THE DOUBLE-EDGED ROLE OF MEAN BIAS IN FP4 LLM PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

FP4 training promises substantial memory and compute savings for large language models, but remains fragile because blockwise quantization is dictated by extreme activation magnitudes, which inflate dynamic range and compress long-tail signals. We identify a counterintuitive source of this failure: dominant activation outliers are not merely arbitrary sparse events, but are largely induced by a coherent rank-one mean bias, whose direction aligns with the leading anisotropic spectral component. This mean component strengthens during training, is amplified and reshaped by attention and FFN operators, and increasingly dominates top activation magnitudes. Crucially, this discovery reveals that a seemingly complex outlier-suppression problem admits a truly simple solution: isolate the coherent mean before quantization. We therefore propose Averis, a mean–residual splitting quantization method that separates the mean component using only reductions and elementwise subtractions before FP4 quantization. Across Qwen3-0.6B Dense and Qwen3-7B-A1.5B MoE, each trained on 100B tokens, Averis enables robust W4A4G4 FP4 training, reducing BF16 loss gaps to 1.19%/0.479% versus 2.05%/0.737% for NVIDIA's recently released Hadamard-based outlier-smoothing method, while limiting downstream gaps to 0.89/0.68 points. With only 2.20% end-to-end overhead over vanilla NVFP4, about 30% of NVIDIA's Hadamard-based design, Averis provides a hardware-efficient path to stable low-bit LLM training. Code is available at: `https://anonymous.4open.science/r/averis-504D`.

1 INTRODUCTION

Training large language models (LLMs) with low-bit quantization of parameters, activations, and gradients offers substantial gains in memory footprint, and computational efficiency. In the FP4 regime, NVIDIA's NVFP4 format reduces memory consumption by $1.8\times$ compared with FP8. For General Matrix Multiplications (GeMMs), the reported $7\times$ acceleration compares NVFP4 on Blackwell with FP8 on Hopper (Alvarez et al., 2025; Devleker & Ghodsian, 2025). In our end-to-end Qwen3-8B (Yang et al., 2025) training profile on B300 GPUs, FP8 and NVFP4 achieve $1.38\times$ and $1.61\times$ per-step speedups, respectively, over the BF16 profile. However, pushing the training frontier down to FP4 often comes with training-loss gaps and downstream performance degradation (Wang et al., 2025; Chmiel et al., 2025; Tseng et al., 2025; Cao et al., 2025).

The central challenge in FP4 training is the presence of extreme outlier magnitudes in training tensors, especially activation matrices that participate in both forward and backward GeMMs (Ashkboos et al., 2024; 2025). Under blockwise quantization, a few activation outliers can set the block scale, inflate the dynamic range, and compress the long-tail signal of ordinary entries into narrow numerical bins, degrading both forward representation fidelity and backward gradient propagation (Xiao et al., 2023; Ashkboos et al., 2024; Li et al., 2025; Cao et al., 2025). Existing mitigation strategies mainly fall into two categories. One operates in the element space, typically using tiled Hadamard transforms to smooth outliers within each tile (Ashkboos et al., 2024; 2025), but with limited ability to contract the overall dynamic range and eliminate dominant outliers. The other operates in the spectral space, relying on explicit spectral control such as singular value decomposition (SVD), as in Metis (Cao et al., 2025), which typically achieves lower quantization error but remains computationally intensive and poorly aligned with modern accelerator hardware.

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.

We identify a counterintuitive structural source of FP4 training instability: dominant activation outliers are not merely arbitrary sparse events, but are largely induced by a coherent rank-one mean bias. This bias corresponds to the feature-wise mean vector of the activation matrix, forms a shared offset across tokens, and aligns with the leading spectral component. Empirically, the mean component strengthens as training progresses while maintaining this spectral alignment. Operator-level analysis further shows that attention and FFN modules actively amplify and reshape the coherent mean component.

We further show that this rank-one mean bias is the dominant source of extreme activation magnitudes. Empirically, in deep layers the top-0.1% activation entries shift from residual-dominated at early training to mean-dominated at late training, with their median squared mean-share surging from initially negligible levels to approximately 95%. Theoretically, under an empirically validated Gaussian residual model, a columnwise mean that is large relative to the residual variance exponentially amplifies far-tail exceedance probabilities relative to zero-mean fluctuations, providing a mathematical account of how the largest activations become mean-driven as the bias strengthens.

Crucially, this discovery reveals that a seemingly complex outlier-suppression problem admits a truly simple solution: isolate the coherent mean before quantization. We therefore propose *Averis* (Averaging-Induced Residual Splitting), which explicitly factors activations and output gradients into column-mean and residual components, and quantizes them separately in both forward and backward GeMMs. By removing the dominant bias-induced dynamic range from the residual stream, Averis reduces quantization error using only mean reductions and elementwise subtractions.

