DEMYSTIFYING MAXIMUM ATTENTION-LOGIT DYNAMICS IN MUON TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon has been widely adopted in language-model training due to its superior efficiency, yet the origin of its excess late-stage attention-logit growth remains unclear. Our work systematically investigates the maximum attention-logit dynamics in Muon training. First, by comparing different attention implementations and numerical precisions, we show that excess late-stage growth appears only with FlashAttention (FA) in Brain Floating Point 16 (BF16), indicating that it is not intrinsic to the optimizer design but originates from numerical errors. We further find that, after an early transient, the resulting gradient errors become persistently rank-one, with stable directions and consistent signs, causing them to accumulate coherently rather than cancel and producing structured high-logit heads. Second, motivated by this structure, we analyze a representative retrieval task with additive rank-one gradient errors and prove that the precision-induced logit gap is negligible in the early stage but becomes dominant later, directly linking the observed errors to the two-stage logit growth. Finally, we identify the FA intermediate quantities most responsible for this growth and mitigate it by selectively increasing their precision. Experiments show that these interventions provide a new trade-off between training performance and logit magnitude relative to existing methods.

1 INTRODUCTION

Muon has shown strong efficiency relative to Adam in Large Language Models (LLMs) pretraining and is increasingly adopted in large-scale training [1, 2]. By exploiting the matrix structure of parameters to normalize gradients, it improves the learning of tail information in pretraining data [3, 4] and achieves the same validation loss in nearly half the training time of Adam [5].

Despite its superior efficiency, Muon has been observed to exhibit excess attention-logit growth during late-stage LLM training [6]. QK-Clip mitigates this behavior by controlling the magnitude of the query and key parameters [6], but does not analyze its underlying cause. The mechanism driving this excess growth therefore remains unclear. In this work, we systematically investigate the maximum attention-logit dynamics in Muon training. Specifically, we ask the following questions:

1. *What is the primary source of the excess attention-logit growth observed in Muon training?*
2. *How can the effect of this source on attention-logit growth be characterized theoretically?*
3. *How can this mechanism guide a principled mitigation of the excess logit growth?*

To answer the first question, we compare the maximum attention-logit dynamics of Muon across attention implementations, including MATH and FlashAttention (FA), and precisions, including Brain Floating Point 16 (BF16) and 32-bit Floating Point (FP32). We find that late-stage excess logit growth appears only with low-precision FA, identifying its numerical errors as the primary source of the excess. We then analyze the resulting gradient discrepancies. After an early transient, these errors become approximately rank one, where the queries and keys across tokens concentrate around shared directions. Moreover, their signs and dominant directions remain stable across training, causing the errors to accumulate coherently rather than cancel. This structure predicts two signatures in high-logit heads: a common-mode lift, in which the minimum, mean, and maximum logits rise together, and column-selective attention concentrated on a small set of recurring, often frequent, tokens. Our experiments confirm both predictions.

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.

To answer the second question, we study the attention dynamics in a representative retrieval task, explicitly modeling the persistent rank-one numerical error observed above in low-precision training. By comparing the high- and low-precision dynamics, we prove that this structured error causes the late-stage excess maximum-logit growth. In the early stage, the relative maximum-logit discrepancy remains small, scaling at most as $n/(A+Bn)^2$ at step n for some constants $A, B > 0$. In contrast, in the late stage, the discrepancy grows quadratically while the high-precision maximum logit grows only logarithmically, yielding a relative ratio of order $n^2/\log n$. This analysis establishes a direct theoretical link between the observed rank-one gradient errors and the late-stage excess logit growth.

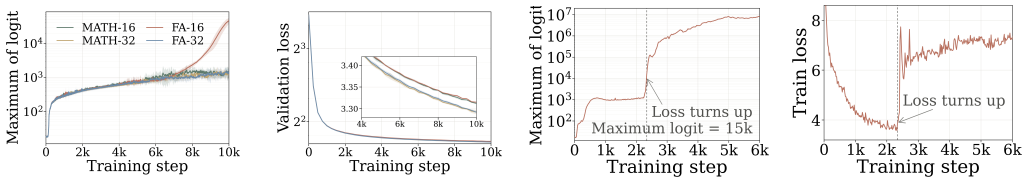
To answer the third question, we selectively increase the precision of the intermediate quantities identified above, revealing a new trade-off among maximum-logit control and validation performance. Precision ablations identify the attention output `O` as the dominant numerical-error source, with the attention probabilities `P` contributing a smaller additional effect. Retaining `O` in 16-bit Floating Point (FP16) or FP32 substantially suppresses excess logit growth, while retaining `P` in FP32 can provide further reduction. These kernel-level interventions offer a different trade-off between performance and logit magnitude from existing techniques such as QK-Clip [6] and QK-Norm [7].

In summary, we identify a clear mechanism for Muon's excess logit growth: low-precision FA induces persistent rank-one gradient errors that Muon coherently accumulates, driving excess maximum-logit growth and structured high-logit heads. Selectively increasing the precision of key intermediate quantities mitigates this growth while offering a new performance–logit trade-off.

2 PRELIMINARIES

Causal self-attention captures the token relationship within sequences by self-weighted sums. Let $x_i \in \mathbb{R}^d$, $i = 1, \ldots, T$, denote the hidden-state embedding at position i. Attention weights are obtained from scaled inner products of learned projections of these embeddings. First, each embedding is mapped to a query and key as $q_i = W_Q x_i + b_Q$ and $k_i = W_K x_i + b_K$, where $W_Q, W_K \in \mathbb{R}^{d_h \times d}$ and $b_Q, b_K \in \mathbb{R}^{d_h}$. Second, under Rotary Positional Embeddings (RoPE) [8], positional information is injected by rotating them to $R_i q_i$ and $R_i k_i$, where $R_i \in \mathbb{R}^{d_h \times d_h}$ is unitary. Specifically, writing $R_{ij} = R_i^\top R_j$, the attention logit and its scaled version are $L_{ij} = q_i^\top R_{ij} k_j$ and $S_{ij} = \tau L_{ij}$, respectively, where $\tau = d_h^{-1/2}$. The attention weights are normalized versions of S_{ij}, i.e., $P_{ij} = \exp(S_{ij}) / \sum_{\ell=1}^{T} \exp(S_{i\ell})$. Finally, the output is $O_i = \sum_{j=1}^{T} P_{ij} v_j$, where $v_j = W_V x_j + b_V$, $W_V \in \mathbb{R}^{d_h \times d}$, and $b_V \in \mathbb{R}^{d_h}$. During training, backpropagation through the attention scores gives $\mathrm{d}q_i = \tau \sum_{j \le i} \mathrm{d}S_{ij} R_{ij} k_j$ and $\mathrm{d}k_j = \tau \sum_{i \ge j} \mathrm{d}S_{ij} R_{ij}^\top q_i$, where $\mathrm{d}X$ denotes the gradient of the loss $\mathcal{L}$ with respect to X. For NoPE, the rotations are removed, i.e., $R_i = I$ and $L_{ij} = q_i^\top k_j$. For our later analysis, given a set of token sequences $\mathcal{B}$ and attention head h, we define the maximum attention logit as $\mathrm{Logit}^h_{\max}(\mathcal{B}) = \max_{b \in \mathcal{B}} \max_{1 \le j \le i \le T} L^h_{ij}(b)$, where $L^h_{ij}(b)$ denotes the attention logit at positions (i, j) for sample $b = (x_1, \cdots, x_T)$.

Attention implementations specify how attention is executed on hardware, with two common choices being MATH and FA. MATH materializes the score matrix S and probability matrix P in full, whereas the I/O-aware FA uses tiled computation and online softmax to avoid materializing these and maintain a small set of intermediate quantities [9, 10]. In low-precision FA, e.g., BF16, some selected intermediate quantities are rounded to the target precision, but softmax statistics and accumulations are always maintained in FP32. Specifically, in the forward pass, FA computes S and P tile-wise, maintaining the row-wise softmax statistics and output accumulator in FP32 and storing the resulting row-wise log-sum-exp values for backward. In this process, P is rounded before the low-precision PV multiplication, and the normalized accumulator is rounded when O is stored. The backward pass likewise proceeds tile-wise using the stored log-sum-exp values. With RoPE, let $\widetilde{Q}, \widetilde{K}, V, O \in \mathbb{R}^{T \times d_h}$ stack $(R_i q_i)^\top, (R_i k_i)^\top, v_i^\top$, and $O_i^\top$ row-wise. Then the backward pass first reconstructs P tile-wise and computes $\mathrm{d}P = \mathrm{d}O V^\top$ together with the row-wise softmax correction $D = \mathrm{rowsum}(\mathrm{d}O \circ O)$, where $\circ$ denotes element-wise multiplication. It then forms $\mathrm{d}S = P \circ (\mathrm{d}P - D\mathbf{1}^\top)$, where $\mathbf{1} \in \mathbb{R}^T$ is the all-ones vector. Finally, it propagates through the query–key products as $\mathrm{d}\widetilde{K} = \tau, \mathrm{d}S^\top \widetilde{Q}$ and $\mathrm{d}\widetilde{Q} = \tau, \mathrm{d}S \widetilde{K}$. Here, causal-masked entries of P and $\mathrm{d}S$ are zero. Rotating the gradients back gives $\mathrm{d}q_i = R_i^\top \mathrm{d}\widetilde{q}_i$ and $\mathrm{d}k_i = R_i^\top \mathrm{d}\widetilde{k}_i$, recovering the causal-attention gradients above. Suppressing the batch index, the corresponding parameter gradi-



(a) Maximum logit (b) Validation loss (c) Maximum logit in [12] (d) Training loss in [12]

Figure 1: Maximum attention logits and losses in successful FA-16 and collapsed runs. (a)–(b) Under RoPE, FA-16 departs from the intrinsic logit trajectory at late stages while maintaining validation loss comparable to FA-32 (five-seed average). (c)–(d) In the collapsed run from [12], collapse begins at $\mathrm{Logit}_{\max} \approx 1.5 \times 10^4$, before the subsequent extreme growth, showing that collapse and excess logit growth are distinct phenomena.

ents are $\mathrm{d}W_Q = \sum_{i=1}^{T} \mathrm{d}q_i x_i^\top$, and $\mathrm{d}W_K = \sum_{i=1}^{T} \mathrm{d}k_i x_i^\top$. For NoPE, setting $R_i = I$ reduces the above expressions to $\widetilde{Q} = Q$ and $\widetilde{K} = K$. See Appendix A for more related works.

Muon has recently emerged as a matrix-aware optimizer for large-scale LLMs training [5, 1]. It updates matrix parameters using a spectrally normalized gradient estimate. For a matrix parameter $W_t \in \mathbb{R}^{m\times n}$, let $G_t = \mathrm{d}W_t$ denote its backpropagated gradient. Muon first forms a momentum-based gradient estimate $M_t = \beta M_{t-1} + (1-\beta)G_t$, and $M_0 = 0$, where $\beta \in [0, 1)$. It then normalizes the nonzero singular values of M_t to one for the parameter update. In practice, Muon approximates this normalization using five-step Newton–Schulz map (NS5) as follows. With $(\omega_0, \omega_1, \omega_2)$ as $(3.4445, -4.7750, 2.0315)$ and $\varepsilon_{\mathrm{NS}} > 0$, initialize $X_0 = M_t/(\|M_t\|_F + \varepsilon_{\mathrm{NS}})$ and iterate $Y_r = X_r X_r^\top$, $X_{r+1} = \omega_0 X_r + (\omega_1 Y_r + \omega_2 Y_r^2)X_r$ for $r = 0, \ldots, 4$. We denote the resulting X_5 as $\mathrm{NS5}(M_t)$ and update $W_{t+1} = W_t - \eta_t\, \mathrm{NS5}(M_t)$, where η_t is the learning rate.

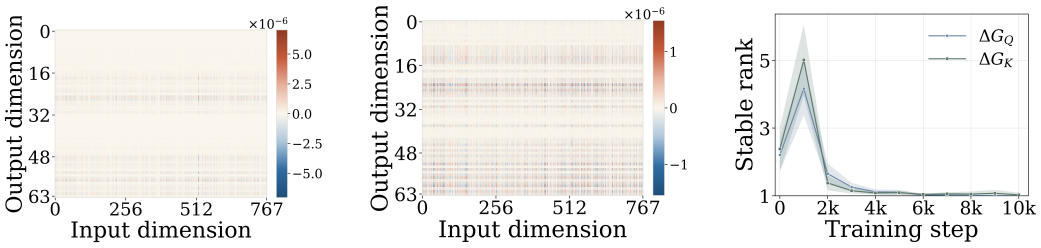
3 Empirical Analyses of Maximum Logit Growth

Experimental setting Throughout the paper, we train 124M- and 1B-parameter GPT-2 models on FineWeb [11] for 10,000 steps. For each model size, we sweep learning rates over $\{1, 2, 5\} \times \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ and select the best according to validation loss. Our main experiments consider RoPE, NoPE, and sinusoidal positional encoding, without additional query–key normalization unless otherwise specified. For clarity, the main text focuses on RoPE, with the remaining positional-encoding results reported in Appendix I.3. Additional experimental details are provided in Appendix H.

3.1 Excess Logit Growth Originates From Numerical Errors

To disentangle the growth related to optimizer design from implementation and numerical effects, we compare four matched settings combining MATH or FA attention with BF16 or FP32 training: MATH-16, MATH-32, FA-16, and FA-32. Among them, MATH-32 has the smallest numerical errors, as it incurs minimal quantization error and is free from the online-computation error introduced by FA. As shown in Figure 1(a), MATH-16, MATH-32, and FA-32 follow a common trajectory: their maximum logits increase rapidly from approximately 20 and soon enter an approximately linear growth regime, eventually reaching values on the order of 10^3. We refer to the trajectory of MATH-32, which is also shared by MATH-16 and FA-32, as the *intrinsic logit growth*, since it persists in the setting with minimal numerical error.

Against this intrinsic trajectory, FA-16, which combines FA with low-precision arithmetic and is commonly used in large-scale LLM training [6, 2], exhibits a clear two-stage behavior, as shown in Figure 1(a). Initially, its maximum attention logit is nearly indistinguishable from that of MATH-32; after approximately step 6,000, it departs from the intrinsic trajectory and accelerates to values on the order of 10^4. In addition to RoPE, the same two-stage behavior appears under sinusoidal positional encoding and NoPE, as shown in Appendix I.3. We call the additional growth of FA-16 above the intrinsic trajectory the *excess logit growth*. Its absence in MATH-16 and FA-32 rules out BF16 arithmetic and FA alone, respectively, suggesting that numerical errors specific to their low-precision combination contribute to the late-stage excess. Notably, this divergence in attention-logit dynamics is not mirrored in validation loss: FA-16 and MATH-16 have nearly identical validation-loss trajectories, as do FA-32 and MATH-32 (Figure 1(b)). This yields the following observation.



(a) ΔG_Q at step 6,000 (b) ΔG_K at step 6,000 (c) Stable rank

Figure 2: Empirically observed rank-one gradient errors in a representative head with excess logit growth. (a)–(b) Direct views of ΔG_Q and ΔG_K at step 6,000, both exhibiting a pronounced approximately rank-one structure. (c) The stable ranks of the two gradient errors approach one after an early transient and remain close to one thereafter.

> **Observation 1 (Intrinsic and excess logit growth).** With low-precision FA, the maximum attention-logit growth exhibits two stages: an initial regime that closely tracks the intrinsic logit growth, followed by late-stage excess logit growth associated with numerical errors specific to low-precision FA.

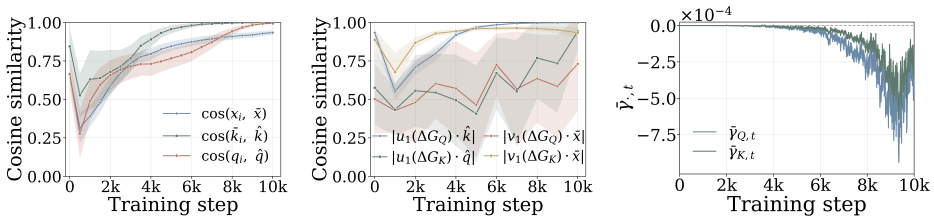
Having identified numerical errors in low-precision FA as the source of excess logit growth, we next ask whether this phenomenon is another manifestation of training collapse, a known consequence of FA numerical errors. We define collapse as a sharp increase in training loss that fails to recover or becomes non-finite [12]. The distinction is evident from both successful and collapsed runs. First, large attention logits need not cause collapse: in the RoPE FA-16 run, $\mathrm{Logit}_{\max}$ reaches approximately 4×10^4 at step 10,000 while the validation loss remains close to that of MATH-16, and further grows to 3.46×10^5 by step 20,000 without collapse (Figure 1(a)). Second, in the collapsed run of [12], collapse begins near step 2,340, when $\mathrm{Logit}_{\max} \approx 1.5 \times 10^4$, substantially below the 4×10^4 already reached without collapse in our successful run; logits near 10^7 appear only after the loss has failed (Figures 1(c) and 1(d)). Thus, although both can arise from numerical errors in low-precision FA, excess logit growth and training collapse are distinct and not tightly coupled.

