MECHANISM-DRIVEN MONITORS FOR PREEMPTIVE DETECTION OF LLM TRAINING INSTABILITY

Ruixuan Huang[1], **Yipei Wang**[2], **Wenyi Fang**[2], **Hantao Huang**[3], **Yifan Huang**[3], **Ansheng You**[3], **Zhenxing Zhang**[3], **Shuai Wang**[1], **Fan Wu**[2], **Yang Zheng**[2]

[1]HKUST [2]Huawei [3]Independent Researcher

ABSTRACT

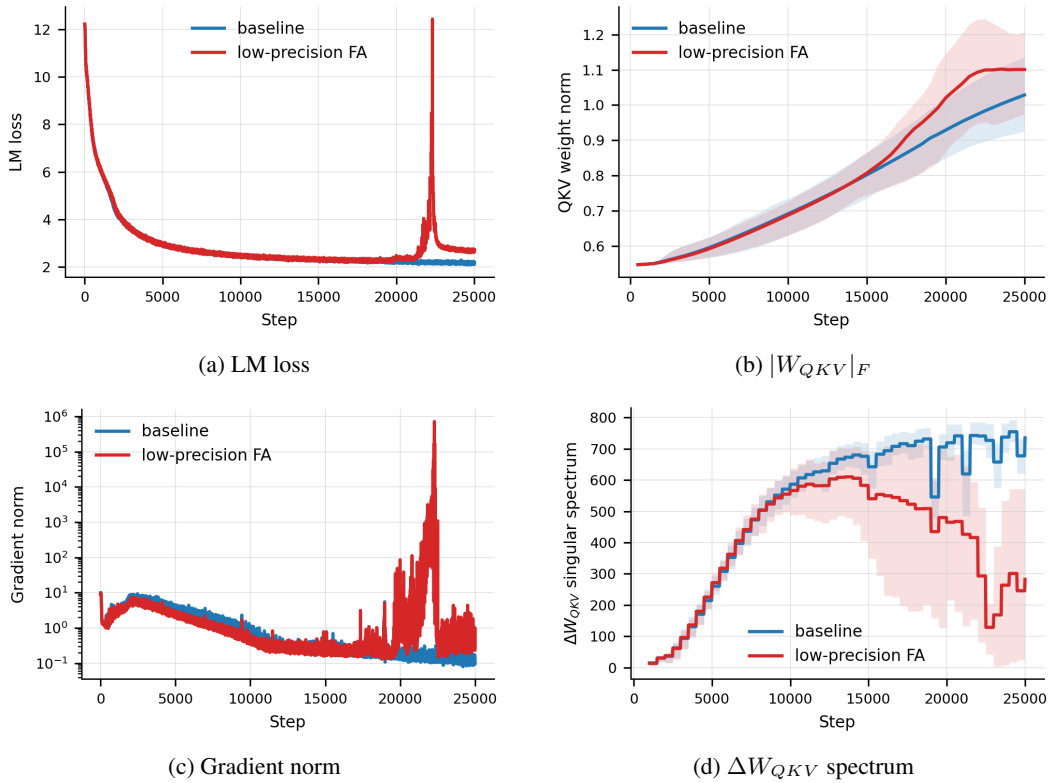
Frontier large language model training consumes massive accelerator fleets and long wall-clock computation, making stability failures costly when they occur. After a numerical or a hyperparameter fault has already destabilized the training dynamics, it may continue for thousands of steps while loss and gradient norms still appear normal. We study mechanism-driven detection of training instability by deriving internal monitors from the functional role of each critical module and from the earliest computational sites where failures are expected to produce measurable signatures. For low-precision flash attention, we monitor the spectral entropy of a QK bilinear decomposition, whose first-order term becomes abnormal before the loss fully collapses. For MoE routers, we derive indicators from their role in expert selection. Our fault-injection experiments on low-precision attention, large learning-rate, and combined faults show that these signals provide distinct signatures for different failures, triggering thousands of steps before loss divergence.

1 INTRODUCTION

Frontier large language model (LLM) training typically occupies thousands of accelerators for weeks to months (Chowdhery et al., 2022; Smith et al., 2022). Parameter counts now reach hundreds of billions to over a trillion (Yang et al., 2025; Kimi Team, 2025; DeepSeek-AI, 2026); pre-training corpora span tens of trillions of tokens (GLM-4.5 Team, 2025; Meituan LongCat Team, 2025; Qwen Team, 2026). DeepSeek-V3 gives an explicit cost accounting, where 14.8T pre-training tokens required 2.788M H800 GPU-hours and a reported $5.576M in direct rental cost, excluding prior research and ablation experiments (DeepSeek-AI, 2024). At this scale, training stability has become an engineering concern of the training system. GLM-130B describes unexpected 100B-scale training challenges, especially loss spikes and divergence (Zeng et al., 2023); DeepSeek-V3 highlights FP8 mixed-precision training and explicitly reports no irrecoverable loss spikes or rollbacks (DeepSeek-AI, 2024); and Kimi K2 introduces MuonClip with QK-clip to address training instability and reports 15.5T-token pre-training with zero loss spike (Kimi Team, 2025).

The risk of training instability often comes from two sources. The first is numerical precision error. For example, flash attention (FA) exhibits substantially larger BF16 numeric deviation than baseline attention in isolated forward passes (Golden et al., 2024), and low-precision FA can corrupt weight updates through biased rounding errors and gradually derail training dynamics (Qiu & Yao, 2026). The second is hyperparameter interaction, such as the coupling among global batch size (GBS), learning rate schedule and MoE auxiliary loss. However, before the global symptoms appear, a training run may already have entered an unstable state in its weights or optimizers, while training silently continues for thousands of steps before the symptoms become visible. Exhaustive ablation only increases these sunk costs. A useful monitor should therefore identify which subsystem has been destabilized before loss divergence appears.

Current training stability monitoring mainly relies on global training curves and symptom-level indicators. Loss, gradient norms, and weight norms are the most delayed indicators. Once a loss spike or divergence appears, the fault may already have been written into weights or optimizer state. Attention entropy, maximum attention logit, and spectral indicators further characterize attention-side instability symptoms (Zhai et al., 2023; Takase et al., 2025; Golden et al., 2024; Kimi Team, 2025).



(a) LM loss

(b) $|W_{QKV}|_F$

(c) Gradient norm

(d) ΔW_{QKV} spectrum

Figure 1: Monitoring signals over the first 25,000 steps of a training run. (a)–(c) are standard symptom-level indicators: LM loss, QKV weight norm, and gradient norm. (d) shows an internal update monitor used in this paper. In (b) and (d), the solid curve is the layer-wise average and the shaded band spans the 10th–90th percentile across all layers.

Edge-of-stability analysis explains high-level loss dynamics, but does not identify which module failed first (Cohen et al., 2021). Max-logit signals are difficult to expose in production FA because they require kernel modification and recomputation (Kimi Team, 2025). Hessian or curvature diagnostics can provide finer geometric information, but are too expensive to run as routine online checks at frontier scale (Yao et al., 2020; Kalra et al., 2026).

Our core idea is mechanism-driven monitoring. For each critical module, we ask what is this module supposed to compute, and where would a malfunction first leave an attributable trace. We apply this principle to two modules. For low-precision FA, we monitor weight updates, where low-precision backward errors first enter the model state (Section 3). We decompose the two-snapshot increment of the QK operator and monitor the spectral entropy of ΔW. Figure 1 shows an example under low-precision FA training, where the ΔW spectrum collapses thousands of steps earlier than loss, gradient norms and weight norms.

For MoE routers, the intended computation is discriminative and non-collapsed expert selection (Section 4). Therefore, we monitor router weight similarity and centered conditioning as weight indicators that characterize whether the effective expert-selection axes become redundant. For the behavior, we monitor per-token routing entropy. It reads the full softmax distribution and can therefore capture the collapse of routing behavior before downstream discrete quantities such as top-k counts, capacity overflow, or load-balance statistics change.