Averis enables robust W4A4G4 training by quantizing all GeMM matrices to FP4. Across Qwen3-0.6B Dense and Qwen3-7B-A1.5B MoE, each trained on 100B tokens, Averis reduces the BF16 loss gaps to 1.19% and 0.479%, compared with 2.05% and 0.737% for NVIDIA's recently released Hadamard-based outlier-smoothing method. The gains also translate to downstream performance under NVFP4 forward evaluation: the accuracy gaps are 0.89 and 0.68 points for Dense and MoE, respectively. On NVIDIA Blackwell GPUs, Averis incurs only 2.20% end-to-end overhead over vanilla NVFP4, about 30% of the Hadamard-based design, and delivers up to $1.57\times$ end-to-end speedup over BF16. Moreover, combining Averis with Hadamard reduces the BF16 loss gaps to 0.94% on Dense and 0.289% on MoE, while increasing the downstream averages by 0.16 and 0.28 points over Averis alone, respectively. Overall, Averis demonstrates a practical and hardware-efficient path to stable FP4 LLM training.

2 Analysis

2.1 Mean Bias Phenomenon

Notation. Let $\mathcal{X} \in \mathbb{R}^{b\times s\times m}$ denote an activation tensor, where b is the batch size, s is the sequence length, and m is the hidden dimension. For GeMM-level analysis, we reshape $\mathcal{X}$ into a matrix $\mathbf{X} \in \mathbb{R}^{l\times m}$ with $l = bs$. We write the singular value decomposition of $\mathbf{X}$ as $\mathbf{X} = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^\top$, where $r = \min(l, m)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. We define the feature-wise mean vector as $\boldsymbol{\mu}_{\mathbf{X}} \triangleq \frac{1}{l}\mathbf{X}^\top \mathbf{1} \in \mathbb{R}^m$, where $\mathbf{1} \in \mathbb{R}^l$ denotes the all-ones vector, the corresponding rank-one mean matrix as $\mathbf{M}_{\mathbf{X}} \triangleq \mathbf{1}\boldsymbol{\mu}_{\mathbf{X}}^\top \in \mathbb{R}^{l\times m}$, and the centered activation matrix as $\widetilde{\mathbf{X}} \triangleq \mathbf{X} - \mathbf{M}_{\mathbf{X}}$. We also define the normalized all-ones direction $\mathbf{e} \triangleq \mathbf{1}/\sqrt{l} \in \mathbb{R}^l$ and the alignment coefficients $\beta_k \triangleq \langle \mathbf{u}_k, \mathbf{e}\rangle$.

Mean bias: a common directional shift in representation space. We define *mean bias* as a coherent shift of token representations toward a shared direction. Figure 1(B) illustrates this phenomenon using the activations from the deepest layer of Qwen3-0.6B dense model at late stage: token-wise cosine similarities with the mean direction are almost uniformly positive, whereas those with a non-mean direction, chosen as the second right singular vector $\mathbf{v}_2$, alternate in sign. This contrast indicates that the mean direction captures a shared representation-level offset rather than ordinary variance. The same pattern holds across additional layers and training stages, as shown in Appendix A.

Mean bias aligns with the dominant anisotropic spectral direction. We further find that this coherent mean direction is tightly coupled to the leading spectral spike. Substituting the SVD of $\mathbf{X}$ into the definition of $\boldsymbol{\mu}_{\mathbf{X}}$ gives $\boldsymbol{\mu}_{\mathbf{X}} = \frac{1}{\sqrt{l}} \sum_{k=1}^{r} \sigma_k \beta_k \mathbf{v}_k$. This expansion shows that the mean direction is controlled jointly by spectral anisotropy, through the singular values σ_k, and by token-wise coherence, through the alignment coefficients β_k. In LLM activations, the spectrum is strongly anisotropic, with a dominant leading singular value as shown in Figure 1(A). When this leading

mode also has large alignment with the all-ones token direction, while non-leading modes exhibit stronger sign cancellation, the mean vector concentrates along $\mathbf{v}_1$. Consistently, Figure 1(C) shows that $\boldsymbol{\mu}_{\mathbf{X}}$ has a cosine similarity approaching 0.99 with $\mathbf{v}_1$, much higher than with other right singular directions.

The mean direction and its alignment with the leading singular vector remain stable across the evaluated microbatch sizes, sequence lengths, and packing layouts; on padded inputs, we exclude padded positions from the mean reduction and normalize by the valid-token count (Appendix H).