3.2 Numerical Gradient Errors Are Consistently Low-rank

To understand how numerical errors contribute to logit growth, we next examine how the gradient errors induced by low-precision FA accumulate over training. In the following, for any quantity Z, let $\Delta Z = Z^{\mathrm{lo}} - Z^{\mathrm{hi}}$ denote the discrepancy between paired low-precision (BF16) and high-precision (FP32) FA evaluations of the same network input and model state, differing only in the precision of the FA kernels.

Empirically rank-one gradient errors. We begin by directly inspecting the gradient errors of the query and key parameters in the logit-exploding head. As shown in Figures 2(a) and 2(b), the gradient numerical errors of W_Q and W_K, denoted by $\Delta G_Q = \Delta \mathrm{d}W_Q$ and $\Delta G_K = \Delta \mathrm{d}W_K$, already exhibit a pronounced rank-one structure at step 6,000. This motivates us to examine whether this approximately rank-one structure persists throughout training once established. To quantify this, we track stable rank [13], $\mathrm{srank}(Z) = \|Z\|_F^2 / \|Z\|_2^2$. Figure 2(c) shows that the stable ranks of both ΔG_Q and ΔG_K approach one after an initial transient and remain close to one thereafter. Thus, although the numerical errors arise from many token-level operations within the attention kernel, their aggregate effect on the query and key gradients is dominated by a rank-one component.

Localizing the dominant numerical source. To understand what gives rise to the dominant rank-one component and excess logit growth, we identify the primary source of numerical error through controlled precision ablations that selectively retain intermediate quantities in FP32, as detailed in Appendix B.1. The ablations show that preserving the attention probabilities P used in the PV multiplication *alone* leaves the excess logit growth largely unchanged, whereas preserving the resulting attention output O in higher precision suppresses most of the excess. Combining the two interventions removes the remaining excess and brings the trajectory close to the full-precision reference. These results identify the low-precision storage of O as the dominant error source, while the pre-



(a) Directional concentration (b) Directional alignment (c) Sign consistency

Figure 3: Verification of the rank-one gradient-error structure and sign consistency. (a) Token-dependent input, key, and query factors concentrate around their mean directions. (b) The leading singular vectors of ΔG_Q and ΔG_K align with these directions. (c) The gradient-error sign remains consistent throughout training.

cision of P contributes a smaller additional error that further promotes logit growth. We therefore focus in what follows on how the rounding error in O propagates into the query and key gradients.

Output rounding induces a sum of rank-one terms in the gradient error. Consider paired BF16 and FP32 evaluations with identical $Q, K, V, P, \mathrm{d}O$, and input features, differing only in the stored attention output O due to rounding. As shown in Section 2, the output discrepancy first enters the backward pass through the row-wise correction, $\Delta D_i = \langle \mathrm{d}O_i, \Delta O_i \rangle$, and consequently induces $\Delta \mathrm{d}S_{ij} = -P_{ij}\Delta D_i$. For RoPE, defining the attention-weighted key $\overline{k}_i = \sum_{j \le i} P_{ij} R_{ij} k_j$, the resulting query- and key-gradient discrepancies are sums of rank-one matrices as follows.

$$\Delta G_Q = -\tau \sum_{i=1}^{T} \Delta D_i \, \overline{k}_i x_i^\top, \qquad \Delta G_K = -\tau \sum_{i=1}^{T} \Delta D_i \sum_{j \le i} P_{ij} (R_{ij}^\top q_i) x_j^\top. \tag{1}$$

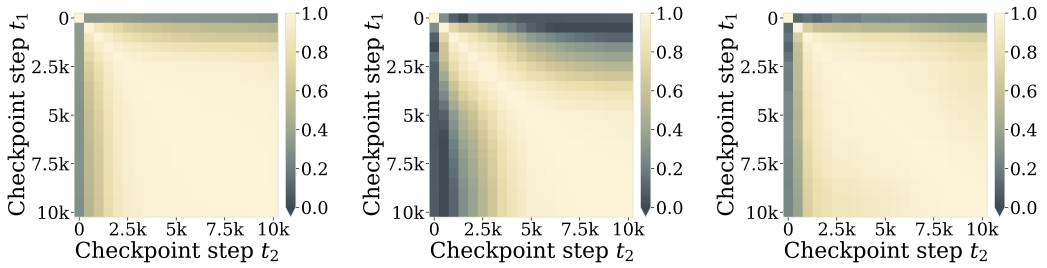
A detailed derivation is given in Proposition 1 in Appendix B.2. For NoPE, the same expressions follow by setting $R_{ij} = I_{d_h}$. Thus, the rounding error in O propagates into both parameter gradients as weighted sums of rank-one outer products.

Mean directions explain the rank-one structure. A natural hypothesis for the near-rank-one structure of ΔG_Q and ΔG_K is that the rank-one components in Eqn. 1 concentrate around common "mean" directions. To test this hypothesis, we first define these directions. Given a token sequence $(x_1, \ldots, x_T)$, let $\mu_x = T^{-1} \sum_{i=1}^{T} x_i$ and define the normalized mean feature direction as $\overline{x} = \mu_x / \|\mu_x\|_2$, together with $\mu_q = W_Q \mu_x + b_Q$ and $\mu_k = W_K \mu_x + b_K$. For RoPE, let $\overline{R} = T^{-1} \sum_{i=1}^{T} R_i$ denote the average rotation over the sequence positions, and define $\widehat{q} = \overline{R}^\top \mu_q / \|\overline{R}^\top \mu_q\|_2$ and $\widehat{k} = \overline{R} \mu_k / \|\overline{R} \mu_k\|_2$. For NoPE, this reduces to $\overline{R} = I_{d_h}$. The vectors $\overline{x}$, $\widehat{q}$, and $\widehat{k}$ therefore provide predicted mean directions for the token-dependent factors in Eqn. 1. Figures 3(a) and 3(b) verify this concentration: the token-dependent factors exhibit high average cosine similarity with $\overline{x}$, $\widehat{k}$, and $\widehat{q}$, while the leading singular directions of ΔG_Q and ΔG_K align with the corresponding predicted directions. This motivates the following approximation.

$$\Delta G_Q = \gamma_Q \widehat{k} \overline{x}^\top + R_Q, \; \Delta G_K = \gamma_K \widehat{q} \overline{x}^\top + R_K, \; \gamma_Q = \langle \Delta G_Q, \widehat{k} \overline{x}^\top \rangle_F, \; \gamma_K = \langle \Delta G_K, \widehat{q} \overline{x}^\top \rangle_F, \tag{2}$$

where R_Q and R_K are the Frobenius-orthogonal residuals of ΔG_Q and ΔG_K with respect to their respective rank-one components.

Consistent push across steps. To determine whether these errors accumulate coherently across steps, we track the signs of γ_Q and γ_K together with the temporal stability of $\widehat{k}$, $\widehat{q}$, and $\overline{x}$. At each low-precision training step, we compute a paired high-precision gradient, evaluate $\gamma_{c,t}$ from Eq. 2, and track its momentum average $\overline{\gamma}_{c,t} = \beta \overline{\gamma}_{c,t-1} + (1-\beta)\gamma_{c,t}$ for $c \in Q, K$, initialized at zero using the Muon momentum coefficient β. Figure 3(c) shows that both momentum averages remain consistently negative throughout training, while the cosine similarities in Figure 4 show that $\overline{x}$, $\widehat{k}$, and $\widehat{q}$ vary slowly after the early transient. Thus, the gradient errors repeatedly push along nearly fixed rank-one directions with consistent signs.



(a) Input direction (b) Key direction (c) Query direction

Figure 4: Pairwise cosine similarities of (a) $\overline{x}$, (b) $\widehat{k}$, and (c) $\widehat{q}$ evaluated between different training steps. Lighter cells indicate higher similarity. These directions remain stable after the initial transient.

Observation 2 (Structured numerical error). After an initial transient of a few thousand steps, the low-precision FA errors in the query and key parameter gradients develop a persistent signed component along slowly varying rank-one directions. Momentum accumulates these errors coherently, producing a consistent push along query–key directions.

3.3 Low-Rank Gradient Errors Induce Structured High-Logit Heads

Since the persistent low-rank gradient errors in Section 3.2 align with the concentrated input direction $\overline{x}$, we expect two characteristic structures in heads with large maximum logits. First, because token representations share a strong common component along $\overline{x}$, these errors should induce a common upward shift across tokens, causing the minimum, mean, and maximum logits to grow together. Second, frequent tokens that contribute strongly to $\overline{x}$ may exhibit more pronounced logit and attention patterns. We next test these predictions empirically.

Common-mode logit lift. We test predicted common shifts by tracking the minimum, mean, and maximum attention logits of a representative high-logit head in Figure 5. After the initial transient, all three statistics rise together to the same order of magnitude. Thus, the growing maximum reflects a broad upward displacement of the score matrix rather than an isolated extreme produced by a small number of query–key pairs. We refer to this shared shift as a *common-mode lift*.

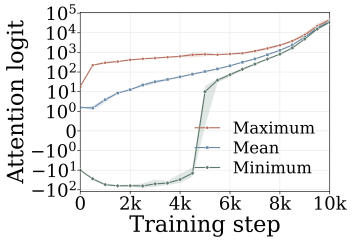


Figure 5: Logit statistics of the representative largest-logit head.

Column-selective attention. To examine the predicted token-dependent structure, we inspect the attention map of the same head and aggregate attention mass by token identity. Figure 6 shows pronounced vertical stripes, indicating that many queries attend to a small set of key tokens. In this representative RoPE head, approximately 40% of the attention mass is assigned to "`.`", 32% to "$\backslash n$", and 17% to "`,`", together accounting for nearly 90% of the total attention mass. Although the dominant token identities vary with context, the inspected examples consistently show attention concentrated on a few key columns, often corresponding to frequent tokens. This pattern shows substantial overlap with the corpus-level token-frequency ranking, as detailed in Appendix I.1.

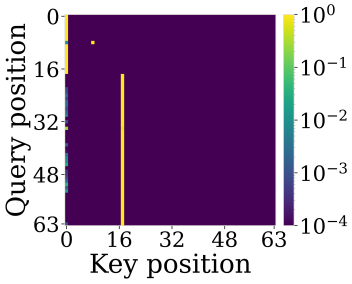


Figure 6: Attention map of the representative head at step 10,000.

Observation 3 (Common lift with selective attention). The largest-logit head exhibits a broad upward shift in its attention logits, while its normalized attention remains concentrated on a small set of key positions and tokens.

4 THEORETICAL ANALYSES OF MAXIMUM LOGIT GROWTH

The previous section empirically links excess attention-logit growth to persistent structured gradient errors induced by low-precision FA. In this section, we prove in a representative model that these errors *cause* the excess logit growth observed in the second growth stage of Section 3.

We study a representative retrieval task in which the model retrieves information from context tokens paired with the query token. Specifically, let $\mathcal{C} = \{c_1, \dots, c_M\}$ and $\mathcal{Z} = \{z_1, \dots, z_M\}$ denote paired context-token and query-token embeddings, where $c_m, z_m \in \mathbb{R}^d$ for $m \in [M]$. A prompt $(x_1, \dots, x_T, x_{T+1})$ of length $T+1$ is sampled as follows. First, the query token x_{T+1} is sampled uniformly from $\mathcal{Z}$. Let m denote its index, so that $x_{T+1} = z_m$. The T context tokens $x_1, \dots, x_T$ are then sampled i.i.d. uniformly from $\mathcal{C}$, conditioned on the event that at least one of them equals c_m. The target output is Wc_m, where $W \in \mathbb{R}^{d_h \times d}$ extracts the relevant information from the paired context embedding. Thus, given a prompt $(x_1, \dots, x_T, x_{T+1})$ with query $x_{T+1} = z_m$, the model is trained to retrieve the corresponding context token c_m and output Wc_m.

For ease of analysis and to highlight the main tension, we consider a NoPE attention with trainable $W_Q, W_K \in \mathbb{R}^{d_h \times d}$ and a fixed language-model head $F \in \mathbb{R}^{d_h \times M}$. Given the prompt $(x_1, \dots, x_T, x_{T+1})$, the attention first computes $L_j = L_{T+1,j} = (W_Q x_{T+1})^\top (W_K x_j)$ and then the attention weights $P_j = \exp(\tau L_j)/\sum_{l=1}^{T} \exp(\tau L_l)$, where $\tau = d_h^{-1/2}$. Finally, the attention output is $o = \sum_{j=1}^{T} P_j W_V x_j$, where $W_V \in \mathbb{R}^{d_h \times d}$ is fixed to W for simplification. Define $f_m = Wc_m$ for $m \in [M]$. Since the attention is targeted to output Wc_m, given $x_{T+1} = z_m$, the language model head is fixed as $F = [f_1, \cdots, f_M]$, and the cross entropy loss is $\mathcal{L}(W_Q, W_K) = \mathbb{E}[\mathbb{E}[-\log(\exp(\lambda f_m^\top o)/\sum_{l=1}^{M} \exp(\lambda f_l^\top o))|x_{T+1} = z_m]]$, where the expectation is over the prompt distribution of $x_1, \cdots, x_{T+1}$, and $\lambda > 0$ is the reverse temperature.

To study the precision-dependent dynamics of maximum logits and loss, we optimize $W_Q^{\mathfrak{b}}$ and $W_K^{\mathfrak{b}}$ under two precision regimes indexed by $\mathfrak{b} \in \{\mathrm{hi}, \mathrm{lo}\}$, using corresponding high- and low-precision gradient oracles. At step n, the oracle for $W_{Q,n}^{\mathfrak{b}}$ returns $G_{Q,n}^{\mathfrak{b}} = \nabla_{W_Q} \mathcal{L}_n^{\mathfrak{b}} - \mathbf{1}\{\mathfrak{b} = \mathrm{lo}\} \cdot \Gamma E_{Q,n}$, where $\nabla_{W_Q} \mathcal{L}_n^{\mathfrak{b}} \in \mathbb{R}^{d_h \times d}$ is the exact gradient evaluated at $(W_{Q,n}^{\mathfrak{b}}, W_{K,n}^{\mathfrak{b}})$, $\Gamma > 0$ controls the scale of the low-precision error, and $E_{Q,n} \in \mathbb{R}^{d_h \times d}$ denotes its direction. We consider Muon with fixed learning rate $\eta > 0$, zero momentum, and zero weight decay, updating $W_{Q,n+1}^{\mathfrak{b}} = W_{Q,n}^{\mathfrak{b}} - \eta\, \mathrm{NS5}(G_{Q,n}^{\mathfrak{b}})$, where NS5 is defined in Section 2. The update for W_K is defined analogously. The parameters are initialized as $W_{Q,0}^{\mathfrak{b}}$ and $W_{K,0}^{\mathfrak{b}}$.

We next impose several assumptions to capture the practical dynamics observed in Section 3, with their consistency with empirical observations verified in Appendix C.

Assumption 1 (Concentrated token embeddings)**.** *Token embeddings are concentrated around a shared direction $\overline{x}$: $c_m = \sigma_{\mathrm{com}} \overline{x} + \sigma_{\mathrm{type}} e_{c,m}$ and $z_m = \sigma_{\mathrm{com}} \overline{x} + \sigma_{\mathrm{type}} e_{z,m}$, where $\overline{x}$ is a unit vector and $\sigma_{\mathrm{com}}, \sigma_{\mathrm{type}} > 0$. The type-specific features satisfy $\overline{x}^\top e_{c,m} = 0$ and $e_{c,m}^\top e_{c,m'} = \mathbf{1}\{m = m'\} - 1/M$, with $e_{z,m}$ satisfying the same conditions. Accordingly, we assume $W\overline{x} = 0$ and $f_m^\top f_{m'} = \mathbf{1}\{m = m'\} - 1/M$ for $f_m = Wc_m$.*

This assumption captures the concentration of token embeddings observed in Section 3.2. Specifically, embeddings of different tokens share a common component, while their type-specific components are nearly orthogonal; the extracted information lies outside the shared component. The correction term $1/M$ is introduced for analytical convenience, and our results readily extend to the setting with exactly orthogonal type-specific components.

Assumption 2 (Task-aligned initialization)**.** *Let $\Phi_Q, \Phi_K \in \mathbb{R}^{d_h \times d}$ be any matrices specifying the initial query–key geometry: $\Phi_Q^\top \Phi_Q = \sum_{m=1}^{M} e_{z,m} e_{z,m}^\top$, $\Phi_K^\top \Phi_K = \sum_{m=1}^{M} e_{c,m} e_{c,m}^\top$, and $\Phi_Q e_{z,m} = \Phi_K e_{c,m}$ for all $m \in [M]$. We assume that parameters are initialized as $W_{Q,0}^{\mathfrak{b}} = u_0 \Phi_Q$ and $W_{K,0}^{\mathfrak{b}} = u_0 \Phi_K$, where $u_0 > 0$ denotes the initialization scale.*

The first two conditions correspond to an initialization that is isometric on the query- and context-type feature subspaces, since the two sums are the corresponding orthogonal projectors under Assumption 1 (Lemma 2). The condition $\Phi_Q e_{z,m} = \Phi_K e_{c,m}$ further encodes the correspondence between matched query and context tokens. This structure is motivated by the approximately projector-

like query and key initialization observed in experiments and, together with the other assumptions, confines the parameter dynamics to a low-dimensional subspace, simplifying the analysis.