We further analyze how learning rate and GBS interact through stable-winner reinforcement. A larger learning rate amplifies coherent margin growth, while a smaller batch size increases margin noise; both can reduce router entropy and accelerate expert-use collapse. Our fault-monitoring experiments demonstrate the separate roles of the two monitor families, and combined faults inherit both signatures without obscuring their attribution.

2 Related Work

Training-stability monitors. Existing work has proposed monitors to detect or mitigate training instability. On the attention side, max-logit clipping, introduced for ViT-22B (Dehghani et al., 2023) and studied through small-scale Transformer proxies (Wortsman et al., 2024), catches softmax explosion directly. Kimi K2 adds QK-clipping and per-head MuonClip (Kimi Team, 2025). Other approaches target attention-entropy collapse via σReparam (Zhai et al., 2023), loss spikes via spectral-norm control (Takase et al., 2025), or Flash-Attention output distributions (Golden et al., 2024). On the MoE router side, LongCat-Flash monitors the average cosine similarity among expert router weights and the gradient-norm ratio between the load-balancing objective and the language-modeling objective on average expert probabilities (Meituan LongCat Team, 2025).

Attention Circuit Analyses. The attention QK-circuit has been analyzed primarily as a static object. Bao et al. (2024) characterize attention localization through the eigenspectrum variance of $W_q^\top W_k$, and Pan et al. (2024) examine singular-vector correspondence on the QK kernel for vision transformers. Researches show that attention maps and QK kernels can exhibit strong low-rank structure. Bhojanapalli et al. (2020) study the rank-deficiency bottleneck of $W_q W_k^\top$ at small d_k, while Dong et al. (2021) prove doubly-exponential rank collapse in pure self-attention with depth. Recent works also show that weight updates contain informative low-rank structure. LoRA (Hu et al., 2022), GaLore (Zhao et al., 2024), and the Muon optimizer family (Liu et al., 2025) exploit low-rank or spectral structure in updates for parameter-efficient adaptation or optimization, and Yunis et al. (2024) survey spectral evolution of weights as a window onto training dynamics. Mechanistically, Qiu & Yao (2026) identify low-precision FA failure as biased rounding accumulating into similar low-rank update directions. These works motivate using ΔW itself as an analysis object.

3 Attention Updates Monitoring

Flash Attention (Golden et al., 2024) fuses the softmax-scaled $QK^\top$ computation in on-chip memory and not materializes the full $N \times N$ logit matrix. This brings dramatic memory and throughput gains, and modern LLM training and inference now rely on it as the prevalent attention implementation. However, FA is reported as a source of training instability, where low-precision arithmetic in its backward pass can deposit persistent, biased errors into weight updates (Kimi Team, 2025; Qiu & Yao, 2026). Moreover, the fused implementation blocks the most natural symptom-level monitor used at scale, namely tracking the maximum attention logit (max-logit) to catch softmax explosion (Kimi Team, 2025). Reading max-logits out of a production FA kernel requires either invasive kernel modification or a recompute pass, both unacceptable in a large training run.

Runtime training monitors are viable only for quantities that require no kernel modification or activation recomputation. For instance, use gradients or weights W directly (Fang et al., 2023). However, in practice, gradient-based indicators are dominated by mini-batch noise across consecutive steps, and W-based indicators are diluted by initialization energy. The natural remaining target is the parameter update ΔW itself, which is exactly the level at which low-precision FA faults have been shown to deposit their persistent damage (Qiu & Yao, 2026).

In fact, modern LLMs are severely over-parameterized, with intrinsic dimension far below parameter count (Jacot et al., 2018; Chizat et al., 2019; Lee et al., 2019), so a fault that perturbs along currently low-impact directions is absorbed silently—loss lags not by how long corruption takes to occur, but by how long it takes the corrupted directions to become task-loaded. Among parameter-side quantities, the update ΔW is preferable to the raw weight W on signal-to-noise grounds: singular-value statistics of W_t are diluted by initialization energy (Appendix A), whereas the increment $\Delta W_{t,\delta} = W_t - W_{t-\delta}$ removes this background and exposes the update geometry directly.

3.1 The Intrinsic Low-Precision Issue of Flash Attention

As low-precision arithmetic becomes standard in LLM training, Qiu and Yao (Qiu & Yao, 2026) show that low-precision FA induces biased scalar errors in $\delta = \mathrm{rowsum}(dO \odot O)$, and that these biased scalars multiply structurally coherent rank-one update atoms. We adopt their per-step source model as the basis for update-side monitoring. Following and simplifying their notation, let $X \in \mathbb{R}^{N \times d}$ be the hidden-state matrix entering the query projection, $K = XW_k$, and

$P = \mathrm{softmax}(QK^\top/\sqrt{d_k})$ be the attention probability matrix. For a token-step sample j, let X_j and $(PK)_j$ denote the corresponding rows. Following their mechanism, index token-step samples by j and write the update-side source as $e_j R_j$, where e_j is the biased scalar error induced through δ and $R_j = X_j^\top (PK)_j$ is the associated rank-one update atom, up to the attention scale and sign. Here e_j corresponds to Qiu and Yao's biased coefficient $(\delta_{lp} - \delta_{hp})[T]$, and R_j to their common low-rank error direction $\mathbf{R} \approx (\mathbf{PK})[T]^\top X[T]$ (their Claim 2, Equation 3). The monitoring premise is that biased scalar coefficients and coherent atoms produce a low-rank mean component in accumulated update windows:

> **Observation 1 (accumulation consequence of Qiu–Yao).** Index token-step samples by j and write $R_j = X_j^\top (PK)_j \in \mathbb{R}^{d \times d_k}$. If $M = \mathbb{E}[e_j R_j]$ has effective rank $r \ll d_k$ (this is the substantive premise; since $M = \mathbb{E}[e_j R_j] \in \mathbb{R}^{d \times d_k}$, $\mathrm{rank}(M) \le d_k$ automatically, supported by the empirical finding in Qiu & Yao (2026) that the atoms R_j share common column structure across tokens and training steps; see their Figure 4) and the centered fluctuations $e_j R_j - M$ are independent (or martingale-difference) with bounded second moment, then over n samples
>
> $$\sum_{j=1}^{n} e_j R_j = nM + O_p(\sqrt{n}). \tag{1}$$
>
> The coherent low-rank component grows linearly in n, while zero-mean residuals grow sublinearly. Once $n\|M\|_2$ dominates the residual, the singular spectrum of the accumulated update is controlled by M.

Observation 1 is a concentration restatement of Qiu–Yao's accumulation mechanism for the windowed setting: it turns their per-step source model into a prediction that accumulated ΔW spectra should develop a low-rank component. Its short proof and a biased-rounding perturbation note are in Appendix B. Observation 1 predicts that spectral concentration will eventually emerge but does not predict the detection-onset step; see Section 6 for the quantitative gap.

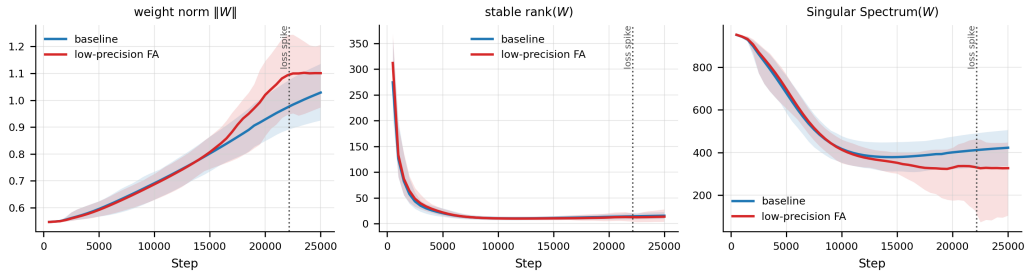
3.2 ΔW Spectral Indicators

Let $\Delta W = W_t - W_{t-\delta}$ for sampling interval δ. The structural state of ΔW can be summarized by various mathematical quantities. Given its singular values $\sigma_1 \ge \cdots \ge \sigma_r$, the *stable rank* $\mathrm{srank}(\Delta W) = \|\Delta W\|_F^2 / \|\Delta W\|_2^2$ (Ipsen & Saibaba, 2024; Roy & Vetterli, 2007) measures the ratio of the squared Frobenius norm to the squared spectral norm, which is the inverse of how much the top-1 singular value dominates the spectrum. However, this metric loses information about the rest of the spectrum, and thus lacks interpretability – reaching full stable rank requires all singular values to be equal, which is not the case in practice. On the other hand, effective rank $\mathcal{S}_\alpha(\Delta W) = \exp\big(-\sum_i p_i \log p_i\big)$ with $p_i = \sigma_i^\alpha / \sum_j \sigma_j^\alpha$ is another way of evaluating the state of the spectrum. Empirically, to balance sensitivity and noise, we use $\alpha = 2$, which is also known as *singular spectrum* (Alter et al., 2000).

