves and bands report the mean and one standard deviation across three initialization seeds.

two-dimensional matrix except the tied embedding/head; an auxiliary AdamW with the same moment and epsilon settings updates that matrix and the normalization gains. The SGD sweep covers five learning rates, with $\eta = 0.05$ used for the mechanism comparison in the main text.

Setting	AdamW	Muon	Plain SGD
Updates	5000	5000	2000
Raw tokens per run	2.62144×10^9	2.62144×10^9	1.048576×10^9
Learning rate	Peak 6×10^{-4}	Peak 6×10^{-4}	$0.001, 0.01, 0.05, 0.1, 0.2$
Schedule	100-step warmup; cosine to 6×10^{-5}	Same as AdamW	Constant
Update rule	$\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$	Momentum 0.95, Nesterov, 5 Newton–Schulz steps, `match_rms_adamw`	No momentum
Weight decay	0.1 on matrix weights	0.1 on matrix weights	0
Global clipping	L2 threshold 1.0	L2 threshold 1.0	None

Table 2: Language-model training recipes. Comparisons are across QK scales within a recipe. In particular, the shorter SGD run is a raw-gradient mechanism control, not a compute-matched optimizer benchmark.

Evaluation and diagnostics. Validation uses a fixed sample stream and 20 evaluation batches every 100 updates and at termination. A fixed eight-sequence probe records per-layer, per-head attention mechanisms every 100 updates and propagation statistics every 1000 updates, including initialization and termination. The training log records the actual learning rate, pre-clipping global and Q/K/V gradient RMS, clipping scale, Q/K/V weight RMS, actual relative update, displacement from initialization, and completion or abort status. Terminal summaries report the mean and one standard deviation across the three seeds. Nonfinite runs remain recorded as aborted and are not assigned a terminal validation loss. All 45 AdamW and 45 Muon runs complete; at the displayed SGD learning rate, 42 of 45 runs complete and all three $\sigma_{qk} = 10$ runs abort.

For causal attention, $\widehat{\gamma}$ averages unbiased centered variances over rows $t \geq 2$ and visible keys. The one-key first row is excluded from this variance estimate. Attention $\widehat{U}$ and $\widehat{J}$ include all rows, with the first row contributing 1 and 0, respectively. This distinction prevents an undefined one-sample variance from entering the causal diagnostic.

Trajectory construction. Figures 9–11 show six representative scales spanning the frozen point, the smallest positive scale, the standard baseline, the warning reference, and two large initializations. The displayed QK gradient and weight RMS combine Q and K as the root mean square of their separately recorded RMS values; the relative update is the arithmetic mean of the recorded Q and K relative updates. Mechanism statistics are first averaged over layers and heads within each seed, then aggregated across seeds. Lines and bands are the three-seed mean and one standard deviation. Optimizer metrics are unsmoothed: the plots retain the first ten updates and then every 25 updates for AdamW and Muon or every 10 for SGD. The symlog update axis exposes these first steps together with the full training horizon. Exactly zero remains visible in the linear panels, but its zero-valued weight, gradient, and logit-variance traces cannot appear on logarithmic ordinates.

Scale-normalizing recipes. Under both AdamW and Muon, $\sigma_{qk} = 10^{-5}$ begins with a minute raw gradient but the largest relative QK updates. Its QK weight RMS grows by orders of magnitude, and both $\widehat{\gamma}/\ln T$ and $\widehat{U}$ rapidly leave their near-uniform initial values and approach the moderate-scale trajectories. At $\sigma_{qk} = 5$ and 10, the ordering reverses: pre-clipping gradients are initially enormous, but relative updates are small because the weights are already large. Those weight scales move only modestly and attention remains concentrated, accompanying slower loss reduction.

Plain-gradient control. Figure 11 removes scale normalization, clipping, and momentum. The 10^{-5} weights then remain essentially at their initial scale, with negligible change in $\widehat{\gamma}$ or $\widehat{U}$; this is the slow absolute-motion side of vanilla gradient descent. Conversely, $\sigma_{qk} = 10$ converts its large raw gradient directly into an enormous relative update and weight norm, and every seed becomes nonfinite at update 2. The full learning-rate grid in Figure 12 shows that this contrast is not an artifact of selecting $\eta = 0.05$: lowering the learning rate makes all scales learn slowly and leaves $\sigma_{qk} = 10$ with extreme loss, while raising it destabilizes the large-scale side first.

Together, the endpoint and trajectory evidence support a conditional practice rule: small but nonzero QK initialization is not a handicap for an optimizer that normalizes update scale, while plain SGD preserves the small raw-gradient bottleneck. Exactly zero is excluded from this recommendation because it is an absorbing frozen point.