Assumption 3 (Persistent rank-one error)**.** *The numerical errors are rank one: $E_{Q,n} = \widehat{k}\overline{x}^\top$ and $E_{K,n} = \widehat{q}\overline{x}^\top$ for all $n \geq 0$, where $\widehat{k}, \widehat{q} \in \mathbb{R}^{d_h}$ are unit error vectors. We further assume $\widehat{k}$ and $\widehat{q}$ are aligned, i.e., $\chi = \widehat{k}^\top \widehat{q} > 0$, and are orthogonal to the column spaces of Φ_Q and Φ_K respectively.*

This assumption idealizes the approximately rank-one, slowly varying gradient errors observed in Section 3.2 as coherent errors along fixed directions, capturing their accumulation during training. The condition $\chi > 0$ ensures that the query and key error directions contribute positively to logit growth. The orthogonality to Φ_Q and Φ_K is motivated by the weak coupling between excess logit growth and task performance observed in Section 3, together with the verification in Appendix C.

Before stating the results, we define several quantities. To characterize the loss dynamics, let $\mathcal{L}_n^{\mathfrak{b}}$ denote the loss at step n under the $\mathfrak{b}$-precision oracle, and define the optimal loss value as $\mathcal{L}^\star = \inf_{W_Q, W_K \in \mathbb{R}^{d_h \times d}} \mathcal{L}(W_Q, W_K)$. For the logit comparison, let $\mathrm{Logit}_{\max,n}^{\mathfrak{b}}$ denote the maximum attention logit at step n under precision $\mathfrak{b} \in \{\mathrm{hi}, \mathrm{lo}\}$, and set $\Delta_n = \mathrm{Logit}_{\max,n}^{\mathrm{lo}} - \mathrm{Logit}_{\max,n}^{\mathrm{hi}}$. We further define $\mathcal{G}_0 = \|\nabla_{W_Q} \mathcal{L}_0^{\mathrm{hi}}\|_F$, $\kappa_{\mathrm{com}} = \sigma_{\mathrm{com}}^2 \chi$, and $\kappa_\Gamma = \psi_5(\Gamma/(\Gamma + \varepsilon_{\mathrm{NS}}))$, where $\psi_5 = \psi^{\circ 5}$ for $\psi(t) = \omega_0 t + \omega_1 t^3 + \omega_2 t^5$ with $(\omega_0, \omega_1, \omega_2) = (3.4445, -4.7750, 2.0315)$.

Theorem 1. *Under Assumptions 1, 2, and 3, the following statements hold.*

(i) ***Convergence to $\mathcal{L}^*$.*** *At both precisions, the population loss $\mathcal{L}_n^{\mathfrak{b}}$ decreases to $\mathcal{L}^*$ as $n \to \infty$.*

(ii) ***Early logit overlap.*** *For any task-gradient level $\mathcal{G}_{\mathrm{E}} \in (0, \mathcal{G}_0)$, there exist $\eta_{\mathrm{E}}, C_{\mathrm{E}} > 0$ such that for every $0 < \eta \leq \eta_{\mathrm{E}}, \Gamma > 0$, and its corresponding step threshold $N_{\mathrm{E}} = N_{\mathrm{E}}(\mathcal{G}_{\mathrm{E}}, \eta) \geq 1$, we have that $|\Delta_{n+1} - \Delta_n| \leq C_{\mathrm{E}} \Gamma^2 \eta$ for $0 \leq n < N_{\mathrm{E}}$, and $|\Delta_n| \leq C_{\mathrm{E}} \Gamma^2 \eta n$ for $0 \leq n \leq N_{\mathrm{E}}$. Moreover, $|\Delta_n| / \mathrm{Logit}_{\max,n}^{\mathrm{hi}} \leq C_{\mathrm{E}} \Gamma^2 \eta n / (u_0 + \eta n)^2$ for $0 \leq n \leq N_{\mathrm{E}}$.*

(iii) ***Late logit separation.*** *As $n \to \infty$, $\mathrm{Logit}_{\max,n}^{\mathrm{hi}} = O(\log n)$, $\Delta_n = \kappa_{\mathrm{com}} \kappa_\Gamma^2 (\eta n)^2 + o((\eta n)^2)$, and $\Delta_{n+1} - \Delta_n = 2\kappa_{\mathrm{com}} \kappa_\Gamma^2 \eta^2 n + o(\eta^2 n)$. In particular, $\Delta_n = \Theta(n^2)$ and $\Delta_{n+1} - \Delta_n = \Theta(n)$.*

The proof is given in Appendix D.3, which also provides the explicit definition of N_{E}. For the loss, both precision regimes converge to the optimal loss $\mathcal{L}^\star$, consistent with the comparable validation losses of FA-16 and FA-32 in Figure 1(b). For the maximum logit, the relative discrepancy remains small in the early stage: $|\Delta_n| / \mathrm{Logit}_{\max,n}^{\mathrm{hi}}$ grows at most linearly with n. In contrast, in the late stage, this relative discrepancy can grow as $\Omega(n^2 / \log n)$, substantially faster than in the early stage. This theoretical separation matches the two-stage behavior observed in Section 3.1, showing that the low-rank gradient errors identified in Section 3.2 can drive the observed maximum-logit dynamics. More experimental results of our representative model are in Appendix G.

5 Mitigating the Precision-Dependent Logit Growth

The previous sections identify low-precision numerical errors as a major source of excess logit growth. We now investigate whether selectively increasing precision can mitigate this growth.

Methods and baselines. Numerical errors from different intermediate quantities contribute unequally to logit growth. The ablations in Appendix B.1 identify low-precision storage of the attention output O as the dominant contributor, with the precision of P having a smaller additional effect. This suggests a direct mitigation strategy: selectively increasing precision along these error-producing paths within the FA kernel. For the attention output, `O32` retains an FP32 copy of O, while `O16` retains an FP16 copy as a lower-cost alternative that reduces rounding error relative to BF16 when values remain within the FP16 representable range. For the attention probabilities, `P32` retains P in FP32 for the PV multiplication and can be combined with either output-storage modification. We compare these methods with QK-Norm [7], which normalizes queries and keys before scoring, and QK-Clip [6], which rescales the query–key projection weights when a head's maximum score exceeds a prescribed threshold. We also tested the low-precision stabilization method of Qiu and Yao [12], which modifies the internal FA computation; however, it frequently encountered numerical issues in our setting and is therefore excluded from the main comparison. Experimental details are provided in Appendix H.

Table 1: Mitigation results under NoPE at step 10,000. Mean ± std over five seeds.

Setting	$\text{Logit}_{\max}$ (↓)	Val. loss (↓)
Vanilla FA-16	156672 ± 2896	3.3915 ± 0.0012
O32	3648 ± 950	3.3681 ± 0.0004
O16	3776 ± 1267	3.3682 ± 0.0006
O32+P32	1284 ± 6	3.3667 ± 0.0004
O16+P32	1296 ± 10	3.3676 ± 0.0009
QK-Clip	4024 ± 11	3.3683 ± 0.0018
QK-Norm	64 ± 0	3.3825 ± 0.0010

Table 2: Mitigation results under RoPE at step 10,000. Mean ± std over five seeds.

Setting	$\text{Logit}_{\max}$ (↓)	Val. loss (↓)
Vanilla FA-16	46720 ± 9382	3.3150 ± 0.0017
O32	1416 ± 158	3.3146 ± 0.0042
O16	1099 ± 107	3.3145 ± 0.0024
O32+P32	996 ± 40	3.3120 ± 0.0003
O16+P32	1330 ± 153	3.3146 ± 0.0016
QK-Clip	802 ± 3	3.3155 ± 0.0024
QK-Norm	62 ± 0	3.3251 ± 0.0244

Training performance and logit magnitude comparison. We evaluate these methods on 124M and 1B models under RoPE and NoPE. Tables 1 and 2 report the 124M results, where all methods substantially suppress excess logit growth. Across both positional encodings, QK-Norm achieves the smallest maximum logit but the highest validation loss. Under NoPE, O16 and O32 reduce $\text{Logit}_{\max}$ from about 1.6×10^5 to below 4×10^3, below QK-Clip while achieving comparable validation loss; adding P32 further reduces it to approximately 1.3×10^3 and improves validation loss. Under RoPE, O16 and O32 reduce $\text{Logit}_{\max}$ by more than an order of magnitude, with lower validation losses than QK-Clip despite larger maximum logits. Notably, O32+P32 consistently improves upon O32 on both metrics under both positional encodings. Overall, these methods provide a distinct trade-off between validation performance and logit magnitude. Results for the 1B model in Appendix I.2 show consistent trends.

Computation and memory cost. To quantify the cost of the proposed methods, we benchmark them on an A40 with batch size 1,024, sequence length 1,024, and attention with 16 heads, dimension 64. Table 3 reports runtime and peak-memory overhead relative to Vanilla FA. Among these methods, O16 is the lowest-cost variant, adding only 0.6% total kernel time and 5.9% peak memory while suppressing most excess growth. O32 incurs modest overheads of 3.4% in runtime and 11.7% in peak memory. Adding P32 further increases the forward cost, but keeps total runtime overhead below 8% for both combinations. Overall, these results reveal an additional cost–effectiveness trade-off: higher precision along the numerical-error path yields lower maximum logits and improved validation loss at the cost of increased computation. In particular, O16 provides a low-cost operating point, while O32+P32 offer higher-precision alternatives at progressively greater cost.

Table 3: Runtime and peak-memory increases relative to Vanilla FA. Mean ± std over five runs.

Variant	Forward time	Backward time	Total time	Peak memory
O32	+5.8±0.2%	+2.7±0.2%	+3.4±0.1%	+11.7±0.0%
O16	+1.8±0.2%	+0.2±0.2%	+0.6±0.1%	+5.9±0.0%
O32+P32	+23.8±0.1%	+2.6±0.2%	+7.6±0.2%	+11.7±0.0%
O16+P32	+21.9±0.2%	+0.1±0.2%	+5.3±0.1%	+5.9±0.0%

6 Conclusion

Our work identifies numerical errors in low-precision FA as the primary source of excess late-stage attention-logit growth in Muon training. These errors are approximately rank one, with stable directions and consistent signs, causing them to accumulate coherently and produce structured high-logit heads. In a representative retrieval task, we prove that persistent rank-one errors lead to early logit overlap followed by late separation, providing theoretical support for the observed dynamics. Finally, we identify a new trade-off among validation performance, logit magnitude, and computational and memory costs: higher-precision storage of the attention output substantially suppresses excess growth, while retaining attention probabilities at higher precision in the forward pass provides further mitigation with additional compute and memory overhead. One limitation of our work is that our experiments focus on LLMs; whether these interventions are also effective in multimodal models remains to be explored.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We have not used generative AI to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, support qualitative and thematic data analysis, design or provide feedback on our research methodology or experiments, or interpret results. Furthermore, using AI tools to generate synthetic datasets, or to clean and reformat datasets, is not applicable to this work. Additionally, we used generative AI tools to edit this research paper to improve readability. We have reviewed all AI-assisted work. All AI-assisted translations and textual edits to add readability were manually reviewed and verified by the authors to guarantee the accuracy of our technical contributions and to preserve the original meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies numerical behaviors in large-scale neural network training and does not involve human subjects, personal data, user-generated content, or sensitive information. All experiments are conducted on publicly available datasets and standard language modeling benchmarks. We do not release or analyze any data containing personally identifiable information. Our study focuses on understanding and mitigating numerical errors in attention computation. The proposed analyses and experiments are intended to improve the reliability and understanding of machine learning systems. We are not aware of any foreseeable ethical concerns arising from the methods or findings presented in this work.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce both our empirical and theoretical results. Section 3 and 5 specifies the model configurations, training settings, precision variants, and evaluation protocols used throughout the experiments, while the appendix H provides additional implementation details, hyperparameters, ablation settings, and descriptions of the data and preprocessing procedures. For the theoretical results, all assumptions are stated explicitly and complete proofs of the main claims are provided in section 4 and appendix D. The supplementary materials contain the code required to run the experiments, including the training pipeline, numerical-precision interventions and logging procedures. Together, these materials are intended to enable independent reproduction of the results presented in this work.

REFERENCES

[1] 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.

[2] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

[3] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Tan. Muon outperforms adam in tail-end associative memory learning. In *International Conference on Learning Representations*, volume 2026, pages 37382–37419, 2026.

[4] Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. *arXiv preprint arXiv:2603.26554*, 2026.

[5] 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.

[6] Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[7] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pages 7480–7512. PMLR, 2023.

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

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

[10] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pages 35549–35562, 2024.

[11] 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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL `https://openreview.net/forum?id=n6SCkn2QaG`.

[12] Haiquan Qiu and Quanming Yao. Why low-precision transformer training fails: An analysis on flash attention. *arXiv preprint arXiv:2510.04212*, 2025.

[13] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21–es, 2007.

[14] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

[15] Xianyan Jia, Shutao Song, Wei He, Yangzihao Wang, Haidong Rong, Feihu Zhou, Liqiang Xie, Zhenyu Guo, Yuanzhou Yang, Liwei Yu, et al. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. *arXiv preprint arXiv:1807.11205*, 2018.

[16] Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019.

[17] Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, et al. Fp8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022.

[18] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

[19] Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. Scaling fp8 training to trillion-token llms. In *International Conference on Learning Representations*, volume 2025, pages 98631–98644, 2025.

[20] Liu Hanqing, Jianjun Cao, Yuanze Li, and Zijian Zhou. Grokking or glitching? how low-precision drives slingshot loss spikes. *arXiv preprint arXiv:2605.06152*, 2026.

[21] Alicia Golden, Samuel Hsia, Fei Sun, Bilge Acun, Basil Hosmer, Yejin Lee, Zachary DeVito, Jeff Johnson, Gu-Yeon Wei, David Brooks, et al. Is flash attention stable? *arXiv preprint arXiv:2405.02803*, 2024.

[22] Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. In *International Conference on Learning Representations*, volume 2026, pages 17389–17416, 2026.

[23] Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

[24] Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

[25] Zhehang Du and Weijie Su. The newton-muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

[26] Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. In *International Conference on Learning Representations*, volume 2026, pages 138323–138360, 2026.

[27] Ruijie Zhang, Yequan Zhao, Ziyue Liu, Zhengyang Wang, Dongyang Li, Yupeng Su, Sijia Liu, and Zheng Zhang. Teon: Tensorized orthonormalization beyond layer-wise muon for large language model pre-training. *arXiv preprint arXiv:2601.23261*, 2026.

[28] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[29] Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

[30] Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

[31] Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

[32] Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in muon. *arXiv preprint arXiv:2601.13474*, 2026.

[33] Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. In *International Conference on Learning Representations*, volume 2026, pages 45672–45718, 2026.

[34] Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. *arXiv preprint arXiv:2602.05725*, 2026.

[35] Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations*, volume 2026, pages 142276–142311, 2026.

[36] Tianyu Ruan, Fengzhuo Zhang, Shuche Wang, and Shihua Zhang. Muon learns more robust and transferable features than adam. *arXiv preprint arXiv:2606.09658*, 2026.

[37] Mitchell Wortsman, Peter Liu, Lechao Xiao, Katie Everett, Alexander Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, et al. Small-scale proxies for large-scale transformer training instabilities. In *International Conference on Learning Representations*, volume 2024, pages 49844–49869, 2024.

[38] 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 *International conference on machine learning*, pages 40770–40803. PMLR, 2023.

[39] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36:75067–75096, 2023.

[40] Suvadeep Hajra. Dense local dependencies induce attention-logit explosion and training instability during long-sequence transformer training, 2026. URL `https://arxiv.org/abs/2505.15548`.

[41] Shuxiao Xie, Shuyang Xie, Yuan Cao, Dezhi Ran, Wei Yang, and Tao Xie. One qk channel, many sources: Guarding low-precision attention collapse. *arXiv preprint arXiv:2608.02091*, 2026.

[42] 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*, pages 4246–4253, 2020.

[43] Ben Anson and Laurence Aitchison. Controlling changes to attention logits. *arXiv preprint arXiv:2511.21377*, 2025.

[44] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[45] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240): 1–113, 2023.

A Related Work

Low-precision training has become a standard route to scaling neural-network training. Early mixed-precision methods use FP16 [14, 15] or BF16 [16] for much of the forward and backward computation while retaining selected computations or states in FP32. Building on this mixed-precision framework, more recent work extends training to 8-bit floating point (FP8) formats [17]. For attention computation in Transformers, low-precision arithmetic is commonly paired with memory-efficient kernels. FlashAttention [9] and FlashAttention-2 [10] reduce memory traffic through IO-aware tiling and online softmax, and FlashAttention-3 [18] further develops FP8 attention kernels with explicit mechanisms for reducing numerical error.

Beyond computational efficiency, recent studies have begun to examine the numerical consequences of low-precision training and memory-efficient attention kernels. Long-horizon language-model experiments show that FP8 training can exhibit instabilities over extended training [19]. In a different setting, Hanqing et al. [20] show that finite-precision errors in softmax cross-entropy backpropagation can break the zero-sum structure of class gradients and induce a systematic parameter drift, leading to abnormal parameter growth and loss spikes. At the attention-kernel level, Golden et al. [21] report substantially larger numerical errors for FlashAttention than for baseline attention in isolated BF16 forward passes. Their proxy analysis further suggests that the resulting effect on model weights is smaller than that of low-precision training. Most closely related, Qiu and Yao [12] attribute a catastrophic loss explosion to the coherent accumulation of biased low-precision errors along low-rank gradient directions. Together, these works show that low-precision arithmetic can induce numerical deviations, systematic parameter drift, and training instability. In contrast, we study sustained attention-logit growth in a regime where the next-token loss can remain well behaved.