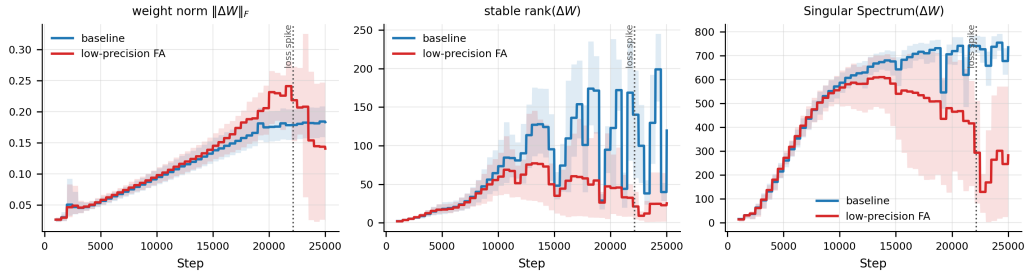
3.3 The ΔW Monitor in Practice

Following the low-precision FA mechanism of Qiu & Yao (2026), we compare the training of LLMs between baseline and a biased low-precision fault injection described in Appendix B. As shown in Figure 1a, the loss curve of the low-precision run diverges at $\sim 22{,}000$, while the baseline run remains stable. Traditionally, LLM training practitioners monitor weight-related metrics such as $\|W\|_F$ and $\mathrm{stable\_rank}(W)$, etc. to detect potential instability. However, Figure 2a shows that because of the lazy regime, analyzing matrix properties of W yields limited insight into the instability of the low-precision run. We stress that the fault certificate is the *deviation* of the ΔW spectrum from the healthy baseline trajectory–not low-rankness per se–since healthy updates already carry low-rank structure that LoRA, GaLore, and the Muon family exploit; this is why we compare the baseline and fault runs rather than reading low rank off a single trace.

The weight metrics for ΔW are plotted in Figure 2b. The singular spectrum of ΔW shows observable spectrum collapse at $10{,}000 \sim 14{,}000$, thousands of steps before the loss diverges. The stable



(a) Monitored metrics of the weight matrix W, including (1) norm, (2) stable rank, (3) singular spectrum



(b) Monitored metrics of the weight update ΔW, including (1) norm, (2) stable rank, (3) singular spectrum

Figure 2: Weight-side spectral monitors under the low-precision FA fault, over the first 25,000 steps. (top) Metrics of the weight matrix W and (bottom) metrics of the weight increment $\Delta W = W_t - W_{t-\delta}$. The raw W statistics are diluted by initialization energy and reveal little, whereas the ΔW increment exposes the update geometry.

rank of ΔW does show some early signals, but the instability nature of stable rank adds noise to the signal. Apart from the singular-value-based metrics, the update norm $\|\Delta W\|_F$ does not show a clear signal of instability until the loss diverges. This suggests that the low-rank structure of ΔW is not explained by a massive energy blow-up or a few overwhelmingly large update entries, but rather a global low-rank structure. Overall, the singular spectrum of ΔW is a more sensitive and explainable metric for detecting early signs of instability in the low-precision FA module.

3.4 The Bilinear Decomposition $\Delta_1 = \Delta_2 + \Delta_3$

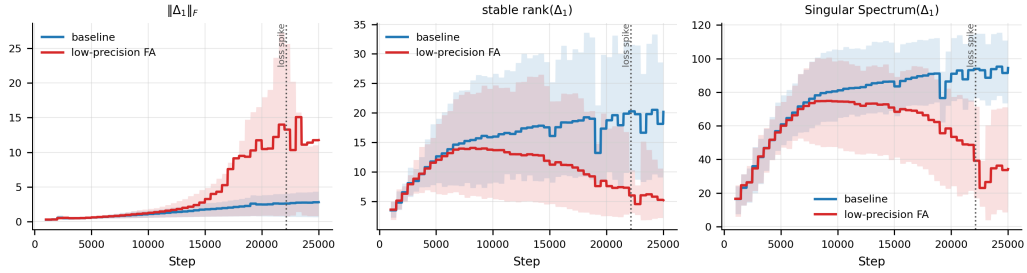
The ΔW monitor of the previous section treats W_q and W_k as independent matrices and computes spectral metrics on ΔW_q, ΔW_k, or their concatenation $[\Delta W_q, \Delta W_k]$. However, in FA, the attention score $QK^\top = X(W_q W_k^\top)X^\top$ depends on W_q and W_k only through the *bilinear form* $F(W_q, W_k) = W_q W_k^\top$, so monitoring the factors separately can miss correlated drifts that cancel or amplify in the product. A natural alternative is to track F itself, but direct spectral analysis of $W_q W_k^\top$ across snapshots offers limited discriminability: slow secular trends dominate, and the signal of interest is buried. This motivates decomposing the *increment* $\Delta_1 = F_t - F_{t-\delta}$ into components with distinct physical and spectral signatures. This increment of F admits an exact decomposition into a first-order term Δ_2 and a second-order term Δ_3.

Proposition 2 (bilinear decomposition). For $W_{q,t}, W_{k,t}$ at two time points and $\Delta W_q = W_{q,t} - W_{q,t-\delta}$, $\Delta W_k = W_{k,t} - W_{k,t-\delta}$, with the un-subscripted factors evaluated at the base point $t - \delta$ (i.e. $W_q := W_{q,t-\delta}$ and $W_k := W_{k,t-\delta}$),

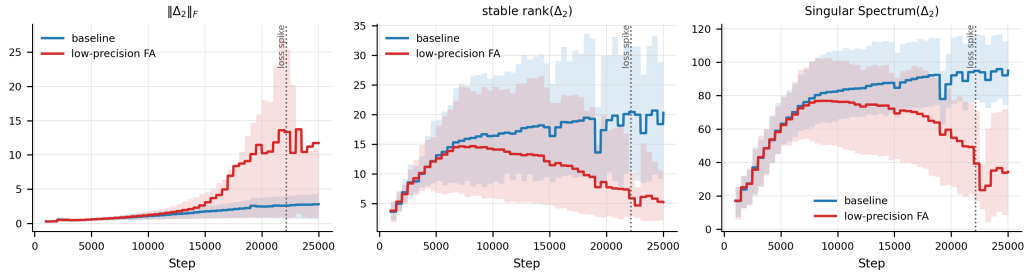
$$\Delta_1 := W_{q,t} W_{k,t}^\top - W_{q,t-\delta} W_{k,t-\delta}^\top = \Delta_2 + \Delta_3, \tag{2}$$

where $\Delta_2 = \Delta W_q W_k^\top + W_q \Delta W_k^\top$ and $\Delta_3 = \Delta W_q \Delta W_k^\top$.

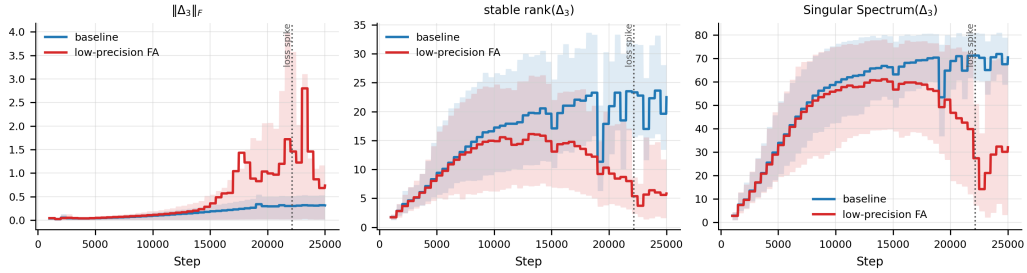
This follows immediately from the bilinearity of F. In particular, $DF[(\Delta W_q, \Delta W_k)] = \Delta_2$ and $\frac{1}{2} D^2 F[(\Delta W_q, \Delta W_k)^{\otimes 2}] = \Delta_3$, with all higher derivatives vanishing identically.