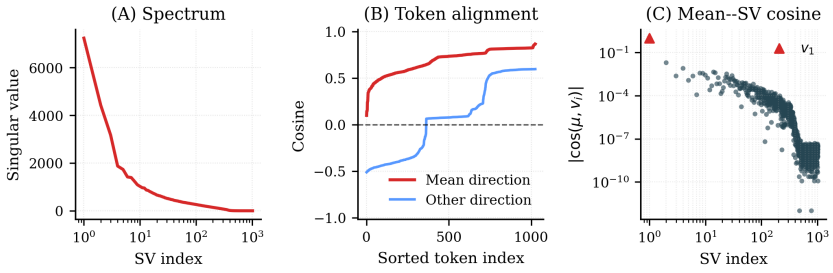


Figure 1: Mean-bias evidence in layer-27 FFN-input activations of Qwen3-0.6B at step 170k. (A) A dominant leading singular value indicates strong anisotropy. (B) Token-wise cosine similarities stay mostly positive along the mean direction but are weaker and more mixed along a non-leading direction. (C) The mean vector aligns most strongly with the top right singular vector.

2.2 Structural Emergence of Mean Bias

We examine the evolution of *mean bias* across training. Globally, the feature-wise mean strengthens over training while staying aligned with the leading anisotropic direction; locally, attention and FFN modules amplify and reshape this coherent component within Transformer blocks.

Empirical Trajectory: Mean Bias Strengthens During Training. We quantify the strength of mean bias by the normalized ratio $R \triangleq \frac{\|\boldsymbol{\mu}_{\mathbf{X}}\|_2}{\sqrt{\|\mathbf{X}\|_F^2/l}}$. Using activation matrices from Qwen3-0.6B at different training stages, we track R and the cosine alignment between the mean vector and the top right singular vector across layers. As shown in Figure 2, the mean-bias energy generally increases as training progresses, most prominently in early and middle layers. Meanwhile, the mean vector remains strongly aligned with the top singular direction across layers and stages, often approaching cosine similarity 1.0. These trends show that mean bias is progressively strengthened during training while remaining a structured rank-one component coupled to the dominant anisotropic direction.

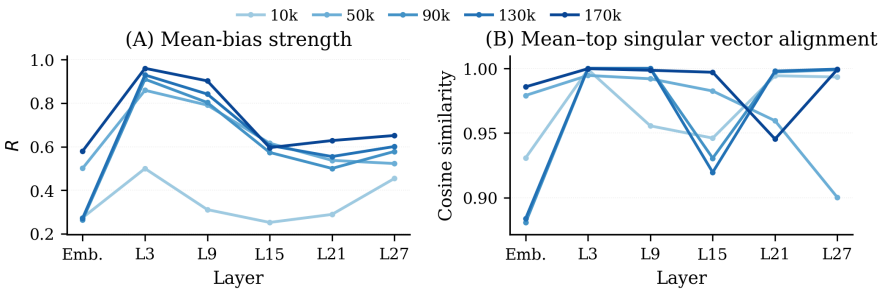


Figure 2: Evolution of mean bias across depth and training in Qwen3-0.6B activations. Left: the normalized mean-bias ratio R increases with training step. Right: the growing mean component remains tightly aligned with the dominant anisotropic direction.

Operator-Level Regeneration and Amplification. Transformer operators actively amplify mean bias, while also reshaping its direction across training stages. We verify this by tracing the mean-bias ratio R across operator stages and measuring adjacent-stage mean-vector cosine. Figure 3(A) shows

that, at the early checkpoint, both attention and FFN increase R, with the FFN also substantially redirecting the mean direction. At the late checkpoint, Figure 3(B) shows that amplification is mainly concentrated in the FFN path, while attention contributes more to directional reshaping.

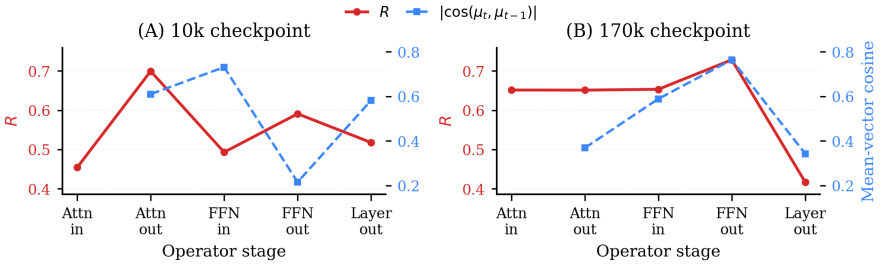


Figure 3: Operator-level evidence for mean-bias amplification. We compare the evolution of mean-bias strength and mean-direction alignment across key operators, showing that nonlinear Transformer components further strengthen the coherent mean component during forward propagation.

2.3 Mean Bias as the Dominant Source of Activation Outliers

We explain the sharp numerical consequence of mean bias: it produces extreme elementwise activations that set blockwise quantization scales, inflate dynamic range, and compress ordinary long-tail signals into narrow numerical bins.