Muon [5] is an optimizer that has recently attracted increasing attention in LLM training for its strong empirical performance and distinctive matrix-structured update rule. Muon forms a momentum matrix for each hidden-layer weight matrix and maps it to an approximately semi-orthogonal update through a Newton–Schulz iteration. Building on this update rule, a broad range of follow-up work has extended Muon along several directions. Among efforts aimed at large-scale training and practical deployment, Liu et al. [1] demonstrate that Muon can scale to large-scale LLM training by incorporating weight decay and carefully adjusting the per-parameter update scale, while MuonBP reduces communication overhead through block-periodic orthogonalization [22]. A complementary set of methods introduces finer-grained adaptivity into Muon's matrix-level update. NorMuon [23] introduces neuron-wise second-moment scaling and row-wise normalization, AdaMuon [24] combines element-wise adaptivity with orthogonalized updates, and Newton–Muon [25] incorporates input-side curvature through layer-input second moments. Other work refines or generalizes the orthogonalization operation itself: Polar Express [26] develops efficient polynomial approximations

to the polar factor used by Muon, while TEON [27] extends matrix-level orthogonalization to structured higher-order tensors.

Beyond these algorithmic extensions, another line of work aims to understand the optimization behavior of Muon. Building on the view that classical optimizers can be interpreted as steepest descent under different norms [28], recent studies characterize Muon or gradient orthogonalization through spectral-norm geometry and non-Euclidean trust-region interpretations [29]. Recent work interprets matrix-aware optimizers from a unifying preconditioning perspective [30], shows that Muon with decoupled weight decay implicitly enforces a spectral-norm constraint on the weight matrices [31], and formalizes the preconditioning effect of spectral orthogonalization in matrix factorization and in-context learning of linear transformers [32]. Recent theory further establishes convergence guarantees directly for the finite-step Newton–Schulz iteration used in practical Muon implementations [33].

A complementary line studies Muon's advantage through its learning dynamics and the representations it induces. Wang et al. [3] show that Muon's advantage in Transformers is concentrated in associative-memory learning under heavy-tailed data. In a related associative-memory model, Li et al. [34] show that Muon learns frequency components more uniformly than gradient descent and derive improved optimization scaling laws. From a generalization perspective, Vasudeva et al. [35] show that spectral updates promote more balanced learning across principal components on imbalanced data. More recently, Ruan et al. [36] show that Muon learns more robust and transferable features than Adam and SGD and explain these by a more balanced learning under a multi-component feature model. These works provide complementary explanations of Muon's empirical advantage from the perspectives of optimization geometry, spectral dynamics, associative memory, and feature learning. In contrast, our work studies how a structured numerical discrepancy generated by low-precision FlashAttention is averaged by momentum, transformed by the practical five-step Newton–Schulz map, and accumulated in the query and key parameters.

Attention-logit growth and QK spectral dynamics have been repeatedly observed in various settings of large-scale Transformer training. In an 8B-parameter Vision Transformer experiment, Dehghani et al. [7] observe attention logits exceeding 5×10^4, producing nearly one-hot attention together with unstable losses and gradients. Wortsman et al. [37] reproduce attention-logit growth in smaller models at aggressive learning rates and study such small-scale behavior as a proxy for large-scale instability. Beyond documenting this phenomenon, several works investigate the mechanisms that produce large or concentrated attention scores. Zhai et al. [38] relate training instability to low-entropy, highly concentrated attention and derive a lower bound on the attention entropy related to the spectral norm of the attention logits. Bondarenko et al. [39] show that attention heads attempting to implement a no-op can drive softmax inputs toward increasingly extreme values, thereby inducing activation outliers. More recently, Hajra [40] attribute one form of attention-logit growth to the interaction between dense local dependencies and the low-rank bottleneck of self-attention. Most closely related to the spectral structure studied here, Xie et al. [41] show that several coherent low-precision faults can converge onto a shared query–key (QK) spectral-runaway channel. Our work complements these studies by investigating the numerical and optimization mechanisms underlying precision-dependent attention-logit growth in low-precision FlashAttention training.

Attention-logit stabilization and control have been studied by modifying different components of the attention computation. QKNorm [42] normalizes queries and keys before their dot product to reduce softmax saturation, and related QK normalization [7] has been used to stabilize large-scale Transformer training. QK-Guard [41] monitors a downstream attention-logit signal and activates parameter-free per-head QK normalization after the signal crosses a prescribed threshold. Other methods directly modify the query–key parameters or their updates. Zhai et al. [38] propose σReparam to prevent the attention-entropy collapse, Anson and Aitchison [43] use parameter-dependent learning rates for the query and key weights to control changes in attention logits, and QK-Clip [6] rescales the query and key projection weights when a per-head maximum exceeds a prescribed threshold. Several methods instead modify the logits or attention rule itself. Gemma 2 [44] applies logit soft-capping to attention logits, while clipped softmax and gated attention mitigate extreme scores associated with no-op attention heads [39]. Long Short-attention (LS-attention) [40] combines local and full attention heads to mitigate logit explosion in long-sequence training. Relatedly, PaLM [45] introduces the z-loss, an auxiliary regularization term that penalizes the log normalizer of the vocabulary softmax to improve training stability; the attention z-loss evaluated in our work applies the same idea to the row-wise attention log-sum-exp. In contrast, our main

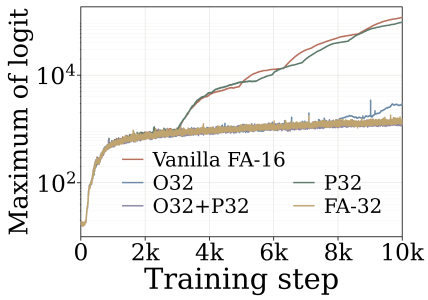


Figure 7: Trajectories of maximum logit for output-path precision ablations under NoPE. Preserving O at higher precision suppresses most of the excess growth, while higher-precision P alone has limited effect; combining the two brings the trajectory close to the FA-32 reference.

intervention repairs the numerical discrepancy at its FlashAttention precision boundaries without changing the mathematical attention function or learned parameterization.

B Appendix for Section 3.2

B.1 Ablation of the Attention-Output Path

We isolate numerical errors along the attention-output path by selectively retaining higher-precision intermediate quantities while keeping all other components unchanged. We compare five settings: Vanilla FA-16, `O32`, `P32`, `O32+P32`, and FA-32. In Vanilla FA-16, P is downcast before the PV multiplication, and the resulting attention output O is stored in BF16, the training dtype. Relative to Vanilla FA-16, `O32` retains an FP32 copy of O_i for computing $D_i = \langle \mathrm{d}O_i, O_i \rangle$ in the backward pass, whereas `P32` retains the softmax probabilities P in FP32 for the PV multiplication. `O32+P32` combines these two modifications. FA-32 serves as the full-precision reference.

As shown in Figure 7, Vanilla FA-16 exhibits pronounced late-stage excess logit growth. `P32` alone leaves the excess largely unchanged, whereas `O32` suppresses most of it, indicating that low-precision storage of the attention output is the dominant contributor along this path. Combining the two modifications, `O32+P32`, almost entirely removes the remaining excess, yielding a trajectory that is nearly indistinguishable from the FA-32 reference. These results motivate our focus below on how errors in the stored attention output propagate through the backward pass.

B.2 Propagation of the Stored-Output Error

The following proposition isolates how an error in the stored attention output propagates into the query and key parameter gradients.

Proposition 1 (Gradient structure induced by output-storage error)**.** *Consider two backward evaluations with identical $Q, K, V, P, \mathrm{d}O$, and input features, differing only in the stored attention output O. Let*

$$\Delta O_i = O_i^{\mathrm{lo}} - O_i^{\mathrm{hi}},$$

and define $G_Q = \mathrm{d}W_Q$ and $G_K = \mathrm{d}W_K$. Then, in FA*, the following statements hold:*

1. *The discrepancies in the row-wise correction and score gradient satisfy*

$$\Delta D_i = \langle \mathrm{d}O_i, \Delta O_i \rangle, \qquad \Delta \mathrm{d}S_{ij} = -P_{ij} \Delta D_i. \tag{3}$$

2. *For RoPE, define $\overline{k}_i = \sum_{j \le i} P_{ij} R_{ij} k_j$. The induced parameter-gradient discrepancies satisfy*

$$\Delta G_Q = -\tau \sum_i \Delta D_i \, \overline{k}_i x_i^\top, \qquad \Delta G_K = -\tau \sum_i \Delta D_i \sum_{j \le i} P_{ij} (R_{ij}^\top q_i) x_j^\top. \tag{4}$$

For NoPE, the same identities hold with $R_{ij} = I_{d_h}$.

Proof. Since only the stored output differs, $\mathrm{d}O$ is fixed, and hence

$$\Delta D_i = \langle \mathrm{d}O_i, O_i^{\mathrm{lo}} - O_i^{\mathrm{hi}} \rangle = \langle \mathrm{d}O_i, \Delta O_i \rangle.$$

Moreover, $\mathrm{d}P$ and P are unchanged. From $\mathrm{d}S_{ij} = P_{ij}(\mathrm{d}P_{ij} - D_i)$, we therefore obtain $\Delta \mathrm{d}S_{ij} = -P_{ij}\Delta D_i$. For RoPE, the attention backward gives $\mathrm{d}q_i = \tau \sum_{j \le i} \mathrm{d}S_{ij} R_{ij} k_j$, $\mathrm{d}k_j = \tau \sum_{i \ge j} \mathrm{d}S_{ij} R_{ij}^\top q_i$. Substituting equation 3 yields $\Delta \mathrm{d}q_i = -\tau \Delta D_i \, \overline{k}_i$, $\Delta \mathrm{d}k_j = -\tau \sum_{i \ge j} P_{ij} \Delta D_i R_{ij}^\top q_i$. Finally, using $G_Q = \sum_i \mathrm{d}q_i x_i^\top$, $G_K = \sum_j \mathrm{d}k_j x_j^\top$, and rearranging the sums gives equation 4. □

C Experimental Support for Theoretical Assumptions

Approximate Orthogonality of Centered Features. Assumption 1 requires the centered feature vectors to be approximately orthogonal on average. To examine this condition, we compute the pairwise angles between centered token features and visualize their distribution using a KDE plot. As shown in Figure 8, the angles are tightly concentrated around $90°$, with a mean angle of $89.46°$. This concentration indicates that the centered features exhibit near-orthogonality, supporting Assumption 1.

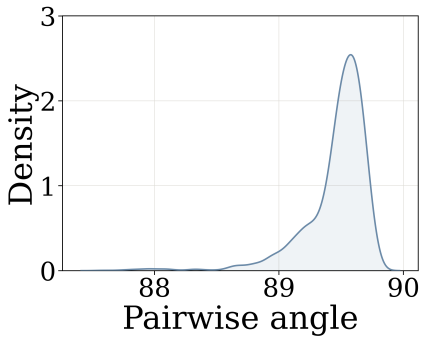


Figure 8: Distribution of pairwise angles between centered token features. The mean angle is $89.46°$, indicating approximate orthogonality.

Approximate Projection Structure at Initialization. Assumption 2 models the initialized query and key matrices as scaled projections on their active subspaces. To examine this property, we analyze the Gram matrices at step 0. For each weight matrix $W \in \{W_Q, W_K\}$, we compute its thin SVD $W = U\Sigma V^\top$ and define the row-space projector $P = VV^\top$. The closest scaled projection within this subspace is cP, where $c = \mathrm{tr}(W^\top W)/\mathrm{rank}(W^\top W)$. We then measure the agreement between $W^\top W$ and cP using the projection isotropy score $S_{\mathrm{proj}} = \mathrm{tr}(W^\top W)^2 / \left(\mathrm{rank}(W^\top W)\|W^\top W\|_F^2\right)$. The score equals 1 for an exact scaled projection. At initialization, the scores are 0.9206 ± 0.0033 for W_Q and 0.9222 ± 0.0044 for W_K, indicating that both Gram matrices are close to scaled projections on their row spaces.

Alignment of Dominant Query and Key Error Directions. Assumption 3 models the dominant low-precision error as a rank-one perturbation whose directions are determined by $\hat{q}$ and $\hat{k}$. The rank-one structure and the correspondence between the error and the directions $\hat{q}^\top x$ and $\hat{k}^\top x$ have been verified in Section 3.2. Here, we examine the remaining directional condition required by the theory. Firstly, we show the positive alignment between the query and key directions. As shown in Figure 9, after an initial transition period of several thousand steps, the cosine similarity between $\hat{k}$ and $\hat{q}$ becomes positive and continues increasing toward 1, indicating that the dominant query and key directions become strongly aligned in the late training stage.

Secondly, we examine the empirical motivation for the orthogonality assumption. Our theoretical analysis assumes that the numerical-error directions are orthogonal to the task-relevant subspace. As an empirical sanity check, we measure the cosine similarities between the optimization gradients G_Q, G_K and numerical error directions $\Delta G_Q, \Delta G_K$ for W_Q and W_K. As shown in Figure 10,

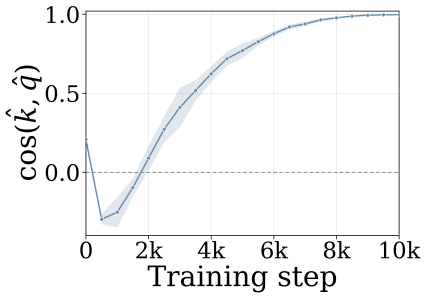


Figure 9: Cosine similarity between the dominant query and key directions $\hat{q}$ and $\hat{k}$ during training. The two directions become positively aligned after the transition stage and approach near-perfect alignment in the late stage.

these similarities approach zero during training, suggesting that the gradient and noise directions are nearly orthogonal in practice. This observation provides empirical support for our orthogonality assumption.

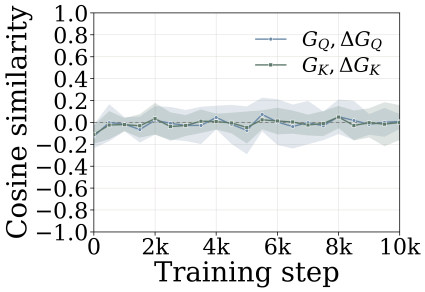


Figure 10: Cosine similarities between the optimization gradients and numerical noise directions for W_Q and W_K. The similarities approach zero during training, suggesting near orthogonality between the two directions.

D Proof of Theorem 1

D.1 Notation

We use the model introduced in Section 4. Throughout the appendix, write $r = M - 1$, $\kappa_m = \tau\sigma^2_{\text{type}}$, $\kappa_{\text{com}} = \sigma^2_{\text{com}}\chi$, $C_h = 1 - 1/M$, and let $\gamma_{\text{hi}} = 0$, $\gamma_{\text{lo}} = \Gamma$. As in Section 4, M is the number of token types and T is the number of context tokens; we assume $M, T \geq 2$ throughout. We write $y \in [M]$ for the query-type index, so that $x_{T+1} = z_y$ and $x_j = c_{i_j}$ for $j \in [T]$. Given y, the joint context law is obtained from T independent uniform samples from $\mathcal{C}$ by conditioning on at least one occurrence of c_y. Write $Z_{M,T} = 1 - (1 - 1/M)^T$ for the probability of this conditioning event. $E_Q = \widehat{k}\overline{x}^\top$ and $E_K = \widehat{q}\overline{x}^\top$; these are independent of the training step by Assumption 3.

D.2 Proof of Theorem 1(i)

Proof. Our proof contains four steps. We first identify a lower bound on the population loss over all trainable parameters. We then show that the updates preserve a two-dimensional subspace. Next, we express the scores and loss on this subspace and show that the lower bound is sharp. Finally, we prove that both training sequences approach the optimal loss and that their exact gradients vanish.

Step 1. Identify a global lower bound on the loss. For one sample, let $\pi_i = \sum_{j:i_j=i} P_j$ be the total attention mass assigned to context tokens of type i. The masses satisfy $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. Since $o = \sum_i \pi_i f_i$, the fixed readout gives $f_i^\top o = \pi_i - 1/M$. Therefore, the loss of each sample

satisfies

$$\log\left(1+\sum_{i\neq y} e^{\lambda(\pi_i-\pi_y)}\right) \geq \log(1+re^{-\lambda}) = \ell_\infty, \tag{5}$$

because $\pi_i - \pi_y \geq -1$ for every $i \neq y$. Averaging over the population gives $\mathcal{L}(W_Q, W_K) \geq \ell_\infty$, and hence $\mathcal{L}^\star \geq \ell_\infty$.

Step 2. Show that the updates preserve a two-dimensional subspace. For $g > 0$ and $\gamma \geq 0$, define

$$h_{\text{task}}(g,\gamma) = \psi_5\left(\frac{g}{\sqrt{rg^2+\gamma^2}+\varepsilon_{\text{NS}}}\right), \qquad h_{\text{com}}(g,\gamma) = \psi_5\left(\frac{\gamma}{\sqrt{rg^2+\gamma^2}+\varepsilon_{\text{NS}}}\right). \tag{6}$$

These are the coefficients by which NS5 updates the type-dependent and shared directions, respectively, where $\psi_5 = \psi^{\circ 5}$, $\psi(t) = \omega_0 t + \omega_1 t^3 + \omega_2 t^5$, and $(\omega_0, \omega_1, \omega_2) = (3.4445, -4.7750, 2.0315)$.