(a) Δ_1: exact QK-product increment



(b) Δ_2: first-order QK-product increment



(c) Δ_3: second-order Q/K update interaction

Figure 3: QK-product increment monitors under the low-precision FA fault. Δ_1 is the exact increment of $W_q W_k^\top$, Δ_2 is its first-order term, and Δ_3 is the second-order interaction between ΔW_q and ΔW_k.

Magnitude regime. In the early-to-mid training regime where $\|W\|_F \gg \|\Delta W\|_F$, and absent cancellation between the two first-order terms $\Delta W_q W_k^\top$ and $W_q \Delta W_k^\top$, we have $\|\Delta_2\|_F \gg \|\Delta_3\|_F$ by a factor of order $\|W\|_F / \|\Delta W\|_F$. We therefore monitor *shape*, not magnitude: the singular-spectrum entropy of Δ_2, the dominant first-order signal. The second-order term Δ_3 remains part of the exact decomposition, but its spectral shape can still expose Q/K update coupling once the interaction becomes coherent.

Exact low-rank spectral computation. Although Δ_1, Δ_2, and Δ_3 are formally $d \times d$ QK-product increments, their nonzero singular spectra can be computed exactly from small cores. For any $A, B \in \mathbb{R}^{d \times r}$ with thin decompositions $A = Q_A R_A$ and $B = Q_B R_B$,

$$AB^\top = Q_A (R_A R_B^\top) Q_B^\top, \tag{3}$$

so the nonzero singular values of $AB^\top$ are those of the $r \times r$ core $R_A R_B^\top$. Applied to a single attention head,

$$\begin{aligned} \Delta_3 &= \Delta W_q \Delta W_k^\top, \\ \Delta_2 &= [\Delta W_q, W_q][W_k, \Delta W_k]^\top, \\ \Delta_1 &= [W_{q,t}, W_{q,t-\delta}][W_{k,t}, -W_{k,t-\delta}]^\top, \end{aligned} \tag{4}$$

with ranks at most d_k, $2d_k$, and $2d_k$, respectively. Thus monitoring does not require materializing a dense $d \times d$ product when $d \gg d_k$; it only requires the singular spectrum of a head-dimensional core.

Size	Full eigvalsh (ms)	Core+eig. (ms)	Speedup	Rel. diff
100	0.55	3.80	0.14x	9.89×10^{-6}
200	39.23	4.54	8.64x	2.78×10^{-7}
500	66.04	3.86	17.12x	2.52×10^{-7}
1000	166.37	3.38	49.18x	2.50×10^{-7}
2000	1512.59	30.09	50.26x	2.29×10^{-7}
4000	3602.54	124.01	29.05x	2.26×10^{-7}

Table 1: Single Ascend 910B NPU timing for full-matrix eigendecomposition versus compressed-core computation of the same singular-spectrum quantities.

The architectural rank cap itself is not the anomaly–the signal is spectral concentration among the nonzero singular modes. On a single Ascend 910B NPU, the compressed-core computation gives large speedups at realistic hidden sizes while preserving the spectrum to small relative error (See Table 1).

Empirical ordering of the QK-product increments. Empirically, Δ_1 and Δ_2 detect the low-precision FA fault almost simultaneously, while Δ_3 deviates later; all three precede the raw ΔW spectrum. This ordering is consistent with the scale separation above (Figure 3). Since

$$\Delta_1 = \Delta_2 + \Delta_3, \qquad \|\Delta_2\|_F = O(\|W\|_F\|\Delta W\|_F), \qquad \|\Delta_3\|_F = O(\|\Delta W\|_F^2), \tag{5}$$

the early-to-mid training regime $\|\Delta W\|_F \ll \|W\|_F$ implies $\Delta_1 \approx \Delta_2$. Thus the exact QK-product increment and its first-order part expose the fault at nearly the same time. The interaction term Δ_3 is weaker because it is second order, but it still lives directly in Q/K update-coupling space and can become visible before spectral concentration is obvious in the separate factor updates $\Delta W_q, \Delta W_k$. This also explains why the QK-product increment curves are smoother: they aggregate the update through the functional QK product, suppressing factor-wise noise while preserving coherent Q/K drift.

4 MoE Router Monitoring

The MoE module is central to the frontier transformer architecture; most >30B LLMs are now MoE-based (Meituan LongCat Team, 2025; Kimi Team, 2025; Wang et al., 2024). In a top-k MoE layer (Jacobs et al., 1991; Shazeer et al., 2017), a lightweight router gating function selects which experts process each token. Concretely, with the router weight matrix $W_R = [w_1, \ldots, w_n] \in \mathbb{R}^{d\times n}$, each token x receives expert scores $s = W_R^\top x$, a softmax produces a routing distribution, and the top-k entries dictate which of the n experts are actually invoked. Although W_R typically holds well below 0.1% of the parameters of a single MoE layer, the choices it makes determine which of the remaining 99.9% are exercised on any given token. This asymmetry between trivial parameter count and outsized influence on capacity utilization makes the router the natural place to look for MoE-specific stability pathologies. Because the router is a small linear map that is usually independent from any parallelism scheme, its internal state, even activations, can be monitored without cross-device communication.

A healthy router maintains diversity along both the expert and token axes. Its weight columns should span distinct directions so that per-token routing distributions do not collapse. We study router stability through internal-state indicators that quantify this diversity directly.

4.1 Router Conditioning and Weight Similarity

The router selects experts through a softmax gate, which is shift-invariant: $\mathrm{softmax}(s) = \mathrm{softmax}(s - c)$ for any constant c. Setting $c = \bar{w}^\top x$, where $\bar{w} = \frac{1}{n}\sum_i w_i$ is the mean of router weights, shows that the routing decision depends only on the centered weights $(w_i - \bar{w})^\top x$. The ratio between the maximum deviation of router weights and the mean of router weights, i.e., $\varepsilon := \frac{\max_i \|w_i - \bar{w}\|}{\|\bar{w}\|}$ (defined for $\bar{w} \neq 0$, with all router columns nonzero), is a natural **conditioning ratio** for the router: it measures how large the discriminative deviations $(w_i - \bar{w})$ are relative

Model	GPT-OSS-20b	GPT-OSS-120b	Qwen3-35B-A3B	GLM-4.7-Flash	DeepSeek-V2	DeepSeek-V4-flash	DeepSeek-V4-pro
$\mathrm{sim}(W_R)$	-0.021 ± 0.007	0.000 ± 0.005	0.513 ± 0.083	0.268 ± 0.028	0.026 ± 0.055	0.179 ± 0.071	0.140 ± 0.024

Table 2: Router weight similarity $\mathrm{sim}(W_R) = \mathbb{E}_{i\neq j}[\cos(w_i, w_j)]$ across open-source MoE checkpoints, reported as mean±std over all MoE layers (number of experts $n = 32, 128, 256, 64, 160, 256, 384$, respectively).

to the common mode $\bar{w}$ that softmax discards. Note that softmax removes $\bar{w}$ exactly, so a small ε does not corrupt the routing decision itself; rather, it signals an ill-conditioned, near-redundant parameterization, in which the large common mode $\bar{w}$ dominates the stored weights, leaving the discriminative component $(w_i - \bar{w})$ with poor relative conditioning. Similarly, Meituan LongCat Team (2025) mention that they use router weight similarity $\mathrm{sim}(W_R) = \mathbb{E}_{i\neq j}[\text{cosine\_similarity}(w_i, w_j)]$ as an indicator during the LLM training, and we can see that the pairwise weight similarity is lower-bounded by a monotone function of this conditioning ratio, through the following proposition

Proposition 3 (Conditioning Ratio Lower-Bounds Router Weight Similarity).

$$\mathrm{sim}(W_R) \geq 1 - \frac{n}{n-1}\,\varepsilon^2 \tag{6}$$

This shows that the conditioning ratio ε controls a lower bound on the router weight similarity: as $\varepsilon \to 0$ the similarity approaches 1, i.e. the expert columns collapse onto the common mean and the router becomes redundant and non-discriminative (the high-similarity, low-stability regime). We defer the proof to Appendix C.