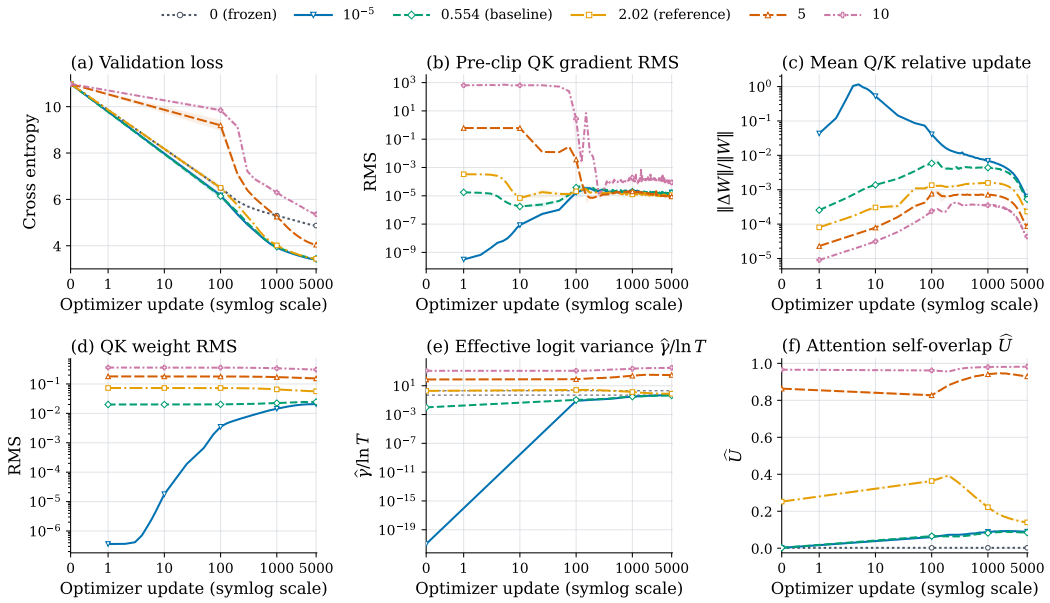

Figure 9: **AdamW turns a tiny nonzero initialization into rapid QK motion.** Validation loss, pre-clipping gradient RMS, actual relative update, weight RMS, effective logit variance, and attention self-overlap are shown over 5000 updates. The 10^{-5} trajectory rapidly grows toward the moderate-scale state, whereas $\sigma_{qk} = 5$ and 10 retain concentrated attention and higher loss. Horizontal lines in panel (e) mark $\widehat{\gamma}/\ln T = 1/2$ and 2.

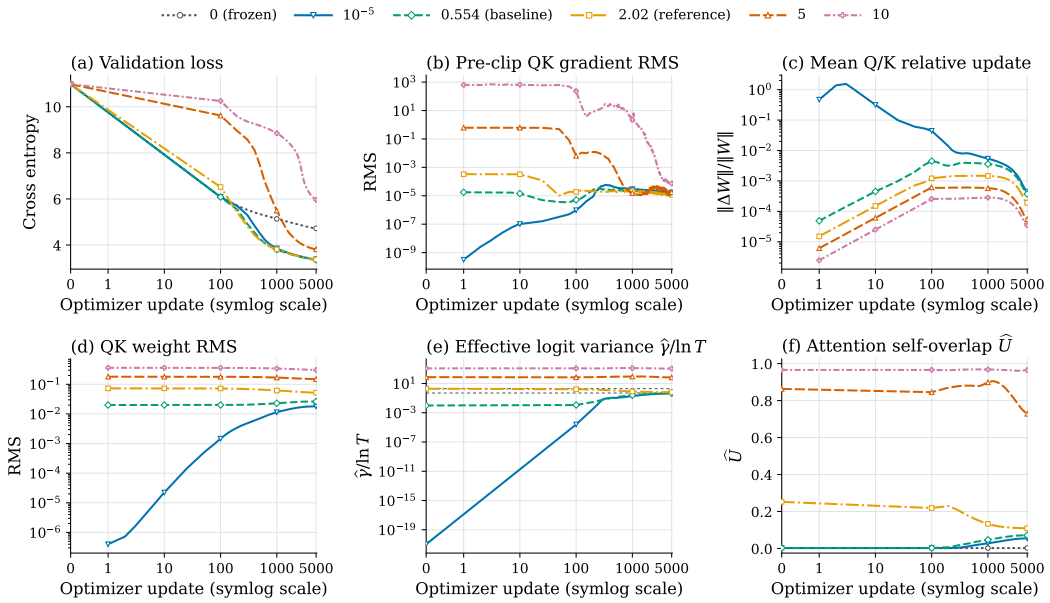

Figure 10: **Muon exhibits the same small–large asymmetry.** Despite its tiny initial raw gradient, $\sigma_{qk} = 10^{-5}$ receives the largest relative updates and rapidly changes its weight and attention state. Large initializations receive much smaller relative updates and remain concentrated; their validation losses improve more slowly and finish worse. Plotting and aggregation match Figure 9.