The following proposition shows that, for each $A \in \{Q, K\}$, the training dynamics remain in the two-dimensional subspace spanned by Φ_A and E_A, and gives the corresponding coordinate updates.

Proposition 2 (Closed population dynamics)**.** *At each precision $\mathfrak{b} \in \{\text{hi}, \text{lo}\}$, there are scalar sequences $u_n^{\mathfrak{b}} > 0$ and $a_n^{\mathfrak{b}} \geq 0$ such that*

$$W_{A,n}^{\mathfrak{b}} = u_n^{\mathfrak{b}} \Phi_A + a_n^{\mathfrak{b}} E_A, \tag{7}$$

with $u_0^{\mathfrak{b}} = u_0$ and $a_0^{\mathfrak{b}} = 0$. The unrestricted population gradients take the form

$$\nabla_{W_A} \mathcal{L}_n^{\mathfrak{b}} = -g(u_n^{\mathfrak{b}}) \Phi_A, \tag{8}$$

for a smooth positive scalar function g on $(0, \infty)$. The same full-gradient identity holds at every state $W_A = u\Phi_A + aE_A$ with $u > 0$ and $a \geq 0$. The coordinate recurrences are

$$\begin{aligned} u_{n+1}^{\mathfrak{b}} &= u_n^{\mathfrak{b}} + \eta h_{\text{task}}\big(g(u_n^{\mathfrak{b}}), \gamma_{\mathfrak{b}}\big), \\ a_{n+1}^{\mathfrak{b}} &= a_n^{\mathfrak{b}} + \eta h_{\text{com}}\big(g(u_n^{\mathfrak{b}}), \gamma_{\mathfrak{b}}\big). \end{aligned} \tag{9}$$

Moreover, $a_n^{\text{hi}} = 0$ for every n.

Since Φ_A has r unit singular values, $g(u)$ is the magnitude of each task-gradient singular value, and $\sqrt{r}\, g(u)$ is the Frobenius norm of the task gradient.

Step 3. Express the scores and loss on the two-dimensional subspace and show that the lower bound is sharp. By Proposition 2, the training dynamics remain in the family $W_Q = u\Phi_Q + aE_Q$ and $W_K = u\Phi_K + aE_K$. We first characterize the scores and loss on this family. By definition of the model, we have

$$W_Q z_y = \sigma_{\text{type}} u \Phi_Q e_{z,y} + \sigma_{\text{com}} a \widehat{k}, \qquad W_K c_i = \sigma_{\text{type}} u \Phi_K e_{c,i} + \sigma_{\text{com}} a \widehat{q}. \tag{10}$$

Since $\Phi_Q e_{z,j} = \Phi_K e_{c,j}$ for each type j and these maps preserve the type-feature inner products, the raw attention logit between query type y and context type i is $L(i;y) = \kappa_{\text{com}} a^2 + (\kappa_m/\tau) u^2 \big(\mathbf{1}\{i = y\} - 1/M\big)$, and the corresponding scaled score is $S(i;y) = \tau\kappa_{\text{com}} a^2 + \kappa_m u^2 \big(\mathbf{1}\{i = y\} - 1/M\big)$.

Set $m = \kappa_m u^2$ and $s_{\text{com}} = \kappa_{\text{com}} a^2$. For every $i \neq y$, the scaled score gap satisfies $S(y;y) - S(i;y) = m$, so m is the margin, whereas s_{com} is a type-independent common shift in the raw attention logits. The common shift cancels from the attention softmax and therefore does not affect the loss, but it does shift the raw logits themselves. Since every example contains a matching context token, its maximum raw attention logit is $s_{\text{com}} + C_h m/\tau$.

We now express the loss in terms of the margin. For each example with target y, let $\mathbf{N} = (N_1, \ldots, N_M)$ denote the type counts among all T context tokens. Given y, they are multinomial counts from T uniform draws over the M types, conditioned on $N_y \geq 1$. Thus the context contains N_i tokens of type i, with $\sum_i N_i = T$ and $N_y \geq 1$. After cancelling the common scaled score in the attention softmax, the total attention mass assigned to type i is

$$\pi_i(m) = \frac{N_i e^{m\mathbf{1}\{i=y\}}}{N_y e^m + T - N_y}. \tag{11}$$

By the definitions of the value features and readout, the class-i logit is $\lambda(\pi_i(m) - 1/M)$. The common offset $-\lambda/M$ cancels from the cross-entropy. Thus, for fixed counts $\mathbf{N}$ and target y, the example loss depends on the model parameters only through the margin m and is given by $\ell_{\mathrm{ex}}(m; \mathbf{N}, y) = \log \sum_{i=1}^{M} e^{\lambda \pi_i(m)} - \lambda \pi_y(m)$. The population loss is therefore

$$\ell(m) = \mathbb{E}_{\mathbf{N}|y} \left[\ell_{\mathrm{ex}}(m; \mathbf{N}, y) \right], \tag{12}$$

where the expectation is under the conditional count law above and, by permutation symmetry, has the same value for every target y. Hence the original population objective restricted to the two-dimensional family satisfies

$$\mathcal{L}(u\Phi_Q + aE_Q, u\Phi_K + aE_K) = \ell(\kappa_m u^2). \tag{13}$$

Thus ℓ is exactly the original population loss restricted to this parameter family. Combining this representation with the full-gradient identity in Proposition 2 gives $g(u) = -(\kappa_m u/r)\ell'(\kappa_m u^2)$. Moreover, Lemma 3 shows that ℓ is smooth and $\ell'(m) < 0$ for every finite $m \geq 0$. At training step n, we have $m_n^{\mathfrak{b}} = \kappa_m (u_n^{\mathfrak{b}})^2$ and $s_{\mathrm{com},n}^{\mathfrak{b}} = \kappa_{\mathrm{com}} (a_n^{\mathfrak{b}})^2$. Then $\mathcal{L}_n^{\mathfrak{b}} = \ell(m_n^{\mathfrak{b}})$, $L_n^{\mathfrak{b}}(i; y) = s_{\mathrm{com},n}^{\mathfrak{b}} + (m_n^{\mathfrak{b}}/\tau)\big(\mathbf{1}\{i = y\} - 1/M\big)$, and $\mathrm{Logit}_{\max,n}^{\mathfrak{b}} = s_{\mathrm{com},n}^{\mathfrak{b}} + (C_h/\tau) m_n^{\mathfrak{b}}$. Moreover, $s_{\mathrm{com},n}^{\mathrm{hi}} = 0$ for every n.

It remains to show that the lower bound from Step 1 is sharp. For this purpose, consider the parameter family with $a = 0$ and take the parameter limit $u \to \infty$, so that $m = \kappa_m u^2 \to \infty$. Because every example contains at least one matching token, (11) gives $\pi_y(m) \to 1$ and $\pi_i(m) \to 0$ for every $i \neq y$. Substituting these limits into (12) yields $\ell(m) \to \ell_\infty$. Together with the lower bound from Step 1 and (13), we obtain

$$\mathcal{L}^\star = \inf_{W_Q, W_K} \mathcal{L}(W_Q, W_K) = \ell_\infty = \log\big(1 + (M-1)e^{-\lambda}\big). \tag{14}$$

Step 4. Prove convergence to the global optimum. Step 3 has identified $\mathcal{L}^\star$. It remains to show that the actual training dynamics drive the margin to infinity. We use the following consequence of the scalar recurrences.

Proposition 3 (Unbounded margin progress)**.** *At each precision, the task coordinate $u_n^{\mathfrak{b}}$ diverges to infinity, the population loss $\mathcal{L}_n^{\mathfrak{b}} = \ell(m_n^{\mathfrak{b}})$ converges monotonically to the global optimum $\mathcal{L}^\star = \ell_\infty$, and the exact gradient coefficient $g(u_n^{\mathfrak{b}})$ converges to zero.*

Its proof is in Appendix E.2. Together with the global characterization in Step 3, the proposition proves that the actual training losses decrease strictly to $\mathcal{L}^\star$, not merely to the optimum within the two-parameter family. Also $\|\nabla_{W_A} \mathcal{L}_n^{\mathfrak{b}}\|_F = \sqrt{r}\, g(u_n^{\mathfrak{b}}) \to 0$ for both query and key. This proves part (i). □

D.3 Proof of Theorem 1(ii)

Proof. Our proof contains three steps. We first use the fixed gradient level to obtain a parameter interval and a discrete training horizon, both independent of the error magnitude. We then compare the coordinate updates and bound each logit-gap increment uniformly in Γ. Finally, we sum the increments and divide by a lower bound on the high-precision maximum.

Step 1. Choose the interval from a fixed gradient level. Write $\mathcal{G}(u) = \sqrt{r}\, g(u)$, $\mathcal{G}_n^{\mathfrak{b}} = \mathcal{G}(u_n^{\mathfrak{b}})$, and $\mathcal{G}_0 = \mathcal{G}(u_0)$. Proposition 2 identifies these quantities as exact gradient norms. Let $F_\gamma(u) = h_{\mathrm{task}}(g(u), \gamma)$ and $\kappa_n = h_{\mathrm{com}}(g(u_n^{\mathrm{lo}}), \Gamma)$, so the coordinate increments are $\eta F_{\gamma_{\mathfrak{b}}}(u_n^{\mathfrak{b}})$ and $\eta\kappa_n$. Fix $0 < \mathcal{G}_{\mathrm{E}} < \mathcal{G}_0$, independently of Γ, and define $u_{\mathrm{E}} = \inf\{u > u_0 : \mathcal{G}(u) \leq \mathcal{G}_{\mathrm{E}}\}$. The function $\mathcal{G}$ is continuous and positive, and Lemma 3 gives $\mathcal{G}(u) \to 0$ as $u \to \infty$. Therefore $u_0 < u_{\mathrm{E}} < \infty$, $\mathcal{G}(u_{\mathrm{E}}) = \mathcal{G}_{\mathrm{E}}$, and $\mathcal{G}(u) > \mathcal{G}_{\mathrm{E}}$ on $[u_0, u_{\mathrm{E}})$.

Let $U_5 = 1.203$ denote the uniform upper bound of the five-step NS polynomial ψ_5 on normalized inputs in $[0, 1]$. Lemma 4 gives $0 < F_\gamma(u) \leq U_5$ for every $u > 0$ and $\gamma \geq 0$. Thus each task-coordinate update is at most ηU_5, so traversing the parameter interval from u_0 to u_{E} requires at least the scaled time $(u_{\mathrm{E}} - u_0)/U_5$. We therefore define this guaranteed time interval and the corresponding discrete horizon by

$$T_{\mathrm{E}} = \frac{u_{\mathrm{E}} - u_0}{U_5}, \qquad \eta_{\mathrm{E}} = \frac{T_{\mathrm{E}}}{2}, \qquad N_{\mathrm{E}} = \left\lceil \frac{T_{\mathrm{E}}}{\eta} \right\rceil - 1. \tag{15}$$

For $0 < \eta \le \eta_{\mathrm{E}}$, this integer is at least one and satisfies $T_{\mathrm{E}} - \eta \le \eta N_{\mathrm{E}} < T_{\mathrm{E}}$. Summing the exact updates gives $u_n^{\mathfrak{b}} \le u_0 + U_5 \eta n < u_{\mathrm{E}}$ for $n \le N_{\mathrm{E}}$, at either precision and for every $\Gamma > 0$. Hence both gradient norms exceed $\mathcal{G}_{\mathrm{E}}$ throughout this horizon. Neither u_{E} nor N_{E} depends on Γ.

Step 2. Bound the updates uniformly in the error magnitude. Write $\alpha_5 = \psi_5'(0) = \omega_0^5$. We use the following exact comparison of the normalized task inputs.

Lemma 1 (Task perturbation and parameter dependence)**.** *For $u > 0$ and $\gamma \ge 0$, define $\mathcal{D}_\gamma(u) = \sqrt{r g(u)^2 + \gamma^2} + \varepsilon_{\mathrm{NS}}$ and $\rho_\gamma(u) = \frac{g(u)}{\mathcal{D}_\gamma(u)}$. With $M_5(u) = \max_{\rho_\Gamma(u) \le t \le \rho_0(u)} |\psi_5'(t)|$, the following bounds and identity hold:*

1. $|F_\Gamma(u) - F_0(u)| \le M_5(u)\Gamma^2 / [2\sqrt{r}\,\mathcal{D}_0(u)^2]$.
2. $F_0'(u) = \varepsilon_{\mathrm{NS}} \psi_5'(\rho_0(u)) g'(u) / \mathcal{D}_0(u)^2$.
3. $0 \le h_{\mathrm{com}}(g(u), \Gamma) \le \alpha_5 \Gamma / \mathcal{D}_\Gamma(u)$.

The proof is in Appendix F.1. It rationalizes the normalization denominators before applying the mean-value theorem and does not require ψ_5 to be close to its linear approximation. On the fixed interval $I_{\mathrm{E}} = [u_0, u_{\mathrm{E}}]$, define the following constants

$$C_{\mathrm{E}}^{\mathrm{pert}} = \frac{882}{2\sqrt{r}\,(\mathcal{G}_{\mathrm{E}} + \varepsilon_{\mathrm{NS}})^2}, \qquad C_{\mathrm{E}}^{\mathrm{com}} = \frac{\alpha_5}{\mathcal{G}_{\mathrm{E}} + \varepsilon_{\mathrm{NS}}}, \qquad \Lambda_{\mathrm{E}} = \max_{u \in I_{\mathrm{E}}} |F_0'(u)|. \tag{16}$$

Here $C_{\mathrm{E}}^{\mathrm{pert}}$ bounds the direct perturbation of the task update, $C_{\mathrm{E}}^{\mathrm{com}}$ bounds the common-direction update, and Λ_{E} controls the sensitivity of the unperturbed task update to changes in u. Since $|\psi_5'| \le 882$ on $[0, 1]$ and $\mathcal{D}_\Gamma \ge \mathcal{D}_0 \ge \mathcal{G}_{\mathrm{E}} + \varepsilon_{\mathrm{NS}}$, the lemma gives $|F_\Gamma - F_0| \le C_{\mathrm{E}}^{\mathrm{pert}} \Gamma^2$, $|F_0'| \le \Lambda_{\mathrm{E}}$, and $0 \le h_{\mathrm{com}} \le C_{\mathrm{E}}^{\mathrm{com}} \Gamma$ on I_{E}, uniformly over $\Gamma > 0$. All three constants are finite and independent of Γ and η.

We next collect these bounds into constants that control the accumulated task-coordinate mismatch, the accumulated common coordinate, the task-update mismatch, and finally the one-step logit-gap increment:

$$\begin{gathered} C_{\mathrm{E}}^u = C_{\mathrm{E}}^{\mathrm{pert}} T_{\mathrm{E}} e^{\Lambda_{\mathrm{E}} T_{\mathrm{E}}}, \qquad C_{\mathrm{E}}^a = C_{\mathrm{E}}^{\mathrm{com}} T_{\mathrm{E}}, \qquad C_{\mathrm{E}}^F = C_{\mathrm{E}}^{\mathrm{pert}} + \Lambda_{\mathrm{E}} C_{\mathrm{E}}^u, \\ C_{\mathrm{E}}^{\mathrm{abs}} = \kappa_{\mathrm{com}} (2 C_{\mathrm{E}}^a C_{\mathrm{E}}^{\mathrm{com}} + \eta_{\mathrm{E}} (C_{\mathrm{E}}^{\mathrm{com}})^2) + \frac{2 C_h \kappa_m}{\tau} [u_{\mathrm{E}} C_{\mathrm{E}}^F + U_5 C_{\mathrm{E}}^u + \eta_{\mathrm{E}} U_5 C_{\mathrm{E}}^F]. \end{gathered} \tag{17}$$

Specifically, C_{E}^u bounds the task-coordinate difference after accumulation over the early interval, C_{E}^a bounds the common coordinate, C_{E}^F bounds the resulting difference between the two task-update coefficients, and $C_{\mathrm{E}}^{\mathrm{abs}}$ combines these contributions into a uniform bound on the logit-gap increment.

The following proposition turns these uniform update estimates into the required gap bounds.

Proposition 4 (Uniform comparison above a fixed gradient floor)**.** *With the constants and horizon above, for every $\Gamma > 0$ and $0 < \eta \le \eta_{\mathrm{E}}$, we have*

1. $|u_n^{\mathrm{lo}} - u_n^{\mathrm{hi}}| \le \Gamma^2 C_{\mathrm{E}}^{\mathrm{pert}} \eta n e^{\Lambda_{\mathrm{E}} \eta n}$, $0 \le a_n^{\mathrm{lo}} \le \Gamma C_{\mathrm{E}}^{\mathrm{com}} \eta n, 0 \le n \le N_{\mathrm{E}}$,
2. $|\Delta_{n+1} - \Delta_n| \le C_{\mathrm{E}}^{\mathrm{abs}} \Gamma^2 \eta$, $0 \le n < N_{\mathrm{E}}$,
3. $|\Delta_n| \le C_{\mathrm{E}}^{\mathrm{abs}} \Gamma^2 \eta n$, $0 \le n \le N_{\mathrm{E}}$.