We measure $\mathrm{sim}(W_R)$ across open-source MoE checkpoints; results are summarized in Table 2. It can be found that different MoE architectures have very different router weight similarity, and the operating point is strongly architecture-dependent— from near-orthogonal router columns in the GPT-OSS family ($\mathrm{sim}(W_R)\approx 0$) to highly aligned columns in Qwen3-35B-A3B ($\mathrm{sim}(W_R)\approx 0.51$). Note that high similarity *does not necessarily imply a collapsed prediction, but is a risk of low stability due to high redundancies.*

Its computational complexity can be reduced to $O(nd)$ by

$$\mathrm{sim}(W_R) = \frac{n\|R\|_2^2 - 1}{n-1}, \text{ where } R = \frac{1}{n}\sum_{i=1}^{n}\frac{w_i}{\|w_i\|} \tag{7}$$

and is therefore a very affordable metric to monitor during training.

4.2 The Effective Component of Routers Is Learning-Rate Sensitive

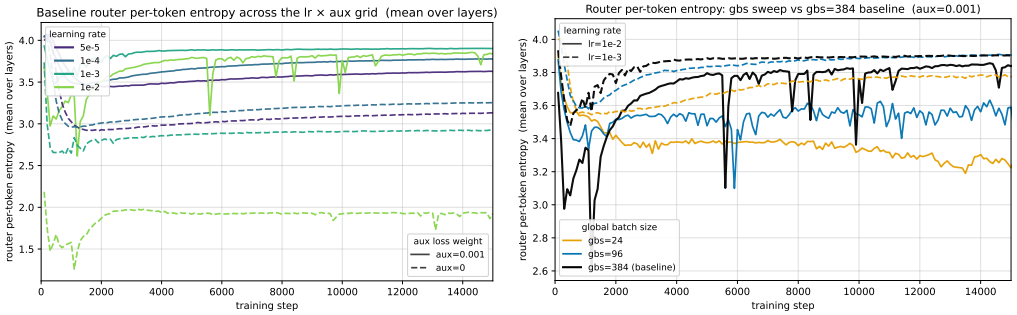
The similarity analysis above isolates the router directions that distinguish experts. In matrix form, let $C_n = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $W_{R,c} = W_R C_n = [w_1 - \bar{w}, \ldots, w_n - \bar{w}]$. Only this centered component changes the centered logits $\delta(x) = W_{R,c}^\top x$; the common-mode component adds the same scalar to every expert score and is removed by softmax. Token dispatch is therefore controlled by centered margins $m_{ij}(x) = (w_i - w_j)^\top x$, rather than by the raw router norm.

The router can fail at either extreme. When the routing distribution is nearly uniform ($H(p) \to \log n$), many experts have nearly tied scores and the router is uncertain, a failure mode studied by Wu et al. (2024). When the distribution is nearly a point mass ($H(p) \to 0$), tokens are routed in a singleton-like way and the model loses expert diversity. Modern MoE systems therefore try to keep expert use balanced through auxiliary or loss-free balancing mechanisms and capacity controls (Fedus et al., 2021; Wang et al., 2024); however, those quantities are realized only after top-k assignment or aggregation over a batch.

Per-token entropy is a more sensitive readout because it is continuous in the full softmax distribution,

$$H(p(x)) = -\sum_i p_i(x)\log p_i(x). \tag{8}$$

Maximal violation (MaxVio) (Wang et al., 2024), capacity overflow, and load-balance counts are downstream discrete readouts: they can stay unchanged while a leading expert's probability grows



(a) Layer-average entropy vs LRs

(b) Layer-average entropy vs GBS

Figure 4: Router per-token entropy under different learning rates and GBS.

inside an already-fixed top-k set. Locally, this behavioral entropy drop is tied to centered logit energy,

$$\log n - \mathbb{E}_x\, H(p) \;\approx\; \frac{1}{2n}\,\mathrm{tr}\big(W_{R,c}^{\top} M_x\, W_{R,c}\big), \tag{9}$$

where $M_x = \mathbb{E}[xx^{\top}]$. This link is used only as a local consistency check; the actual collapse certificate is the behavior-side entropy on current tokens.

When a stable-winner feedback loop is present, a large learning rate amplifies it. For a softmax gate with MoE output $y(x) = \sum_i p_i(x)E_i(x)$, define $q_i(x) = \langle \partial \ell/\partial y, E_i(x)\rangle$ and $r_i(x) = \partial \ell/\partial s_i = p_i(x)(q_i(x) - \sum_j p_j(x)q_j(x))$, so $\mathbf{1}^{\top} r = 0$ and $\nabla_{W_R}\ell(x) = x\, r(x)^{\top}$ (Jacobs et al., 1991; Shazeer et al., 2017). We analyze this mechanism under the dense-softmax relaxation of the gate, treating $y(x) = \sum_i p_i(x)E_i(x)$ as a mixture over all experts and thus setting aside top-k hard selection, expert-capacity limits, and the fact that some implementations route gate gradients only through the selected experts. The margin-feedback argument then applies to the soft gate scores s_i that drive selection. Consider a coherent token region Ω in which expert $j^{\star}$ is already a slight winner, and let h be any competing expert. Define the time-indexed winner–competitor margin on a probe token x' as

$$m_{j^{\star}h}(x', t) = s_{j^{\star}}(x', t) - s_h(x', t) = (w_{j^{\star}}(t) - w_h(t))^{\top} x'.$$

For the strongest competitor $h^{\star}$ in this local window, $m_{j^{\star}h^{\star}}$ is the top-two margin. For the mechanism sketch, write one plain stochastic-gradient step on token x as $w_i(t+1) = w_i(t) - \eta x r_i(x, t)$; adaptive optimizers replace this by their preconditioned effective step, but the same margin-coherence argument applies. This update changes the margin on x' by

$$\Delta m_{j^{\star}h}(x', t) = -\eta(x'^{\top} x)\big(r_{j^{\star}}(x, t) - r_h(x, t)\big).$$

Therefore define the time-dependent reinforcement coefficient

$$\gamma_h(t) = -\mathbb{E}_{x,x'\in\Omega}\left[(x'^{\top} x)(r_{j^{\star}}(x, t) - r_h(x, t))\right]. \tag{10}$$

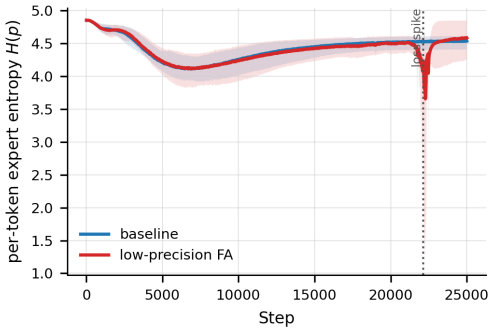
When $\gamma_h(t) > 0$, the update reinforces the current winner on average over nearby tokens in Ω instead of pulling it back toward its competitors. Summing the one-step recurrence gives

$$\mathbb{E}[m_{j^{\star}h}(x', T)] \gtrsim m_{j^{\star}h}(x', 0) + \eta \sum_{t<T} \gamma_h(t). \tag{11}$$

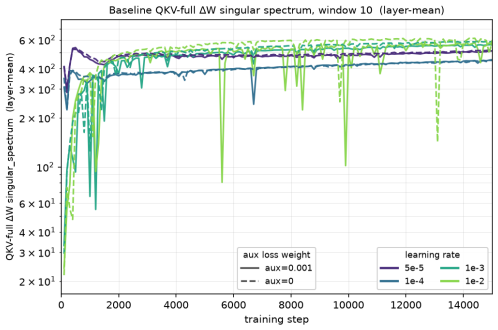
So η scales each reinforcement increment directly. For the strongest competitor, write the gap as $G_T = \eta \sum_{t<T} \gamma_{h^{\star}}(t)$, so $m_{j^{\star}h^{\star}}(x', T) \gtrsim m_{j^{\star}h^{\star}}(x', 0) + G_T$. Since

$$1 - p_{(1)}(x', T) \le (n-1)e^{-m_{j^{\star}h^{\star}}(x', T)} \lesssim (n-1)e^{-m_{j^{\star}h^{\star}}(x', 0)} e^{-G_T}, \tag{12}$$