Empirical outlier attribution. To localize the source of extreme activation values, let $\mathcal{E}_{\text{top}}$ denote the index set of the top-0.1% entries of $\mathbf{X}$ ranked by absolute magnitude. For each $(i, j) \in \mathcal{E}_{\text{top}}$, we measure the component-wise squared contribution proxies $\rho_{ij}^{(\text{mean})} \triangleq \frac{(\mathbf{M}_{\mathbf{X}})_{ij}^2}{X_{ij}^2}$, $\rho_{ij}^{(\text{res})} \triangleq \frac{\widetilde{X}_{ij}^2}{X_{ij}^2}$, which serve as empirical indicators of whether extreme entries are dominated by the coherent mean component or by the residual. Figure 4 reports this attribution for activations of Qwen3-0.6B at early and late checkpoints, comparing a shallow layer with a deep layer. At the early checkpoint, top outliers are largely residual-dominated. As training progresses, the mean contribution shifts toward larger values, where late-stage outliers exhibit a substantial concentration of high mean-share values.

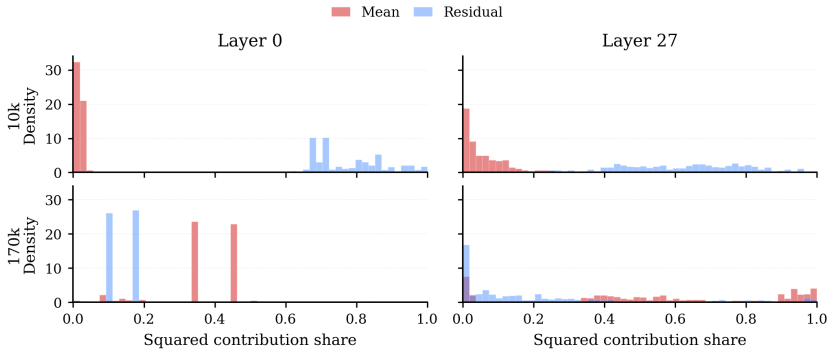


Figure 4: Histograms of activation outlier mean and residual contribution shares (top 0.1% entries) across layers and training steps.

Theoretical model: mean-bias amplification of columnwise outliers. We now formalize this effect at the level of a single feature coordinate. For a fixed column j, let

$$m_j \triangleq (\boldsymbol{\mu}_{\mathbf{X}})_j = \frac{1}{\sqrt{l}} \sum_{k=1}^{r} \sigma_k \beta_k v_{kj}. \tag{1}$$

We also denote the leading-mode contribution to this column mean by $m_j^{(1)} \triangleq \frac{\sigma_1 \beta_1}{\sqrt{l}} v_{1j}$.

Let $I \sim \mathrm{Unif}(\{1, \ldots, l\})$ and define the row-sampled entry

$$Y_j \triangleq X_{Ij} = m_j + \eta_{Ij}, \qquad \eta_{Ij} \triangleq X_{Ij} - m_j.$$

Thus η_{Ij} is the centered residual obtained by sampling one token position from column j.We use the following assumptions.

Assumption 1: Marginal Gaussian residual. For each column j, the row-sampled centered residual is approximated by a Gaussian marginal distribution,

$$\eta_{Ij} \dot{\sim} N(0, \tau_j^2). \tag{2}$$

This assumption concerns only the marginal distribution of a randomly sampled coordinate and does not require the residuals $\{\eta_{ij}\}_{i=1}^{l}$ to be mutually independent. Empirically, Figure 5 shows that the raw activation distribution deviates strongly from a Gaussian fit, whereas the residual after mean removal is substantially closer to Gaussian.

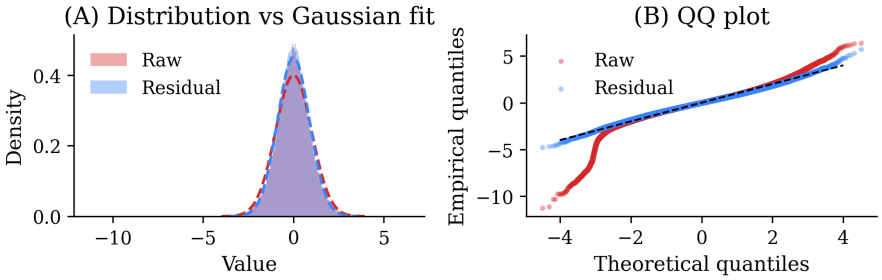


Figure 5: Empirical validation of the marginal Gaussian residual assumption. Raw activations deviate from Gaussianity, while residuals are much closer to a Gaussian fit in both density and QQ plots.

Assumption 2: Diagonal variance approximation. The residual variance of column j is well approximated by its diagonal spectral contribution:

$$\tau_j^2 \approx \tau_{j,\mathrm{diag}}^2 \triangleq \frac{1}{l} \sum_{k=1}^{r} \sigma_k^2 (1 - \beta_k^2) v_{kj}^2. \tag{3}$$

Equivalently, cross-mode covariance terms contribute little to the row-sampling variance. This approximation is empirically validated in Appendix B, which shows that the diagonal estimate closely tracks the empirical residual variance and that the relative cross-term contribution remains small.