Its proof is in Appendix E.3. In particular, the factors multiplying Γ^2 are chosen on a fixed parameter interval, rather than on an interval that changes with Γ.

Step 3. Normalize by the high-precision maximum. Put $\mu_{\mathrm{E}} = \min\{2/3, \mathcal{G}_{\mathrm{E}} / [\sqrt{r}\,(\mathcal{G}_{\mathrm{E}} + \varepsilon_{\mathrm{NS}})]\} > 0$. Since $\mathcal{G}(u_j^{\mathrm{hi}}) > \mathcal{G}_{\mathrm{E}}$ for $j < N_{\mathrm{E}}$, the lower bound $\psi_5(t) \ge \min\{t, 2/3\}$ gives $F_0(u_j^{\mathrm{hi}}) \ge \mu_{\mathrm{E}}$. Summing the high-precision updates and using the raw-logit identity from Step 3 of part (i) yields

$$u_n^{\mathrm{hi}} \ge u_0 + \mu_{\mathrm{E}} \eta n, \qquad \mathrm{Logit}_{\max,n}^{\mathrm{hi}} \ge \frac{C_h \kappa_m}{\tau} (u_0 + \mu_{\mathrm{E}} \eta n)^2 > 0. \tag{18}$$

Therefore, we have

$$\frac{|\Delta_n|}{\mathrm{Logit}^{\mathrm{hi}}_{\max,n}} \le \frac{\tau C_{\mathrm{E}}^{\mathrm{abs}} \Gamma^2 \eta n}{C_h \kappa_m (u_0 + \mu_{\mathrm{E}} \eta n)^2} \le C_{\mathrm{E}}^{\mathrm{rel}} \Gamma^2 \frac{\eta n}{(u_0 + \eta n)^2}, \qquad 0 \le n \le N_{\mathrm{E}}, \tag{19}$$

where $C_{\mathrm{E}}^{\mathrm{rel}} = \tau C_{\mathrm{E}}^{\mathrm{abs}} / (C_h \kappa_m \mu_{\mathrm{E}}^2)$. The second inequality uses $0 < \mu_{\mathrm{E}} \le 2/3 < 1$. Taking $C_{\mathrm{E}} = \max\{C_{\mathrm{E}}^{\mathrm{abs}}, C_{\mathrm{E}}^{\mathrm{rel}}\}$ proves part (ii), with no dependence of the constants or the horizon on Γ. □

D.4 Proof of Theorem 1(iii)

Proof. We bound the margins by logarithmic growth, obtain the quadratic common shift directly from the update recurrence, and then compare the logits and their increments. All asymptotics below are for fixed model parameters, $\Gamma > 0$, and $\eta > 0$.

Step 1. Bound the margins. By Proposition 3, both margins tend to infinity. Fix either precision and suppress its superscript. For all sufficiently large n, $m_n \ge 2$. Since the normalization denominator is at least $\varepsilon_{\mathrm{NS}} > 0$, Lemmas 3 and 4 give

$$0 < u_{n+1} - u_n \le \frac{\eta \omega_0^5}{\varepsilon_{\mathrm{NS}}} g(u_n) \le C u_n e^{-m_n}, \qquad C = \frac{2\eta \omega_0^5 \kappa_m \lambda (T-1)}{r \varepsilon_{\mathrm{NS}}}.$$

Expanding $m_n = \kappa_m u_n^2$ after one update yields

$$0 < m_{n+1} - m_n \le (2C + C^2 e^{-2}) m_n e^{-m_n} \le B e^{-m_n/2}, \quad B = \frac{2}{e}(2C + C^2 e^{-2}),$$

where the last inequality uses $m e^{-m/2} \le 2/e$ for $m \ge 2$. Thus the margin increments are bounded. Setting $Y_n = e^{m_n/2}$ and using $e^t - 1 \le t e^t$ for $t \ge 0$, we obtain

$$Y_{n+1} - Y_n \le \frac{Y_n (m_{n+1} - m_n)}{2} e^{(m_{n+1} - m_n)/2} \le \frac{B}{2} e^{B/2}.$$

Summation gives $Y_n = O(n)$ and hence $m_n = O(\log n)$ at both precisions. No finite-step entry bound is needed.

Step 2. Obtain the common-shift asymptotics. Proposition 3 also gives $g(u_n^{\mathrm{lo}}) \to 0$. Writing $\kappa_n = h_{\mathrm{com}}(g(u_n^{\mathrm{lo}}), \Gamma)$, continuity gives $\kappa_n \to \psi_5(\Gamma/(\Gamma + \varepsilon_{\mathrm{NS}})) = \kappa_\Gamma > 0$. Since $a_0^{\mathrm{lo}} = 0$, summing the recurrence and taking arithmetic means yields

$$a_n^{\mathrm{lo}} = \eta \sum_{j=0}^{n-1} \kappa_j = \eta \kappa_\Gamma n + o(n), \quad s_{\mathrm{com},n}^{\mathrm{lo}} = \kappa_{\mathrm{com}} (a_n^{\mathrm{lo}})^2 = \kappa_{\mathrm{com}} \kappa_\Gamma^2 (\eta n)^2 + o((\eta n)^2). \tag{20}$$

Step 3. Compare the logits and their increments. The high-precision maximum satisfies $\mathrm{Logit}^{\mathrm{hi}}_{\max,n} = (C_h/\tau) m_n^{\mathrm{hi}} = O(\log n)$, since its common shift vanishes. Hence, we have

$$\Delta_n = s_{\mathrm{com},n}^{\mathrm{lo}} + \frac{C_h}{\tau}(m_n^{\mathrm{lo}} - m_n^{\mathrm{hi}}) = \kappa_{\mathrm{com}} \kappa_\Gamma^2 (\eta n)^2 + o((\eta n)^2), \tag{21}$$

because the margin difference is $O(\log n) = o(n^2)$. For the increment, by $a_{n+1}^{\mathrm{lo}} - a_n^{\mathrm{lo}} = \eta \kappa_n$, we obtain

$$\begin{aligned} \Delta_{n+1} - \Delta_n &= \kappa_{\mathrm{com}}(2\eta a_n^{\mathrm{lo}} \kappa_n + \eta^2 \kappa_n^2) + \frac{C_h}{\tau}[(m_{n+1}^{\mathrm{lo}} - m_n^{\mathrm{lo}}) - (m_{n+1}^{\mathrm{hi}} - m_n^{\mathrm{hi}})] \\ &= 2\kappa_{\mathrm{com}} \kappa_\Gamma^2 \eta^2 n + o(\eta^2 n). \end{aligned} \tag{22}$$

Indeed, $a_n^{\mathrm{lo}} = \eta \kappa_\Gamma n + o(n)$ and $\kappa_n \to \kappa_\Gamma$, whereas $\eta^2 \kappa_n^2$ and both margin increments are bounded. Both leading coefficients are positive, so $\Delta_n = \Theta(n^2)$ and $\Delta_{n+1} - \Delta_n = \Theta(n)$, proving part (iii). Since $\mathrm{Logit}^{\mathrm{hi}}_{\max,n} > 0$ and is $O(\log n)$, this also yields $\Delta_n / \mathrm{Logit}^{\mathrm{hi}}_{\max,n} = \Omega\big(n^2/\log n\big)$. □

E Proofs of Supporting Propositions

In the supporting proofs, $\ell_{\mathrm{ex}}(m) = \log \sum_i e^{\lambda \pi_i(m)} - \lambda \pi_y(m)$ denotes the scalar loss for fixed counts and target, so that $\ell(m) = \mathbb{E}_{\mathbf{N}|y} \ell_{\mathrm{ex}}(m)$.

E.1 Proof of Proposition 2

Proof. At any candidate state $W_A = u\Phi_A + aE_A$, direct substitution into the model gives the score and loss identities in Step 3 of the proof of Theorem 1(i); these identities do not require invariance of the training dynamics. Write $g(u) = -(\kappa_m u/r)\ell'(\kappa_m u^2)$. We now show that the actual updates retain this state form. First, we compute the full population gradient at such a state. Second, we show that its task coefficient is positive. Third, we apply NS5 and close an induction over the training steps.

Step 1. Identify the full matrix gradient. Differentiating $\ell(\kappa_m u^2)$ with respect to u would vary W_Q and W_K together and would not by itself determine either full matrix gradient. We instead use the following identity, obtained by differentiating the original example loss before population averaging.

Lemma 2 (Aligned features and the population gradient)**.** *Each Φ_A has r unit singular values, and satisfies*

$$\sum_{i=1}^{M} \Phi_K e_{c,i} e_{z,i}^\top = \Phi_Q, \qquad \sum_{i=1}^{M} \Phi_Q e_{z,i} e_{c,i}^\top = \Phi_K.$$

For each $A \in \{Q, K\}$, the matrices Φ_A and E_A have orthogonal row spaces and orthogonal column spaces, and $\|E_A\|_F = 1$. At $W_Q = u\Phi_Q + aE_Q$ and $W_K = u\Phi_K + aE_K$, where $u > 0$ and $a \geq 0$, the unrestricted population gradients satisfy

$$\nabla_{W_A} \mathcal{L} = \frac{\kappa_m u}{r} \ell'(\kappa_m u^2) \Phi_A, \qquad A \in \{Q, K\}. \tag{23}$$

In particular, no shared-feature or mixed feature term remains after averaging over the population.

The proof is in Appendix F.2. The lemma identifies the full gradient as $-g(u)\Phi_A$, not only its projection onto Φ_A. The corresponding objective is already known from (13) to be $\ell(m)$ with $m = \kappa_m u^2$. We next check the sign of g.

Step 2. Verify that the task update increases the margin. We need $-\ell'(m) > 0$. The following lemma also supplies the gradient decay used in part (iii).

Lemma 3 (Scalar loss and its derivative)**.** *The population function ℓ defined in (12) is smooth. For every finite $m \geq 0$, we have $0 < -\ell'(m) \leq 2\lambda(T-1)e^{-m}$, and $\ell(m)$ converges monotonically to ℓ_∞ as $m \to \infty$.*

Its proof is in Appendix F.3. Since $u > 0$ and $\ell'(\kappa_m u^2) < 0$, we have $g(u) > 0$. Since ℓ is smooth, g is smooth on $(0, \infty)$. At precision $\mathfrak{b}$, the supplied gradient is therefore $-g(u)\Phi_A - \gamma_{\mathfrak{b}} E_A$, with $\gamma_{\mathrm{hi}} = 0$ and $\gamma_{\mathrm{lo}} = \Gamma$.

Step 3. Apply NS5 and prove preservation by induction. The two summands in the supplied gradient have orthogonal row and column spaces. The next lemma states how the matrix iteration acts on them.

Lemma 4 (Scalar NS5 response)**.** *For $g > 0$ and $\gamma \geq 0$, we have*

$$\mathrm{NS5}\left(-g\Phi_A - \gamma E_A\right) = -h_{\mathrm{task}}(g, \gamma)\Phi_A - h_{\mathrm{com}}(g, \gamma) E_A, \qquad A \in \{Q, K\}. \tag{24}$$

Here, NS5 *consists of a normalization step followed by five matrix-polynomial iterations, as defined in Section 2. The polynomial ψ_5 is odd and positive on $(0, \infty)$. For $0 < t \leq 1$, $\min\{t, 2/3\} \leq \psi_5(t) \leq U_5 = 1.203$, $\psi_5(t) \leq \omega_0^5 t$, $|\psi_5'(t)| \leq 882$.*

The proof is in Appendix F.4. At $n = 0$, the desired form holds with $u = u_0$ and $a = 0$. Suppose it holds at step n for one precision. Steps 1–2 give the full supplied gradient at that state. Applying the lemma and subtracting the NS5 update yields

$$W_{A,n+1} = [u_n + \eta h_{\mathrm{task}}(g(u_n), \gamma)]\Phi_A + [a_n + \eta h_{\mathrm{com}}(g(u_n), \gamma)] E_A.$$

The two bracketed coefficients are the same for query and key, so a single pair u_{n+1}, a_{n+1} again describes both matrices. Positivity of ψ_5 gives $u_{n+1} > u_n > 0$ and $a_{n+1} \geq a_n \geq 0$. At high precision, $\gamma = 0$ makes the common increment zero, so $a_n^{\mathrm{hi}} = 0$ for every n. This proves the induction and all claims of the proposition. □

E.2 Proof of Proposition 3

Proof. We first rule out a finite limiting task coordinate, then use the global loss value from (14) to identify the limit along training. Fix one precision and suppress its superscript; its error coefficient γ is either 0 or Γ.

Step 1. Prove that the task coordinate diverges. Proposition 2 gives $u_{n+1} - u_n = \eta h_{\mathrm{task}}(g(u_n), \gamma) > 0$. Thus u_n and $m_n = \kappa_m u_n^2$ are strictly increasing. Suppose $u_n \le R < \infty$ for every n. On $[u_0, R]$, the function $u \mapsto h_{\mathrm{task}}(g(u), \gamma)$ is continuous and positive, so its minimum is some $C_R > 0$. Every update would then satisfy $u_{n+1} - u_n \ge \eta C_R$, and summing gives $u_n \ge u_0 + n\eta C_R$, a contradiction. Hence $u_n \uparrow \infty$ and $m_n \uparrow \infty$.

Step 2. Deduce optimal-loss convergence and the gradient limit. Since m_n increases strictly to infinity and ℓ is strictly decreasing, Lemma 3 gives $\mathcal{L}_n = \ell(m_n) \downarrow \ell_\infty$. Equation (14) identifies this value with the global infimum $\mathcal{L}^\star$ over the optimized matrices. The same lemma bounds the gradient coefficient by

$$0 < g(u_n) \le \frac{2\kappa_m \lambda (T-1)}{r}\, u_n e^{-\kappa_m u_n^2} \longrightarrow 0.$$

The exponential factor dominates the factor u_n, proving the remaining limit. The argument applies separately at both precisions. □

E.3 Proof of Proposition 4

Proof. We first propagate the task-coordinate difference and accumulate the common coordinate. We then substitute these estimates into the exact one-step logit identity. Finally, we sum the one-step bounds. Throughout, both task coordinates lie in I_{E} and $\eta n < T_{\mathrm{E}}$.

Step 1. Compare the coordinates. Let $d_n = |u_n^{\mathrm{lo}} - u_n^{\mathrm{hi}}|$. Adding and subtracting $F_0(u_n^{\mathrm{lo}})$ in the recurrence gives

$$\begin{aligned} d_{n+1} &\le d_n + \eta |F_0(u_n^{\mathrm{lo}}) - F_0(u_n^{\mathrm{hi}})| + \eta |F_\Gamma(u_n^{\mathrm{lo}}) - F_0(u_n^{\mathrm{lo}})| \\ &\le (1 + \eta \Lambda_{\mathrm{E}}) d_n + \eta \Gamma^2 C_{\mathrm{E}}^{\mathrm{pert}}. \end{aligned} \tag{25}$$

The segment between the two task coordinates lies in I_{E}, so the second line uses the derivative and forcing bounds from (16). Since $d_0 = 0$, induction gives

$$d_n \le \Gamma^2 C_{\mathrm{E}}^{\mathrm{pert}} \eta \sum_{j=0}^{n-1} (1 + \eta \Lambda_{\mathrm{E}})^j \le \Gamma^2 C_{\mathrm{E}}^{\mathrm{pert}} \eta n e^{\Lambda_{\mathrm{E}} \eta n} \le \Gamma^2 C_{\mathrm{E}}^{u}. \tag{26}$$

Indeed, substituting the sum for d_n into (25) produces the same sum with one additional term. The remaining inequalities use $1 + x \le e^x$ and $\eta n < T_{\mathrm{E}}$; the case $\Lambda_{\mathrm{E}} = 0$ is included. For the common coordinate, $a_0^{\mathrm{lo}} = 0$ and $0 \le \kappa_j \le \Gamma C_{\mathrm{E}}^{\mathrm{com}}$ imply

$$0 \le a_n^{\mathrm{lo}} = \eta \sum_{j<n} \kappa_j \le \Gamma C_{\mathrm{E}}^{\mathrm{com}} \eta n \le \Gamma C_{\mathrm{E}}^{a}. \tag{27}$$

Step 2. Bound every term in one logit increment. Abbreviate $F_n^{\mathrm{lo}} = F_\Gamma(u_n^{\mathrm{lo}})$ and $F_n^{\mathrm{hi}} = F_0(u_n^{\mathrm{hi}})$. The triangle inequality gives

$$\begin{aligned} |F_n^{\mathrm{lo}} - F_n^{\mathrm{hi}}| &\le |F_\Gamma(u_n^{\mathrm{lo}}) - F_0(u_n^{\mathrm{lo}})| + |F_0(u_n^{\mathrm{lo}}) - F_0(u_n^{\mathrm{hi}})| \\ &\le \Gamma^2 C_{\mathrm{E}}^{\mathrm{pert}} + \Lambda_{\mathrm{E}} d_n \le \Gamma^2 C_{\mathrm{E}}^{F}. \end{aligned} \tag{28}$$

Since $\mathrm{Logit}_{\max,n}^{\mathfrak{b}} = \kappa_{\mathrm{com}} (a_n^{\mathfrak{b}})^2 + (C_h \kappa_m / \tau)(u_n^{\mathfrak{b}})^2$ and $a_n^{\mathrm{hi}} = 0$, expanding each square after one update yields

$$\begin{aligned} \Delta_{n+1} - \Delta_n = {} & \kappa_{\mathrm{com}} (2\eta a_n^{\mathrm{lo}} \kappa_n + \eta^2 \kappa_n^2) \\ & + \frac{C_h \kappa_m}{\tau} \{ 2\eta (u_n^{\mathrm{lo}} F_n^{\mathrm{lo}} - u_n^{\mathrm{hi}} F_n^{\mathrm{hi}}) + \eta^2 [(F_n^{\mathrm{lo}})^2 - (F_n^{\mathrm{hi}})^2] \}. \end{aligned} \tag{29}$$

The common term satisfies $|2\eta a_n^{\mathrm{lo}}\kappa_n+\eta^2\kappa_n^2| \le \Gamma^2\eta(2C_{\mathrm{E}}^a C_{\mathrm{E}}^{\mathrm{com}}+\eta_{\mathrm{E}}\big(C_{\mathrm{E}}^{\mathrm{com}}\big)^2)$. For the task product, add and subtract $u_n^{\mathrm{lo}}F_n^{\mathrm{hi}}$. Since $u_n^{\mathfrak{b}} \le u_{\mathrm{E}}$ and $F_n^{\mathfrak{b}} \le U_5$,

$$\begin{aligned}|u_n^{\mathrm{lo}}F_n^{\mathrm{lo}} - u_n^{\mathrm{hi}}F_n^{\mathrm{hi}}| &\le u_n^{\mathrm{lo}}|F_n^{\mathrm{lo}} - F_n^{\mathrm{hi}}| + |F_n^{\mathrm{hi}}|d_n \\ &\le \Gamma^2(u_{\mathrm{E}}C_{\mathrm{E}}^F + U_5 C_{\mathrm{E}}^u), \\ |(F_n^{\mathrm{lo}})^2 - (F_n^{\mathrm{hi}})^2| &= |F_n^{\mathrm{lo}} - F_n^{\mathrm{hi}}|\,|F_n^{\mathrm{lo}} + F_n^{\mathrm{hi}}| \le 2\Gamma^2 U_5 C_{\mathrm{E}}^F.\end{aligned}$$

Substitution into (29), followed by $\eta \le \eta_{\mathrm{E}}$, proves $|\Delta_{n+1} - \Delta_n| \le C_{\mathrm{E}}^{\mathrm{abs}}\Gamma^2\eta$ with exactly the constant in (17).