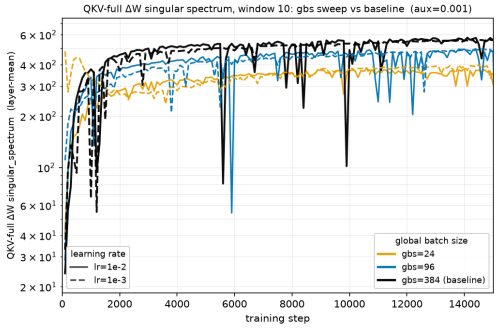
the residual routing mass shrinks exponentially in the accumulated margin gain. This conditional mechanism is supported by Figure 4(a): at fixed GBS, larger learning rates consistently reduce the layer-average router per-token entropy, and the reduction is strongest when the auxiliary load-balancing loss is removed. Figure 4(b) shows a similar pattern for the GBS sweep. Even at a fixed learning rate, decreasing GBS lowers router entropy, with the ordering GBS $= 384 > 96 > 24$



(a) Router entropy



(b) Singular spectrum of ΔW under different LRs



(c) Singular spectrum of ΔW under different GBSs

Figure 5: Visualization of the fault signatures of the two modules. (a) shows the router per-token entropy under low-precision FA, while (b) and (c) show the singular spectrum of ΔW under different learning-rate and GBS settings. The router indicator is insensitive to the low-precision attention fault (a), whereas the ΔW spectrum is insensitive to LR/GBS variation (b, c); the two signatures are therefore separable.

visible for both learning-rate settings. To make the LR–GBS coupling explicit, for global batch size B write the empirical reinforcement coefficient as

$$\widehat{\gamma}_{h,B}(t) = \gamma_h(t) + \xi_{h,B}(t), \qquad \mathbb{E}[\xi_{h,B}(t)] = 0, \qquad \mathrm{Var}[\xi_{h,B}(t)] \approx \frac{\sigma_h^2(t)}{B}.$$

Then the stochastic component of the one-step margin update obeys

$$\mathrm{Var}[\Delta m_{j^\star h}(t)] \approx \eta^2 \,\mathrm{Var}[\xi_{h,B}(t)] \approx \frac{\eta^2 \sigma_h^2(t)}{B}.$$

Thus a smaller GBS increases the per-step margin variance at fixed LR. The local expansion proved in Appendix D shows why entropy drops: near uniform routing, if the centered-logit perturbation induced by mini-batch noise is ϵ with $\mathbb{E}[\epsilon] = 0$, then

$$\mathbb{E}_\epsilon H(\mathrm{softmax}(\delta + \epsilon)) = H(\mathrm{softmax}(\delta)) - \frac{1}{2n}\mathbb{E}\|C_n \epsilon\|_2^2 + O(\|\delta, \epsilon\|^3).$$

Smaller GBS therefore lowers average router entropy at fixed LR. If the trajectory additionally enters the positive-$\gamma_h(t)$ stable-winner regime above, these noise-induced margin excursions may be reinforced over subsequent steps; the current GBS sweep supports this as a conditional mechanism rather than a standalone proof of collapse.

5 Designing Module-Specific Monitors from First Principles

The two monitor families developed in Sections 3 and 4 target different modules with different failure mechanisms. We advocate that the design principle is to understand the module's failure mecha-

nism and derive the corresponding monitor, rather than to seek a universal monitoring architecture. The two case studies below illustrate this principle.

Operator-level faults (Flash Attention). Under the biased low-precision injection described in Section 3, the attention-side indicators exhibit observable spectrum collapse in a consistent order: Δ_2 spectra show observable spectrum collapse at $\sim$5,000 steps, ΔW entropy collapses at $\sim$13,000 steps, and loss spikes only at $\sim$22,000 steps – a lead time of thousands of steps for the earliest indicator. Throughout, the router indicator remains in its healthy ranges until the loss diverges as shown in Figure 5a. The fault selectively damages the attention update path without disturbing routing.

Hyperparameter sensitivity (MoE router). The stable-winner feedback loop derived in Section 4 predicts that larger learning rates and smaller global batch sizes amplify router entropy collapse. Figure 4 confirms this: at fixed GBS, larger learning rates consistently reduce layer-average per-token entropy, and the effect is strongest when the auxiliary load-balancing loss is removed. In our observed runs the singular spectrum of ΔW remains in its healthy ranges under hyperparameter-driven routing changes, as shown in Figure 5b and 5c. These signatures are consistent with the two indicator families responding to disjoint failure mechanisms.

From case studies to a design principle. These two cases are illustrative, not exhaustive: large-model training admits many more failure modes – data distribution shift, optimizer state corruption, depth-scaling instabilities, communication faults – each with its own mechanism. The transferable lesson is not a fixed monitoring architecture but a *design principle*: each fault class has a physical or algorithmic mechanism, and the mechanism determines which internal observable will fire first. A practitioner who understands a module's failure mechanism can derive the corresponding monitor. This argues for systematic investment in *internal* training metrics grounded in module-level mechanisms – interpretability and observability of training dynamics – rather than reliance on loss curves and gradient norms alone.

6 LIMITATIONS

We discuss four open directions. **Attention variants.** The bilinear decomposition in Proposition 2 assumes the explicit W_q, W_k parameterization of multi-head attention. For MLA, GQA, MQA, and DSA, the effective QK operator is mediated by compression projections, shared heads, or dynamic routing, so the Δ_2 proxy must be re-derived for each variant. Low-rank update drift is a systematic consequence of biased backward rounding in low-precision FA, not an artifact of MHA; the spectral monitoring principle transfers, but the concrete algebra and detection thresholds remain variant-specific. **Precision and fault coverage.** The validation suite covers one fault class per category (BF16 bit-shift for operator-level faults, uniform learning-rate scaling for hyperparameter-level faults). Broader coverage of FP8 training, stochastic rounding, and gradient-clipping interactions is future work. The forward-error closure to $\kappa(W_k^\top W_q)$ also has a dimensionality mismatch (operator-space $D \times D$ vs. head-space $d_k \times d_k$) that we have not yet resolved. **Two-stage timing.** The $\approx$ 8,000-step gap between Δ_2 and ΔW entropy collapse is empirically robust but lacks a closed-form prediction. Weyl amplification accounts for the order of the gap but not its precise magnitude, which likely requires anisotropic noise statistics. A quantitative detection-onset analysis via spiked random-matrix theory is ongoing. **Router indicators.** The algebraic reduction of Equation 9 to a purely weight-only quantity $\|W_{R,c}\|_F^2/(2n)$ requires activation isotropy $M_x \propto I$. RMSNorm enforces only $\mathrm{tr}(M_x) = d$, and trained Transformer activations are anisotropic. The resulting divergence between weight-side and decision-side router indicators in concentrated-activation regimes is itself diagnostic, but is not currently used by the monitoring stack.

7 CONCLUSION

We derived internal training monitors for two stability-critical modules of modern LLMs by asking what each module is supposed to compute and where damage from its known failure mechanism would first appear.

For FA, the answer is the spectral geometry of ΔW. Biased low-precision backward errors produce coherent low-rank drift in accumulated weight updates (Observation 1), and the QK-product

decomposition (Proposition 2) exposes this drift through the first-order term Δ_2, computable from head-dimensional cores without materializing full $d \times d$ products. In our controlled fault injection, Δ_2 spectral collapse preceded loss divergence by approximately $17{,}000$ steps, and ΔW singular-spectrum collapse preceded it by approximately $9{,}000$ steps. Router indicators did not respond to this fault.