Assumption 3: Mean-dominant columns with controlled attribution error. For columns that contribute to large outliers, (a) $|m_j|/\tau_j \gg 1$; and (b) for the thresholds t under consideration and a prescribed $\varepsilon > 0$,

$$\varepsilon_j(t) \triangleq \frac{|m_j - m_j^{(1)}|}{\tau_j} \left(\bar{a}_j(t)_+ + 1\right) \le \varepsilon, \qquad \bar{a}_j(t) \triangleq \frac{t - \min\{|m_j|, |m_j^{(1)}|\}}{\tau_j}, \tag{4}$$

where $x_+ = \max\{x, 0\}$. This condition controls the attribution error relative to the residual scale and the threshold, rather than only relative to the mean magnitude.

Theorem 1 (Mean-bias amplification of columnwise outliers)**.** *Fix a column j and let $Y_j = X_{Ij} = m_j + \eta_{Ij}$ with $I \sim \mathrm{Unif}(\{1, \ldots, l\})$. Suppose Assumption 1 holds exactly with $\eta_{Ij} \sim N(0, \tau_j^2)$ and $\tau_j > 0$. Write $s = |m_j|$, $Q = 1 - \Phi$, where Φ is the standard normal CDF, and let $Y_j^{(0)} \sim N(0, \tau_j^2)$ be the zero-mean, variance-matched baseline. For every $t > 0$,*

$$\mathbb{P}(|Y_j| > t) = Q\left(\frac{t - s}{\tau_j}\right) + Q\left(\frac{t + s}{\tau_j}\right). \tag{5}$$

This probability is strictly increasing in $s \ge 0$. Define $A(t) = \mathbb{P}(|Y_j| > t)/\mathbb{P}(|Y_j^{(0)}| > t)$ and $\Delta = s(2t - s)/(2\tau_j^2)$. If $t \ge s$, then

$$0 \le \log(2A(t)) - \Delta \le \log\left(1 + \sqrt{\frac{\pi}{2}}\left(1 + \frac{t}{\tau_j}\right)\right), \qquad A(t) \ge \frac{1}{2} \exp\left(\frac{s^2}{2\tau_j^2}\right). \tag{6}$$

If $0 < t < s$, *then* $\mathbb{P}(|Y_j| > t) > 1/2$. *For* $0 < p \le 1/2$, *let* $q_p(|Y|)$ *satisfy* $\mathbb{P}(|Y| > q_p(|Y|)) = p$. *Then*

$$s - 0.87\tau_j \le q_p(|Y_j|) - q_p(|Y_j^{(0)}|) \le s. \tag{7}$$

These conclusions require no far-tail asymptotic condition.

Proof. See Appendix E for the full bounds and proofs.

The theorem characterizes mean-bias amplification even when t is close to $|m_j|$: the exponent Δ controls $\log(2A(t))$ with an explicit additive error, while the quantile bound describes the effect on activation magnitudes relevant to the quantization dynamic range. The far-tail formula is recovered under the additional conditions $(t - |m_j|)/\tau_j \to \infty$ and $t|m_j|/\tau_j^2 \to \infty$ (Appendix E). Under Assumption 3, replacing m_j by $m_j^{(1)}$ while retaining the same residual variance changes the tail probability by a factor between $e^{-\varepsilon}$ and e^{ε} (Corollary 1). This quantifies attribution to the rank-one mean-bias component $\frac{\sigma_1\beta_1}{\sqrt{l}}\mathbf{1}\mathbf{v}_1^\top$. A small relative mean error alone controls the amplification exponent but need not give a small relative error in the tail probability.

We further verify this implication empirically in Appendix C. For late-stage activations in both shallow and deep layers, subtracting the feature-wise mean substantially contracts the high-magnitude tail of the value distribution.

3 METHOD

Building on the insight above, we identify a simple design principle for low-bit training: since quantization-sensitive outliers are predominantly driven by a coherent rank-one mean component, it is sufficient to isolate this component before quantization. We propose *Averis* (Averaging-Induced Residual Splitting), a mean–residual quantization method that explicitly separates the column-mean component from the residual and quantizes them independently. Averis applies this decomposition to activations and output gradients in the forward and backward GeMMs; although output gradients show weaker mean bias, mean centering still slightly reduces quantization error, as verified in Appendix D. It requires only mean reductions, remaining well aligned with efficient accelerator execution.

Notation and quantization operator. Let $\mathcal{Q}_b(\cdot)$ denote a b-bit quantization operator (e.g., blockwise FP4) applied to a matrix. We write $\bar{\mathbf{M}} \triangleq \mathcal{Q}_b(\mathbf{M})$ for the quantized version of any matrix $\mathbf{M}$. Both mean vectors $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{D}}$ are also quantized to FP4, independently of their residual matrices. Throughout, for any positive integer d, we write $\mathbf{1}_d \in \mathbb{R}^d$ for the all-ones column vector of length d.