Step 3. Sum the increments. Both precisions have the same initial matrices, so $\Delta_0 = 0$. Therefore

$$|\Delta_n| \le \sum_{j=0}^{n-1} |\Delta_{j+1} - \Delta_j| \le C_{\mathrm{E}}^{\mathrm{abs}}\Gamma^2\eta n.$$

This proves the accumulated-gap bound. All constants used in the comparison are independent of Γ and of the allowed step size. □

F Proofs of Supporting Lemmas

F.1 Proof of Lemma 1

Proof. We compare the normalized task inputs exactly, differentiate the unperturbed update, and then bound the common update.

Step 1. Compare the task inputs. Fix $u > 0$ and abbreviate $g = g(u)$, $\mathcal{D}_0 = \sqrt{r}\,g + \varepsilon_{\mathrm{NS}}$, and $\mathcal{D}_\Gamma = \sqrt{rg^2 + \Gamma^2} + \varepsilon_{\mathrm{NS}}$. Rationalizing the square root gives

$$0 \le \frac{g}{\mathcal{D}_0} - \frac{g}{\mathcal{D}_\Gamma} = \frac{g\Gamma^2}{\mathcal{D}_0\mathcal{D}_\Gamma(\sqrt{rg^2+\Gamma^2} + \sqrt{r}\,g)} \le \frac{\Gamma^2}{2\sqrt{r}\,\mathcal{D}_0^2}. \tag{30}$$

The last line uses $\mathcal{D}_\Gamma \ge \mathcal{D}_0$ and $\sqrt{rg^2+\Gamma^2} + \sqrt{r}\,g \ge 2\sqrt{r}\,g$. The mean-value theorem on $[\rho_\Gamma(u), \rho_0(u)]$ proves the task-update estimate. Keeping $\mathcal{D}_0$, rather than replacing it by $\sqrt{r}\,g$, also retains the stabilizer in this bound.

Step 2. Differentiate the exact task update. The derivative of $g/(\sqrt{r}\,g + \varepsilon_{\mathrm{NS}})$ with respect to g is $\varepsilon_{\mathrm{NS}}/\mathcal{D}_0^2$. The chain rule therefore gives $F_0'(u) = \varepsilon_{\mathrm{NS}}\psi_5'(\rho_0(u))g'(u)/\mathcal{D}_0(u)^2$. $g'(u)$ is obtained directly from $g(u) = -(\kappa_m u/r)\ell'(\kappa_m u^2)$ as $g'(u) = (\kappa_m/r)[-\ell'(m) - 2m\ell''(m)]$, where $m = \kappa_m u^2$. Thus the constant Λ_{E} in (16) depends only on the fixed population loss, $\mathcal{G}_{\mathrm{E}}$, and the finite polynomial.

Step 3. Bound the common update. Its normalized input is $\Gamma/\mathcal{D}_\Gamma \in (0,1)$. Positivity and $\psi_5(t) \le \alpha_5 t$ from Lemma 4 imply $0 \le h_{\mathrm{com}}(g(u),\Gamma) \le \alpha_5\Gamma/\mathcal{D}_\Gamma(u)$. This completes the estimates. □

F.2 Proof of Lemma 2

Proof. We first record the feature identities needed for the matrix calculation. We then differentiate a single example with respect to the full query and key matrices. Finally, we average those gradients over the target and context types.

Step 1. Verify the feature identities. Let $\mathcal{E}_K = \mathrm{span}\{e_{c,i} : i \in [M]\}$ and $\mathcal{E}_Q = \mathrm{span}\{e_{z,i} : i \in [M]\}$, and let $\Pi_{\mathcal{E}_A}$ be the orthogonal projector onto the corresponding span. The two feature spans need not be distinct or orthogonal. Within either family, the Gram identity implies that the squared norm of the sum of its M vectors is $M - M = 0$, so the sum is zero. Also, $\sum_i e_{c,i}e_{c,i}^\top$ fixes every $e_{c,j}$ and vanishes on $\mathcal{E}_K^\perp$, and hence equals $\Pi_{\mathcal{E}_K}$. The same argument applies to the query family.

Applying $\sum_i \Phi_K e_{c,i}e_{z,i}^\top$ to $e_{z,j}$ gives $\Phi_K e_{c,j} = \Phi_Q e_{z,j}$. Both this sum and Φ_Q vanish on $\mathcal{E}_Q^\perp$, which proves the first sum identity in the lemma. Interchanging query and key proves the second. The simplex Gram matrix has rank $r = M - 1$, so each Φ_A, being an isometry on that span and zero on its orthogonal complement, has r unit singular values. Finally, $\Phi_A\overline{x} = 0$ and the assumed

orthogonality of $\widehat{k}, \widehat{q}$ to its image give $\Phi_A E_A^\top = 0$ and $\Phi_A^\top E_A = 0$. These are precisely the row- and column-space orthogonalities. The unit norms in E_A give $\|E_A\|_F = 1$.

Step 2. Differentiate before restricting the parameter directions. Let $\mathcal{L}_{\mathrm{ex}}(W_Q, W_K)$ denote the original cross-entropy for one complete example. Its restriction to the two-coordinate family is $\ell_{\mathrm{ex}}(m)$ from (12), but the following derivatives are taken before imposing that restriction. For context position j, write $L_j = (W_Q z_y)^\top W_K c_{i_j}$ and $S_j = \tau L_j$ for the scaled score. Adding a constant to every S_j does not change attention, so $\sum_j \partial \mathcal{L}_{\mathrm{ex}}/\partial S_j = 0$. Evaluate these derivatives at $W_A = u\Phi_A + aE_A$ and put $\xi = \sum_j (\partial \mathcal{L}_{\mathrm{ex}}/\partial S_j) e_{c,i_j}$. The chain rule gives

$$\nabla_{W_Q} \mathcal{L}_{\mathrm{ex}} = \tau \sum_j \frac{\partial \mathcal{L}_{\mathrm{ex}}}{\partial S_j} (W_K c_{i_j}) z_y^\top, \qquad \nabla_{W_K} \mathcal{L}_{\mathrm{ex}} = \tau \sum_j \frac{\partial \mathcal{L}_{\mathrm{ex}}}{\partial S_j} (W_Q z_y) c_{i_j}^\top.$$

Substituting (10) and cancelling the terms proportional to $\sum_j \partial \mathcal{L}_{\mathrm{ex}}/\partial S_j$ yields

$$\begin{aligned} \nabla_{W_Q} \mathcal{L}_{\mathrm{ex}} &= \tau \sigma_{\mathrm{type}} u \Phi_K \xi (\sigma_{\mathrm{com}} \overline{x} + \sigma_{\mathrm{type}} e_{z,y})^\top, \\ \nabla_{W_K} \mathcal{L}_{\mathrm{ex}} &= \tau \sigma_{\mathrm{type}} (\sigma_{\mathrm{com}} a \widehat{k} + \sigma_{\mathrm{type}} u \Phi_Q e_{z,y}) \xi^\top. \end{aligned} \tag{31}$$

The mixed terms in these expressions need not vanish for an individual example. Their population averages are determined next.

Step 3. Average using the type symmetry. Recall $\xi = \sum_j (\partial \mathcal{L}_{\mathrm{ex}}/\partial S_j) e_{c,i_j}$. Group the score derivatives by type as $d_i = \sum_{j: i_j = i} \partial \mathcal{L}_{\mathrm{ex}}/\partial S_j$, then we have $\xi = \sum_i d_i e_{c,i}$, $\sum_i d_i = 0$. Differentiating the score restriction $S_j = \tau s_{\mathrm{com}} + m(\mathbf{1}\{i_j = y\} - 1/M)$ with respect to m gives

$$\ell'_{\mathrm{ex}}(m) = \sum_j \frac{\partial \mathcal{L}_{\mathrm{ex}}}{\partial S_j} \big(\mathbf{1}\{i_j = y\} - 1/M\big) = d_y.$$

Conditional on y, the r nonmatching types remain exchangeable, since the event $N_y \geq 1$ is invariant under their permutations. Thus $\mathbb{E}[d_y \mid y] = \ell'(m)$, all $\mathbb{E}[d_i \mid y]$ for $i \neq y$ are equal, and their sum is $-\ell'(m)$. It follows that

$$\mathbb{E}[\xi \mid y] = \ell'(m) e_{c,y} - \frac{\ell'(m)}{r} \sum_{i \neq y} e_{c,i} = \frac{M}{r} \ell'(m) e_{c,y}.$$

Averaging over the uniform target gives $\mathbb{E}[\xi] = 0$, which removes the two mixed terms in (31). For the remaining query term,

$$\nabla_{W_Q} \mathcal{L} = \frac{\kappa_m u \ell'(m)}{r} \sum_{i=1}^{M} \Phi_K e_{c,i} e_{z,i}^\top = \frac{\kappa_m u \ell'(m)}{r} \Phi_Q.$$

The second sum identity gives $\nabla_{W_K} \mathcal{L} = (\kappa_m u / r) \ell'(m) \Phi_K$. This proves the full-gradient formula. □

F.3 Proof of Lemma 3

Proof. Fix the target and counts. Let $n_{\mathrm{match}} = N_y$ and $n_{\mathrm{other}} = T - N_y$. Equation (11) gives $\pi_y = n_{\mathrm{match}}/(n_{\mathrm{match}} + n_{\mathrm{other}} e^{-m})$ and $\pi_i = N_i e^{-m}/(n_{\mathrm{match}} + n_{\mathrm{other}} e^{-m})$ for $i \neq y$. Hence $\pi'_y = \pi_y(1-\pi_y)$ and $\pi'_i = -\pi_y \pi_i$ for $i \neq y$. Write $\widehat{p}_i = e^{\lambda \pi_i} / \sum_j e^{\lambda \pi_j}$ for the readout probabilities, put $\zeta_i = 1 - \pi_y + \pi_i$ for $i \neq y$ and $\zeta_y = 0$, and let $\nu_1 = \sum_i \widehat{p}_i \zeta_i$. Direct differentiation gives

$$\ell'_{\mathrm{ex}}(m) = \lambda \sum_i \widehat{p}_i \pi'_i - \lambda \pi'_y = -\lambda \pi_y \left[(1 - \widehat{p}_y)(1 - \pi_y) + \sum_{i \neq y} \widehat{p}_i \pi_i \right] = -\lambda \pi_y \sum_{i \neq y} \widehat{p}_i (1 - \pi_y + \pi_i),$$

where we used $\pi'_y = \pi_y(1 - \pi_y)$, $\pi'_i = -\pi_y \pi_i$ for $i \neq y$, and $1 - \widehat{p}_y = \sum_{i \neq y} \widehat{p}_i$. Therefore, by definition we have $-\ell'_{\mathrm{ex}}(m) = \lambda \pi_y \nu_1$.

There are finitely many count vectors and their probabilities do not depend on m, so the population loss is smooth and may be differentiated term by term. If $n_{\mathrm{other}} > 0$, then $0 < \pi_y < 1$ and $\nu_1 > 0$; if $n_{\mathrm{other}} = 0$, the example derivative is zero. Since $M, T \geq 2$, $\mathbb{P}(n_{\mathrm{other}} > 0 \mid y) = 1 - M^{-T}/Z_{M,T} > 0$, so $\ell'(m) < 0$ at every finite m. Moreover, $\zeta_i \leq 2(1 - \pi_y)$ implies $0 \leq -\ell'_{\mathrm{ex}}(m) \leq 2\lambda(1 - \pi_y) \leq 2\lambda(T-1)e^{-m}$. Averaging proves the upper bound. Finally, $\pi_y \to 1$ and $\pi_i \to 0$ for $i \neq y$ as $m \to \infty$, so $\ell(m)$ converges monotonically to $\log(e^\lambda + M - 1) - \lambda = \ell_\infty$. □

F.4 Proof of Lemma 4

Proof. We first show algebraically that each matrix iteration preserves the two gradient directions. We then bound the scalar polynomial and its derivative on the normalized input interval.

Step 1. Apply the matrix polynomial to the two orthogonal directions. By Lemma 2, $\Phi_A E_A^\top = \Phi_A^\top E_A = 0$, $\Phi_A \Phi_A^\top \Phi_A = \Phi_A$, and $E_A E_A^\top E_A = E_A$. Consequently, for $X = b_{\text{task}}\Phi_A + b_{\text{com}} E_A$ with real $b_{\text{task}}, b_{\text{com}}$, we have

$$XX^\top X = {b_{\text{task}}}^3 \Phi_A + {b_{\text{com}}}^3 E_A, \qquad (XX^\top)^2 X = {b_{\text{task}}}^5 \Phi_A + {b_{\text{com}}}^5 E_A.$$

One NS iteration therefore sends X to $\psi(b_{\text{task}})\Phi_A + \psi(b_{\text{com}})E_A$. The initial matrix $-g\Phi_A - \gamma E_A$ has squared Frobenius norm $rg^2 + \gamma^2$. Following the definition in Section 2, the internal normalization step of NS5 produces signed coefficients $-g/(\sqrt{rg^2+\gamma^2}+\varepsilon_{\text{NS}})$ and $-\gamma/(\sqrt{rg^2+\gamma^2}+\varepsilon_{\text{NS}})$. Applying the internal normalization followed by five iterations and using oddness of ψ_5 prove (24).

Step 2. Prove the global scalar bounds. The factorization $\psi(t) = t(\omega_0 + \omega_1 t^2 + \omega_2 t^4)$ shows that ψ preserves sign. Indeed, the quadratic in t^2 has positive leading coefficient and negative discriminant. Thus ψ_5 is positive for every positive input.

Let $I_{\text{NS}} = [0, 1.203]$. Evaluating ψ at the endpoints and at the real roots of ψ' gives $\psi(I_{\text{NS}}) \subset I_{\text{NS}}$ and $\psi([2/3, 1.203]) \subset [2/3, 1.203]$. Also $\psi(t) \geq t$ on $[0, 2/3]$. An iterate starting below $2/3$ therefore cannot decrease until it enters the latter interval, and after entering it cannot fall below $2/3$. This proves $\psi_5(t) \geq \min\{t, 2/3\}$. The inclusion $\psi(I_{\text{NS}}) \subset I_{\text{NS}}$ proves the upper bound 1.203 for inputs in $[0, 1]$. On I_{NS}, since $\omega_1 t^2 + \omega_2 t^4 \leq 0$, hence we have $\psi(t) \leq \omega_0 t$. Applying this five times gives $\psi_5(t) \leq \omega_0^5 t$. Direct differentiation gives $|\psi'| \leq \omega_0$ on $[0, 1]$ and $|\psi'| \leq 4$ on I_{NS}. The chain rule then gives $|\psi_5'| \leq \omega_0 4^4 < 882$.