For MoE routers, the answer is per-token entropy and weight similarity. The conditioning-ratio bound (Proposition 3) and the local weight-entropy link (Equation 9) connect weight-side redundancy to decision-side entropy drop. Learning-rate and batch-size sweeps confirmed the predicted sensitivity: larger learning rates and smaller batch sizes amplify entropy collapse, while ΔW spectral indicators remain unchanged. The two fault signatures do not cross-contaminate.

The design principle itself is the transferable part of this work. Modern LLM architectures continue to introduce modules with their own internal dynamics: persistent memory stores such as Engram (Cheng et al., 2026), manifold-constrained residual connections (Xie et al., 2025), attention-based residual gates (Chen et al., 2026), and learnable structured-sparsity mechanisms (Fang et al., 2024) each carry failure modes that loss curves and gradient norms cannot attribute to a source. As these modules enter production training, each will need monitors derived from its own mechanism, not borrowed from attention or routing. The methodology demonstrated here provides a template for that derivation.

ACKNOWLEDGMENTS

We thank Xuemin Hong, Jun Li, and Honghui Ge for helpful discussions, constructive feedback on our work, and broader support that helped make this project possible.

REFERENCES

Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97 (18):10101–10106, 2000.

Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=aRZjRj41WQ.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 864–873. PMLR, 2020. URL http://proceedings.mlr.press/v119/bhojanapalli20a.html.

Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, and Ming Wei. Attention residuals. *ArXiv preprint*, abs/2603.15031, 2026. URL https://arxiv.org/abs/2603.15031.

Xin Cheng, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Zhewen Hao, Yukun Li, Han Zhang, Huishuai Zhang, Dongyan Zhao, and Wenfeng Liang. Conditional memory via scalable lookup: A new axis of sparsity for large language models. *ArXiv preprint*, abs/2601.07372, 2026. URL https://arxiv.org/abs/2601.07372.

Lénaïc Chizat, Edouard Oyallon, and Francis R. Bach. On lazy training in differentiable programming. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2933–2943, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html.

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, 2022.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

DeepSeek-AI. DeepSeek-V3 technical report, 2024.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR, 2023. URL https://proceedings.mlr.press/v202/dehghani23a.html.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 2021. URL http://proceedings.mlr.press/v139/dong21a.html.

Gongfan Fang, Hongxu Yin, Saurav Muralidharan, Greg Heinrich, Jeff Pool, Jan Kautz, Pavlo Molchanov, and Xinchao Wang. Maskllm: Learnable semi-structured sparsity for large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0e9a05f5ce62284c91e4a33498899124-Abstract-Conference.html.

Wenyi Fang, Hao Zhang, Ziyu Gong, Longbin Zeng, Xuhui Lu, Biao Liu, Xiaoyu Wu, Yang Zheng, Zheng Hu, and Xun Zhang. A survey of metrics to enhance training dependability in large language models. In *2023 IEEE 34th International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pp. 180–185. IEEE, 2023.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2021.

GLM-4.5 Team. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, 2025.

Alicia Golden, Samuel Hsia, Fei Sun, Bilge Acun, Basil Hosmer, Yejin Lee, Zachary DeVito, Jeff Johnson, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Is flash attention stable?, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ilse C.F. Ipsen and Arvind K. Saibaba. Stable rank and intrinsic dimension of real and complex matrices, 2024.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Dayal Singh Kalra, Jean-Christophe Gagnon-Audet, Andrey Gromov, Ishita Mediratta, Kelvin Niu, Alexander H. Miller, and Michael Shvartsman. A scalable measure of loss landscape curvature for analyzing the training dynamics of llms. *ArXiv preprint*, abs/2601.16979, 2026. URL https://arxiv.org/abs/2601.16979.

Kimi Team. Kimi k2: Open agentic intelligence, 2025.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8570–8581, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0d1a9651497a38d8b1c3871c84528bd4-Abstract.html.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, et al. Muon is scalable for LLM training, 2025.

Meituan LongCat Team. LongCat-Flash technical report, 2025.

Xu Pan, Aaron Philip, Ziqian Xie, and Odelia Schwartz. Dissecting query-key interaction in vision transformers. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6216515a5e0b3257c49dcb1647e497d1-Abstract-Conference.html.

Haiquan Qiu and Quanming Yao. Why low-precision transformer training fails: An analysis on flash attention, 2026. ICLR 2026.

Qwen Team. Qwen3.5-Omni technical report, 2026.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530b, a large-scale generative language model, 2022.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike no more: Stabilizing the pre-training of large language models, 2025. COLM 2025.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts, 2024.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie E. Everett, Alexander A. Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=d8w0pmvXbZ.

Haoze Wu, Zihan Qiu, Zili Wang, Hang Zhao, and Jie Fu. GW-MoE: Resolving uncertainty in MoE router with global workspace theory, 2024.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. mHC: Manifold-constrained hyper-connections. *ArXiv preprint*, abs/2512.24880, 2025. URL https://arxiv.org/abs/2512.24880.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report, 2025.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 581–590, 2020. doi: 10.1109/BigData50022.2020.9378171.

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R. Walter. Approaching deep learning through the spectral dynamics of weights, 2024.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=-Aw0rrrPUF.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhai23a.html.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient LLM training by gradient low-rank projection. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hYHsrKDiX7.

APPENDIX

A INITIALIZATION DOMINANCE AND RAW-WEIGHT MONITORING

We justify the initialization-dominance claim used in Section 3. Let $W_t = W_0 + E_t$, where E_t is the total displacement from initialization. Weyl's singular-value perturbation inequality gives, for every singular index i,

$$|\sigma_i(W_t) - \sigma_i(W_0)| \le \|E_t\|_2 \le \|E_t\|_F. \tag{13}$$

Thus, when $\|E_t\|_F \le \varepsilon \|W_0\|_F$ with $\varepsilon < 1$, every singular-value statistic of the raw weight is observed through an initialization-dominated background. This does not say that the network is not learning: in lazy or NTK-like regimes, function values can change while parameter displacement remains small. It only says that a raw-weight monitor has poor signal-to-noise for faults that first alter the update geometry. The update increment $\Delta W_{t,\delta} = W_t - W_{t-\delta}$ removes W_0 exactly, so spectral concentration in the update is not diluted by initialization energy.

B ACCUMULATION CONSEQUENCE FOR FLASH ATTENTION

This appendix proves the concentration form of the accumulation consequence (Observation 1) used for the monitoring window. The source model is their per-step mechanism: low-precision FA supplies biased scalar coefficients e_j multiplying coherent rank-one update atoms $R_j = X_j^\top (PK)_j$. This appendix does not introduce a new FA failure mechanism; it only records the accumulation consequence used by the ΔW monitor.

Index token-step samples by j and set $R_j = X_j^\top (PK)_j \in \mathbb{R}^{d \times d_k}$. Assume $Y_j = e_j R_j - M$ are independent or martingale-difference fluctuations with $M = \mathbb{E}[e_j R_j]$ and $\mathbb{E}\|Y_j\|_F^2 \le \nu^2$. Then

$$A_n := \sum_{j=1}^{n} e_j R_j = nM + Z_n, \qquad Z_n := \sum_{j=1}^{n} Y_j. \tag{14}$$

By orthogonality of the centered increments,

$$\mathbb{E}\|Z_n\|_F^2 \le n\nu^2. \tag{15}$$

Markov's inequality implies $\|Z_n\|_F = O_p(\sqrt{n})$, hence $\|Z_n\|_2 = O_p(\sqrt{n})$. This proves Equation (1). The biased mean nM grows linearly, while the zero-mean residual grows sublinearly.

Finally, suppose M has rank r and singular gap $\sigma_r(M) > 0$. Weyl's inequality gives

$$|\sigma_i(A_n) - n\sigma_i(M)| \le \|Z_n\|_2. \tag{16}$$

When $n\sigma_r(M) \gg \|Z_n\|_2$, the top r singular values of A_n are controlled by M and the remaining singular values are residual-scale. Consequently, stable rank and singular-spectrum entropy of the accumulated update converge toward those of the low-rank mean M. This is the formal sense in which biased low-precision arithmetic becomes a low-rank ΔW fault only under coherence of the rank-one atoms.