Forward pass: activation mean–residual splitting. Let $\mathbf{W} \in \mathbb{R}^{m\times n}$ be a weight matrix and let $\mathcal{X} \in \mathbb{R}^{b\times s\times m}$ be an input activation tensor, where b is batch size, s is sequence length, and m is hidden dimension. As usual, we reshape $\mathcal{X}$ into a matrix $\mathbf{X} \in \mathbb{R}^{l\times m}$ with $l = b \cdot s$ and compute the GeMM output $\mathbf{Y} = \mathbf{X}\mathbf{W} \in \mathbb{R}^{l\times n}$.

We define the *activation column-mean vector* $\mu_{\mathbf{X}} \in \mathbb{R}^m$ by $\mu_{\mathbf{X}} \triangleq \frac{1}{l}\mathbf{1}_l^\top\mathbf{X}$. The residual is $\mathbf{X}_{\mathrm{R}} \triangleq \mathbf{X} - \mathbf{1}_l\,\mu_{\mathbf{X}}$. We then quantize the mean vector and residual independently:

$$\bar{\mu}_{\mathbf{X}} = \mathcal{Q}_b(\mu_{\mathbf{X}}), \qquad \bar{\mathbf{X}}_{\mathrm{R}} = \mathcal{Q}_b(\mathbf{X}_{\mathrm{R}}), \qquad \bar{\mathbf{W}} = \mathcal{Q}_b(\mathbf{W}).$$

The quantized forward computation is

$$\hat{\mathbf{Y}} \;=\; \mathbf{1}_l\,(\bar{\mu}_{\mathbf{X}}\bar{\mathbf{W}}) \;+ \bar{\mathbf{X}}_{\mathrm{R}}\,\bar{\mathbf{W}}. \tag{8}$$

where the first term is broadcast along the token dimension via $\mathbf{1}_l$. In practice, one computes $\mu_{\mathbf{X}}$ first, subtracts it from $\mathbf{X}$ on the fly to form $\mathbf{X}_{\mathrm{R}}$, and never materializes the mean matrix.

Backward pass: output-gradient mean–residual splitting. Let $\mathbf{D} \in \mathbb{R}^{l\times n}$ denote the matrix of loss derivatives with respect to the GeMM output $\mathbf{Y}$, i.e., $\mathbf{D} \triangleq \frac{\partial\mathcal{L}}{\partial\mathbf{Y}}$. We define the *output-gradient column-mean vector* $\mu_{\mathbf{D}} \in \mathbb{R}^n$ by

$$\mu_{\mathbf{D}} \triangleq \frac{1}{l}\mathbf{1}_l^\top\mathbf{D}, \qquad \mathbf{D}_{\mathrm{R}} \triangleq \mathbf{D} - \mathbf{1}_l\,\mu_{\mathbf{D}}.$$

Quantize the components independently:

$$\bar{\mu}_{\mathbf{D}} = \mathcal{Q}_b(\mu_{\mathbf{D}}), \qquad \bar{\mathbf{D}}_{\mathrm{R}} = \mathcal{Q}_b(\mathbf{D}_{\mathrm{R}}).$$

For the input-gradient GeMM, Averis evaluates

$$\widehat{\frac{\partial \mathcal{L}}{\partial \mathbf{X}}} = \mathbf{1}_l \, (\bar{\mu}_{\mathbf{D}} \bar{\mathbf{W}}^\top) + \bar{\mathbf{D}}_{\mathrm{R}} \bar{\mathbf{W}}^\top. \tag{9}$$

Single-GeMM implementation of the mean and residual products. Equations (8) and (9) do not require an independent matrix–vector multiplication for the mean contribution. For the input gradient, let $\mathbf{D}_{\mathrm{R}} \in \mathbb{R}^{l \times n}$ and $\mu_{\mathbf{D}} \in \mathbb{R}^{1 \times n}$ denote the residual and its associated row-vector mean. After quantizing them separately, concatenate along the token dimension and compute a single GeMM:

$$\mathbf{Z} = \begin{bmatrix} \bar{\mu}_{\mathbf{D}} \\ \bar{\mathbf{D}}_{\mathrm{R}} \end{bmatrix} \in \mathbb{R}^{(l+1) \times n}, \qquad \mathbf{H} = \mathbf{Z} \bar{\mathbf{W}}^\top \in \mathbb{R}^{(l+1) \times m}.$$

Broadcasting the first output row over the remaining rows gives

$$\left(\widehat{\frac{\partial \mathcal{L}}{\partial \mathbf{X}}} \right)_{i,:} = \mathbf{H}_{i+1,:} + \mathbf{H}_{1,:}, \qquad i = 1, \ldots, l,$$

which is exactly Equation (9) for the quantized operands. The forward pass uses the same construction.