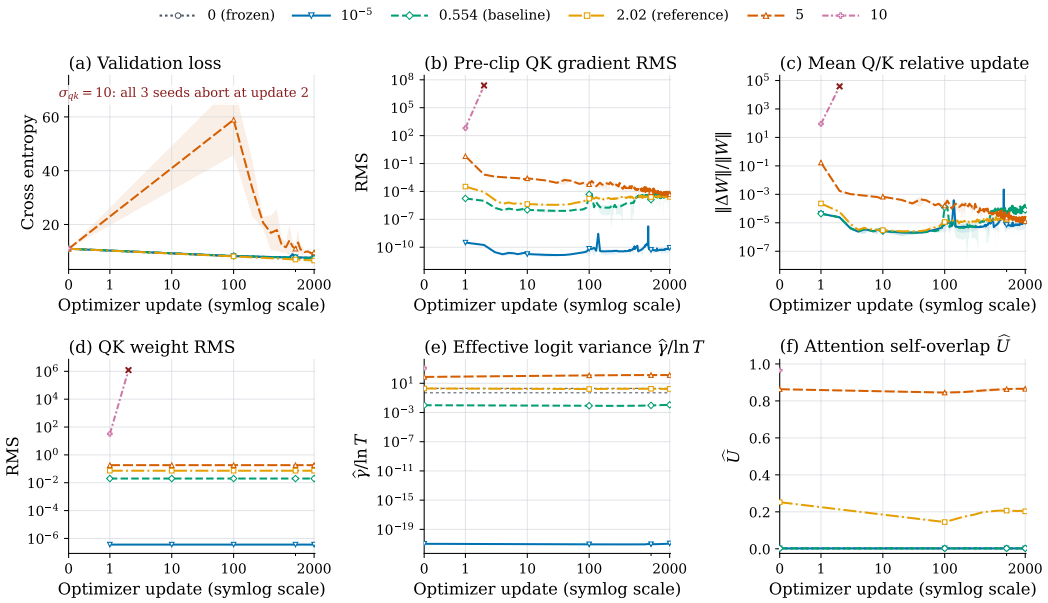

Figure 11: **Plain SGD exposes the unnormalized gradient mechanism.** At $\eta = 0.05$, the smallest positive QK weights and their attention state barely move over 2000 updates. Large initialization instead produces large absolute steps: all three $\sigma_{qk} = 10$ runs abort at update 2, marked by the terminal crosses in the optimizer panels. The frozen zero control remains distinct from the limit $\sigma_{qk} \to 0^+$.

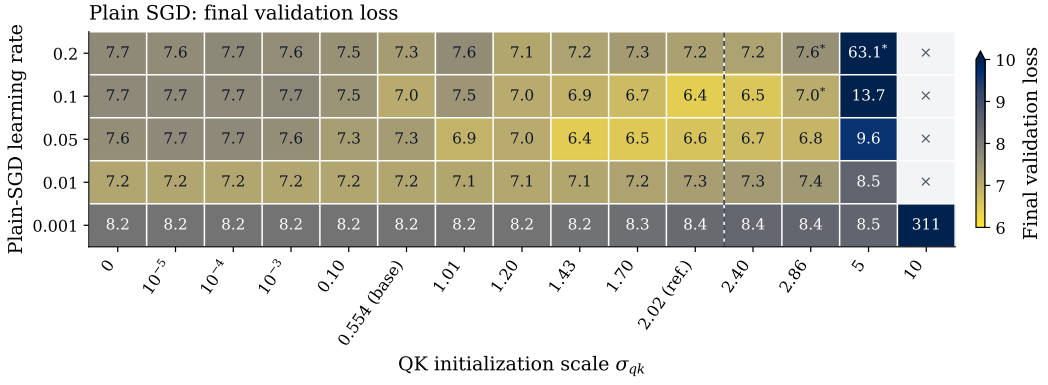

Figure 12: **The plain-SGD tradeoff persists across learning rates.** Cells report mean terminal validation loss over completed seeds for all 15 QK scales and five learning rates; colors saturate above loss 10 so that the useful range remains visible, while cell text gives the unsaturated value. An asterisk denotes two of three completed seeds and a cross denotes no completed seed. The dashed separator marks scales strictly above $\sigma_{qk,+} = (2 \ln 4096)^{1/4} \approx 2.02$.