For the biased rounding-error injection used in our experiments, each selected BF16 entry is reinterpreted as its unsigned 16-bit storage word u. The implementation computes

$$\tilde{u} = (u \gg n) \ll n, \tag{17}$$

and then reinterprets $\tilde{u}$ as a BF16 value before the backward expression consumes it. This low-bit masking operation removes the lowest n storage bits, thereby discarding low-order significand information and inducing a deterministic biased rounding error rather than additive real-valued noise. For example, with $n = 3$,

$$\begin{aligned} \text{original value } 1.1015625 \quad &: \quad \texttt{0|01111111|0001101}_{\text{BF16}} = \texttt{0x3f8d} \\ &\quad\ \downarrow \texttt{(uint16 >> 3) << 3} \\ \text{attacked value } 1.0625 \quad &: \quad \texttt{0|01111111|0001000}_{\text{BF16}} = \texttt{0x3f88}. \end{aligned} \tag{18}$$

The exact tensor and mask width are experimental knobs, but the resulting tensor-level error has the same algebraic role. If the backward path uses attacked tensors $\widehat{O} = O + E_O$ and $\widehat{dO} = dO + E_{dO}$, the scalar source can include perturbations to both O and dO:

$$\widehat{\delta} - \delta = \mathrm{rowsum}(dO \odot E_O + E_{dO} \odot O + E_{dO} \odot E_O). \tag{19}$$

This implementation-specific expansion only changes what contributes to the scalar e_j. Perturbing dO also induces a direct perturbation of $dP = dOV^\top$, so the main text uses only the abstract source form $e_j R_j$ rather than treating the O, dO expansion as a separate mechanism.

C Router Similarity Bound

Proof of the conditioning-ratio similarity bound. Let

$$\varepsilon = \frac{\max_i \|w_i - \bar{w}\|}{\|\bar{w}\|}, \qquad u_i = \frac{w_i}{\|w_i\|}, \qquad e = \frac{\bar{w}}{\|\bar{w}\|}.$$

The statement is meaningful when $\bar{w} \neq 0$ and all router columns are nonzero, which we assume below. Writing the pairwise similarity as the average over ordered distinct pairs,

$$\mathrm{sim}(W_R) = \frac{1}{n(n-1)} \sum_{i \neq j} u_i^\top u_j = \frac{\|\sum_i u_i\|^2 - n}{n(n-1)}. \tag{20}$$

If $\varepsilon \geq 1$, then Equation (20) gives $\mathrm{sim}(W_R) \geq -1/(n-1)$, and since $1 - \frac{n}{n-1}\varepsilon^2 \leq 1 - \frac{n}{n-1} = -1/(n-1)$ (using $\varepsilon^2 \geq 1$), the desired bound follows. It remains to consider $0 \leq \varepsilon < 1$. Let θ_i be the angle between w_i and $\bar{w}$. Because $\|w_i - \bar{w}\| \leq \varepsilon \|\bar{w}\|$, the point w_i lies in the ball of radius $\varepsilon\|\bar{w}\|$ around $\bar{w}$. This ball does not contain the origin, so $\theta_i < \pi/2$. For an acute ray making angle θ_i with $\bar{w}$, the closest point on that ray to $\bar{w}$ has distance $\|\bar{w}\| \sin\theta_i$; hence $\sin\theta_i \leq \varepsilon$ and

$$e^\top u_i = \cos\theta_i \geq \sqrt{1 - \varepsilon^2}.$$

Therefore

$$\left\| \sum_i u_i \right\| \geq e^\top \sum_i u_i \geq n\sqrt{1 - \varepsilon^2}.$$

Substituting this into Equation (20) yields

$$\mathrm{sim}(W_R) \geq \frac{n^2(1 - \varepsilon^2) - n}{n(n-1)} = 1 - \frac{n}{n-1}\varepsilon^2,$$

which proves the proposition. □

D The Router Weight–Entropy Link

We derive Equation (9). Let $z = W_R^\top x \in \mathbb{R}^n$ be the expert logits and $\delta = (I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)z = W_{R,c}^\top x$ the centered logits ($\mathbf{1}^\top \delta = 0$); softmax is invariant under the centering shift, so $p = \mathrm{softmax}(\delta)$.

Lemma (local expansion). For centered δ,

$$\mathcal{H}\big(\mathrm{softmax}(\delta)\big) = \log n - \frac{\|\delta\|_2^2}{2n} + O(\|\delta\|^3). \tag{21}$$

Proof. With $Z = \sum_j e^{\delta_j}$ and $\sum_j \delta_j = 0$, expanding to second order gives $Z = n + \frac{1}{2}\|\delta\|_2^2 + O(\|\delta\|^3)$, hence $\log Z = \log n + \|\delta\|_2^2/(2n) + O(\|\delta\|^3)$. Writing $\mathcal{H} = \log Z - \sum_i p_i \delta_i$ and expanding $p_i = e^{\delta_i}/Z$ to the same order gives $\sum_i p_i \delta_i = \|\delta\|_2^2/n + O(\|\delta\|^3)$. Subtracting yields Equation (21). Equivalently, the Hessian of $-\mathcal{H}$ at the uniform point is $\frac{1}{n}(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$, i.e., I/n restricted to the centered subspace. □

Corollary (weight–entropy link). Substituting $\delta = W_{R,c}^\top x$ and taking expectations over the token distribution,

$$\log n - \mathbb{E}_x\, \mathcal{H}(p) = \frac{1}{2n}\, \mathrm{tr}\big(W_{R,c}^\top M_x\, W_{R,c}\big) + O\big(\mathbb{E}\|\delta\|^3\big), \tag{22}$$

with $M_x = \mathbb{E}[xx^\top]$, since $\mathbb{E}\|W_{R,c}^\top x\|_2^2 = \mathrm{tr}(W_{R,c}^\top M_x W_{R,c})$. Under second-moment isotropy $M_x \approx I_d$ this reduces to the weight-only quantity $\|W_{R,c}\|_F^2/(2n)$.

Corollary (mean-zero logit perturbations). Let $F(z) = \mathcal{H}(\mathrm{softmax}(z))$ and $C_n = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. For centered logits δ near zero and a mean-zero perturbation ϵ with $\mathbb{E}_\epsilon[\epsilon] = 0$,

$$\mathbb{E}_\epsilon F(\delta + \epsilon) = F(\delta) - \frac{1}{2n}\mathbb{E}_\epsilon\|C_n\epsilon\|_2^2 + O(\|\delta, \epsilon\|^3), \tag{23}$$

where the remainder is local and third order in $\|\delta\| + \|\epsilon\|$.

Proof. Softmax shift-invariance gives $F(z) = F(C_n z)$. Applying Equation (21) to $\delta + \epsilon$ and to δ gives

$$F(\delta + \epsilon) - F(\delta) = -\frac{1}{2n}\left(\|C_n(\delta + \epsilon)\|_2^2 - \|C_n\delta\|_2^2\right) + O(\|\delta, \epsilon\|^3).$$

Expanding the quadratic term,

$$\|C_n(\delta + \epsilon)\|_2^2 - \|C_n\delta\|_2^2 = 2(C_n\delta)^\top C_n\epsilon + \|C_n\epsilon\|_2^2.$$

Taking expectation over ϵ, the cross term vanishes because $\mathbb{E}_\epsilon[C_n\epsilon] = C_n\mathbb{E}_\epsilon[\epsilon] = 0$. This yields Equation (23). □

Scope. Three caveats bound the use of this link. First, RMSNorm fixes only the trace $\mathrm{tr}(M_x) = d$, not isotropy, so the weight-only reduction is an extra assumption – this is the fourth limitation in the main text. Second, the expansion is local: it is quantitative near uniform routing (all logits $O(1)$) and degrades to a qualitative, direction-wise monotone statement in the collapsed regime $\mathcal{H} \to 0$, where the quadratic form underestimates the true entropy drop. Third, Equation (23) is a second-order perturbative explanation of why zero-mean logit noise can lower expected entropy; it is not, by itself, a global proof of router collapse. The link is therefore a consistency check between the weight-side and decision-side indicator families, not a substitute for either.