For the weight-gradient GeMM, the exact mean–residual expansion gives

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{X}^\top \mathbf{D} = \mathbf{X}_{\mathrm{R}}^\top \mathbf{D}_{\mathrm{R}} + \mathbf{X}_{\mathrm{R}}^\top (\mathbf{1}_l \mu_{\mathbf{D}}) + (\mathbf{1}_l \mu_{\mathbf{X}})^\top \mathbf{D}_{\mathrm{R}} + (\mathbf{1}_l \mu_{\mathbf{X}})^\top (\mathbf{1}_l \mu_{\mathbf{D}}).$$

Since $\mathbf{X}_{\mathrm{R}}$ and $\mathbf{D}_{\mathrm{R}}$ are column-centered, $\mathbf{X}_{\mathrm{R}}^\top \mathbf{1}_l = 0, \mathbf{1}_l^\top \mathbf{D}_{\mathrm{R}} = 0$, and therefore the two cross terms vanish: $\mathbf{X}_{\mathrm{R}}^\top (\mathbf{1}_l \mu_{\mathbf{D}}) = 0, (\mathbf{1}_l \mu_{\mathbf{X}})^\top \mathbf{D}_{\mathrm{R}} = 0$. This cancellation is exact before quantization; Averis approximates the resulting two-term identity without requiring the quantized residuals to remain exactly zero-mean (see Appendix I for a two-term versus four-term comparison). Thus Averis evaluates the quantized weight-gradient GeMM as

$$\widehat{\frac{\partial \mathcal{L}}{\partial \mathbf{W}}} = \bar{\mathbf{X}}_{\mathrm{R}}^\top \bar{\mathbf{D}}_{\mathrm{R}} + l \, \bar{\mu}_{\mathbf{X}}^\top \bar{\mu}_{\mathbf{D}}. \tag{10}$$

In MoE layers, mean extraction is performed separately for each expert after token dispatch; the expert-wise formulation is given in Appendix G.

4 Experiments

Models and Datasets. We conduct experiments on two model settings: Qwen3-0.6B Dense and Qwen3-7B-A1.5B MoE. The MoE model is a scaled-down variant following Qwen3-235B-A22B. For both models, we use the DCLM (Li et al., 2024) dataset for pretraining. Both are trained on 100B tokens. Full model and training configurations are provided in Appendix F.

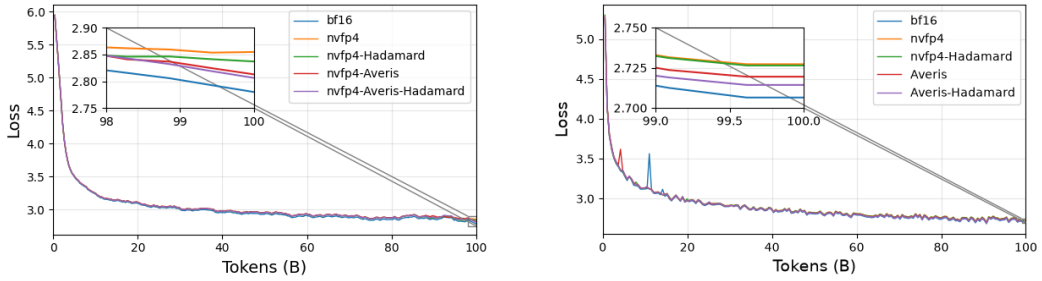
FP4 Training. All low-bit experiments adopt W4A4G4 quantization, where weights, activations, and gradients are represented in the E2M1 NVFP4 format. Stochastic rounding (SR) is applied to the backward gradient GeMMs to mitigate quantization bias, and is orthogonal to other methods. Training-quality experiments for both model scales are conducted with FP4 simulation on Hopper GPUs, while the end-to-end runtime overhead is evaluated on Blackwell GPUs.

Baselines and Evaluation. We compare Averis against full-precision BF16 training, vanilla FP4 training, and NVIDIA's recently released Hadamard-based outlier-smoothing design for FP4 training. We evaluate different methods by the training-loss gap relative to BF16 and downstream task performance. For low-precision settings, downstream evaluation is also performed with NVFP4 quantized forward computation. For downstream evaluation, we consider three task types: question answering (ARC Clark et al. (2018), RACE Lai et al. (2017)), classification (HellaSwag Zellers et al. (2019), PIQA Bisk et al. (2020)), and cloze prediction (LAMBADA Paperno et al. (2016)).

4.1 Main Results

Figure 6 compares the training-loss trajectories of Qwen3-0.6B and Qwen3-7B-A1.5B, and Table 1 reports the final loss and downstream results. On Qwen3-0.6B, Averis substantially narrows the BF16

loss gap, reducing it from 2.70% under vanilla NVFP4 and 2.05% under NVFP4-Hadamard to 1.19%. On Qwen3-7B-A1.5B, the corresponding FP4 loss gaps are 0.766%, 0.737%, 0.479%, and 0.289%. Averis improves the matched FP4 baseline at both model scales, and its combination with Hadamard gives the smallest MoE loss gap.



(a) Training-loss curve for Qwen3-0.6B.

(b) Training-loss curve for Qwen3-7B-A1.5B.