□

G Numerical Illustration of the Two Regimes

We illustrate the two regimes in Theorem 1 with a simple simulation of the reduced scalar dynamics. We use $M = 4$, $T = 11$, $u_0 = 0.1$, $\lambda = 4$, $\kappa_m = 1$, $\kappa_{\text{com}} = 2$, $\eta = 0.002$, $\varepsilon_{\text{NS}} = 10^{-7}$, $\Gamma = 10^{-10}$, $d = d_h = 4$, $\tau = 1/2$, $\sigma_{\text{com}} = \sigma_{\text{type}} = \sqrt{2}$, and $\chi = 1$. The two trajectories differ only in the prescribed gradient perturbation Γ; Figure 11 shows the maximum attention logit under the two precisions. The early-overlap and late-stage excess growth behaviors observed in real training also emerge in the numerical simulation of this retrieval model, as indicated by the corresponding shaded regions in the plot.

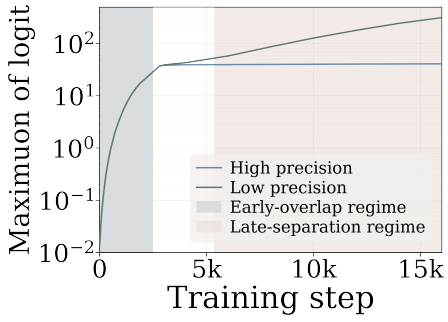


Figure 11: Maximum attention logit in the retrieval model.

H Experimental Details

Our implementation builds on Keller Jordan's `modded-nanoGPT` codebase and uses a FineWeb dataset. All experiments use the same training configuration summarized in Table 4. For QK-Norm, we use a parameter-free RMS normalization layer for both queries and keys. For both QK-Norm

Table 4: Training hyperparameters for the 124M/1B GPT-2.

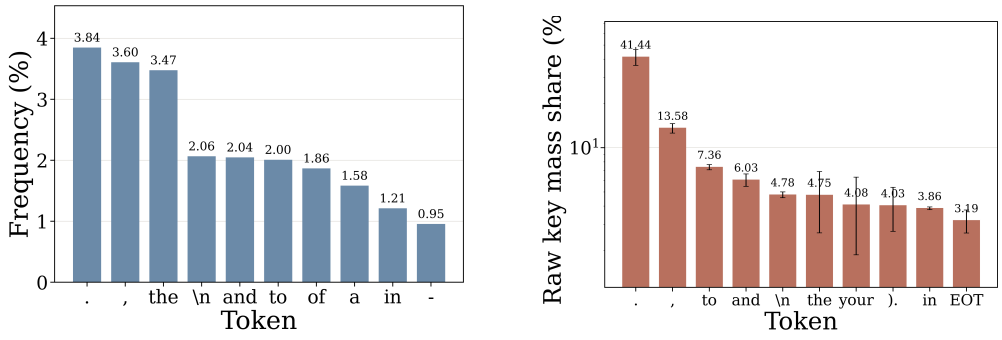
Hyperparameter	Value
Model	12 layers or 40 layers
Sequence length	1,024
Total batch size	1,048,576 tokens
Training steps	10,000
Training tokens	10.49 billion
Precision	BF16 or FP32
Warmup steps	700
Gradient clipping	1.0
Weight decay	0
Muon learning rate	1×10^{-2}
Adam learning rate	1×10^{-3}
Muon momentum	0.95
Muon weight decay	0
Validation interval	250 steps
Checkpoint interval	250 steps
Default seed	$42, 43, \ldots, 46$

and QK-Clip, we re-tune the learning rate and method-specific hyperparameters based on validation loss. Experiments on the 124M model are run on two NVIDIA H100 GPUs, with each run taking approximately 10 hours. Experiments on the 1B model are run on 32 NVIDIA H20 GPUs, with each run likewise taking approximately 10 hours. The kernel-level correction benchmarks are conducted on an NVIDIA A40 GPU and require only a few minutes per configuration.

I Additional Experiments

I.1 Token Frequency and Column-Selective Attention

The attention map in Figure 6 illustrates column-selective attention for a representative example. To examine whether this structure persists at the corpus level and how it relates to token frequency, we further aggregate the attention received by key positions according to their token identities over the test corpus. Figure 12 compares the most frequent tokens in the corpus with the tokens receiving the largest aggregate key-attention mass.



(a) Test-corpus token frequency.

(b) Key-attention mass share.

Figure 12: Corpus-level comparison between token frequency and received key-attention mass for the representative largest-logit RoPE head. (a) The ten most frequent tokens in the test corpus. (b) The ten token identities receiving the largest aggregate attention mass at key positions. EOT denotes the <|endoftext|> token.

The two rankings exhibit substantial overlap. Seven of the ten most frequent corpus tokens also appear among the ten tokens receiving the largest key-attention mass, and their attention-mass shares are often noticeably larger than their corpus frequencies. The distinction is also evident for EOT,

which receives the second-largest key-attention mass despite not appearing among the ten most frequent corpus tokens. This suggests that column selectivity is shaped not only by raw occurrence frequency, but also by token identity and learned structure. Overall, the corpus-level statistics support the observation in section 3.3: the largest-logit head concentrates attention on a small set of recurring token types.

I.2 Experiments for 1B model

We further evaluate the mitigation methods on the 1B model. Table 5 shows that all kernel-level methods effectively suppress excess logit growth and achieve lower validation losses than vanilla FA, QK-Norm, and QK-Clip. Moreover, `O32+P32` consistently improves over `O32` in both metrics across different positional encodings. These results are consistent with the observations in the main experiments.

Table 5: Mitigation results under RoPE at step $10{,}000$. Mean $\pm$ std over five seeds.

Setting	$\mathrm{Logit}_{\max}$ ($\downarrow$)	Val. loss ($\downarrow$)
Vanilla FA-16	46592 ± 10752	2.8631 ± 0.0011
`O32`	1580 ± 252	2.8624 ± 0.0005
`O16`	1568 ± 8	2.8627 ± 0.0005
`O32+P32`	1376 ± 80	2.8619 ± 0.0004
QK-Clip	796 ± 6	2.8630 ± 0.0006
QK-Norm	62.75 ± 0	2.8734 ± 0.0006

I.3 Additional Experiments under NoPE and sinusoidal

The main text uses RoPE to separate precision-dependent logit growth from ordinary task learning and to identify the structure of the numerical gradient errors. We now examine the same signatures under NoPE and sinusoidal. The additional measurements support a common qualitative picture of broad logit lift, nearly rank-one gradient discrepancies, and increasingly persistent reference directions.

Scope and notation. We use the training and paired-gradient conventions of Sections 3 and 3.2. We focus on the max-logit heads, namely layer 1 head 5 (L1H5) in the NoPE setting and layer 11 head 3 (L11H3) in the sinusoidal setting.

I.3.1 Positional architectures

Figure 13 shows the maximum attention-logit trajectories under RoPE, sinusoidal positional encoding and NoPE. They consistently exhibit the same two-stage behavior: the growth of the low-precision FA logits first enter an approximately linear growth regime, and is followed by a later re-acceleration with substantially faster excess growth. NoPE and Sinusoidal logits reach larger logits than RoPE. These results indicate that the extent of late-stage logit growth is architecture dependent, while the slowdown–re-acceleration pattern is consistently observed.

I.3.2 Common-mode lift within the affected head

Figure 14 extends the common-mode measurements in Figure 5 to NoPE and sinusoidal. In both heads, the minimum eventually becomes positive and rises together with the mean and maximum, indicating common-mode lift rather than an isolated extreme.

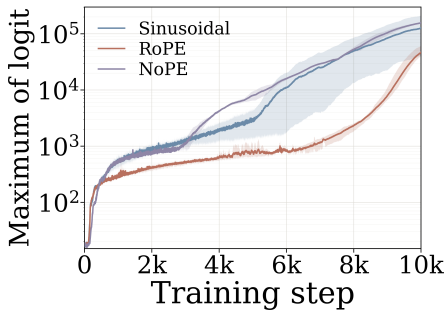
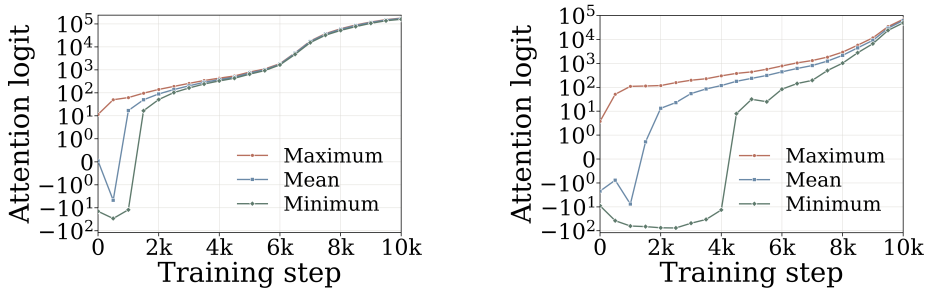


Figure 13: **Two-stage attention-logit dynamics across positional architectures.** RoPE, sinusoidal positional encoding, and NoPE consistently exhibit an early stage in which low-precision FA logit growth slows, followed by a late stage in which the logits re-accelerate and exhibit excess growth.

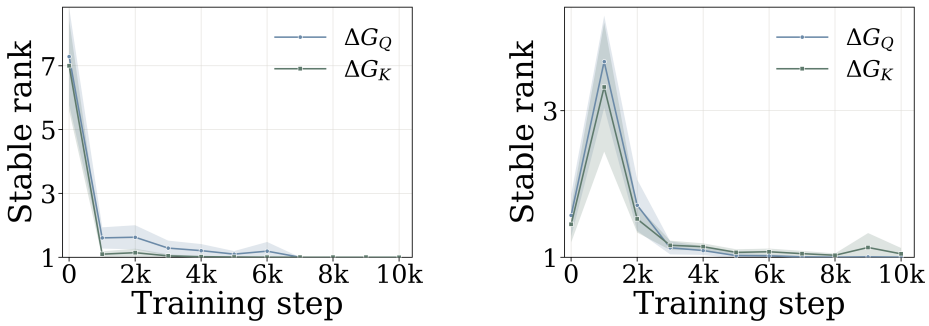


(a) NoPE, L1H5 (b) Sinusoidal, L11H3

Figure 14: **Joint growth of the minimum, mean, and maximum logits.** Statistics over finite causal entries in the indicated heads. Both heads develop a later common lift. These fixed-head quantities need not equal the maximum taken over all attention heads.

I.3.3 Rank-one structure and alignment of the gradient errors

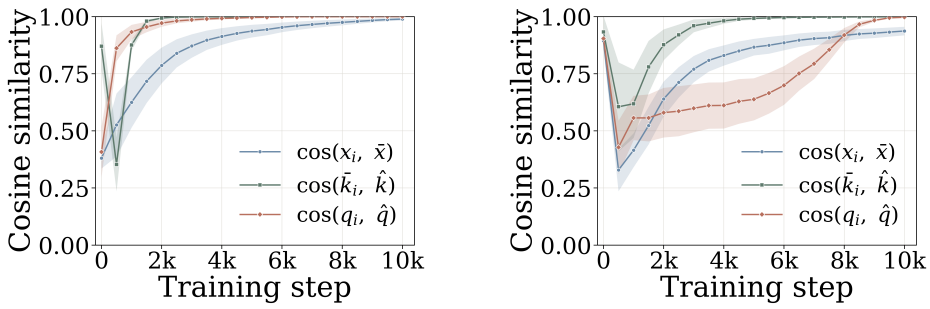
Spectral concentration appears in both architectures. Figure 15 applies $\mathrm{srank}(Z) = \|Z\|_F^2/\|Z\|_2^2$ to the paired discrepancies δG_Q and δG_K. NoPE ranks fall sharply from initialization, while sinusoidal first exhibits a transient peak. Both approach one at later checkpoints, extending the nearly rank-one error geometry beyond RoPE.



(a) NoPE, L1H5 (b) Sinusoidal, L11H3

Figure 15: **Nearly rank-one query- and key-gradient discrepancies beyond RoPE.** Stable ranks of ΔG_Q and ΔG_K during training, using the same definition as Figure 2(c). Despite different transients, both architectures approach stable rank one in the later checkpoints.

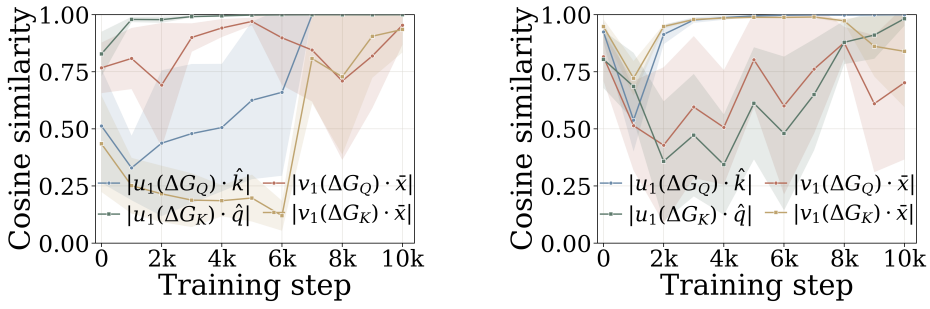
Input and effective query–key factors explain why mean directions are useful candidates. Figure 16 compares x_i with $\overline{x}$, the attention-weighted key $\overline{k}_i$ with $\widehat{k}$, and the query q_i with $\widehat{q}$, using the definitions in Section 3.2. These alignments strengthen over training in both examples. This behavior supports the same factor-based construction of the candidate directions as in the RoPE analysis.



(a) NoPE, L1H5 (b) Sinusoidal, L11H3

Figure 16: **Alignment of the factors in the gradient-error identities.** The three curves compare token inputs, effective keys, and effective queries with $\overline{x}$, $\widehat{k}$, and $\widehat{q}$, respectively. NoPE shows increasingly strong alignment of all three factors. In sinusoidal, the effective key aligns sooner than the query factor, and the input alignment remains below one. The dashed line marks perfect alignment.

Leading singular vectors provide a separate test of the proposed directions. Let $u_1(Z)$ and $v_1(Z)$ be the leading left and right singular vectors of Z. As shown in Figure 17, under the NoPE setting, the singular directions become well-aligned with the predicted directions after thousands of training steps. In the sinusoidal setting, both directions exhibit consistent alignment throughout most of the training trajectory, generally maintaining a cosine similarity greater than 0.5. Consequently, Eqn. 2 remains useful projection decompositions.

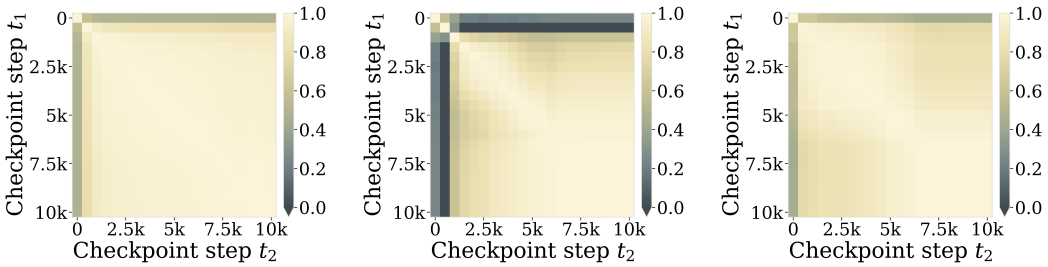


(a) NoPE, L1H5 (b) Sinusoidal, L11H3

Figure 17: **Singular-axis alignment with the proposed rank-one factors.** For each discrepancy, the left singular vector is compared with its predicted head-space direction and the right singular vector with $\overline{x}$.

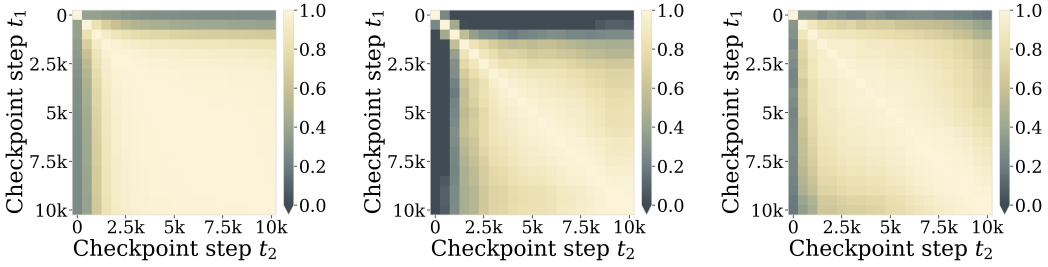
I.3.4 Temporal persistence of the reference directions

Figures 18 and 19 extend the checkpoint-to-checkpoint comparison in Figure 4. For a normalized direction $v_t \in \{\overline{x}_t, \widehat{k}_t, \widehat{q}_t\}$, a heatmap entry compares v_{t_1} and v_{t_2} by cosine similarity. In both architectures, broad high-similarity regions develop away from the initial checkpoints, showing that the reference directions change slowly over much of the later window.



(a) $\overline{x}$, layer 1 (b) $\widehat{k}$, L1H5 (c) $\widehat{q}$, L1H5

Figure 18: **Persistence of the mean-input and query–key directions under NoPE.** Each panel shows cosine similarity between the indicated direction at two checkpoints, with the displayed color bar giving its scale.



(a) $\overline{x}$, layer 11 (b) $\widehat{k}$, L11H3 (c) $\widehat{q}$, L11H3

Figure 19: **Persistence of the corresponding directions under sinusoidal.** Each panel shows cosine similarity between the indicated direction at two checkpoints, with the displayed color bar giving its scale. The same checkpoint-pair diagnostic is applied to the layer-11 mean input and the L11H3 query/key directions.