Figure 6: Training-loss curves across model scales. Left: Qwen3-0.6B. Right: Qwen3-7B-A1.5B.

Downstream Performance. Table 1 shows that under NVFP4 forward evaluation, Averis limits the average downstream gaps relative to BF16 to 0.89 and 0.68 points on Dense and MoE, respectively. Averis-Hadamard further narrows these gaps to 0.73 and 0.40 points, achieving the best downstream average among FP4 methods at both scales.

Complementarity with Hadamard Transformation. Averis is orthogonal to Hadamard-based outlier smoothing. On Qwen3-0.6B, combining Averis with Hadamard transformation further reduces the loss gap from 1.19% to 0.94% and gives the best downstream average among FP4 variants. On Qwen3-7B-A1.5B, combining the two methods reduces the loss gap from 0.479% for Averis and 0.737% for NVFP4-Hadamard to 0.289%, extending the evidence of complementarity to MoE. This suggests that Averis removes the dominant coherent mean-induced dynamic range, while Hadamard transformation can still provide additional element-space smoothing on the residual signal.

Table 1: Training loss and downstream performance comparison on Qwen3-0.6B and Qwen3-7B-A1.5B. Loss gaps are computed from the original, unrounded losses. Downstream results are reported in percentage points. Bold numbers indicate the best result among FP4 methods.

Method	Loss	Loss Gap	ArcC	ArcE	HellaSwag	LAMBADA	PIQA	RACE	Avg
Qwen3-0.6B									
BF16	2.780	–	28.92	53.37	49.83	47.76	71.00	55.31	51.03
NVFP4	2.855	2.70%	27.73	52.48	48.15	44.93	69.80	**55.55**	49.77
NVFP4-Hadamard	2.837	2.05%	27.56	52.99	**48.51**	45.78	**70.02**	55.45	50.05
Averis	2.813	1.19%	**29.44**	**53.66**	48.15	44.44	69.64	55.49	50.14
Averis-Hadamard	**2.806**	**0.94%**	28.92	52.86	48.43	**46.96**	**70.02**	54.60	**50.30**
Qwen3-7B-A1.5B									
BF16	2.706	–	33.96	61.24	58.95	43.24	74.16	32.63	50.70
NVFP4	2.727	0.766%	30.55	60.31	56.88	42.65	73.61	31.29	49.22
NVFP4-Hadamard	2.726	0.737%	31.31	**61.03**	58.09	42.93	**73.88**	31.48	49.79
Averis	2.719	0.479%	32.76	60.19	57.39	**43.33**	73.61	**32.82**	50.02
Averis-Hadamard	**2.714**	**0.289%**	**33.79**	60.77	**58.14**	42.75	73.50	**32.82**	**50.30**

Runtime overhead comparison. We benchmark the preprocessing overhead of Averis against tiled Hadamard transformation. Tiled Hadamard reshapes $\mathbf{X}$ into $[l, m/16, 16]$ and applies a 16×16 Hadamard transform along the last dimension, whereas Averis only computes the feature-wise mean of $\mathbf{X}$. As shown in Table 2, Averis is consistently faster on large activation shapes, achieving mean-latency speedups of $4.47\times$ and $4.72\times$ for $(l, m) = (512 \times 2048, 4096)$ and $(512 \times 2048, 8192)$, respectively. This advantage increases with activation size.

Table 2: Runtime overhead comparison between tiled Hadamard preprocessing and Averis mean extraction on large activation shapes. Latencies are reported in milliseconds. Speedup is computed from mean latency as $T_{\text{Hadamard}}/T_{\text{Averis}}$.

Shape (l, m)	Method	Mean	Std	Speedup
$(512 \times 2048, 4096)$	Tiled Hadamard	9.1614	0.0128	–
$(512 \times 2048, 4096)$	Averis	2.0494	0.0021	$4.47\times$
$(512 \times 2048, 8192)$	Tiled Hadamard	18.8421	0.0103	–
$(512 \times 2048, 8192)$	Averis	3.9927	0.0028	$4.72\times$

End-to-end overhead on Blackwell. Table 3 compares end-to-end training-step latencies on Blackwell for vanilla NVFP4, Averis, and NVFP4-Hadamard. Averis adds only 2.01% and 2.20% overhead over vanilla NVFP4 on Qwen3-0.6B and Qwen3-7B-A1.5B, respectively, corresponding to 29.6% and 28.9% of the Hadamard overhead.

Table 3: End-to-end training-step latency on Blackwell GPUs for Qwen3-0.6B and Qwen3-7B-A1.5B.

Setting	Qwen3-0.6B (ms)	0.6B Overhead	Qwen3-7B-A1.5B (ms)	7B Overhead
NVFP4	5545.30	–	24852.04	–
Averis	5656.80	2.01%	25398.78	2.20%
NVFP4-Hadamard	5922.38	6.80%	26744.92	7.62%

5 Related Work

Outlier dimensions and extreme activation magnitudes have been a central challenge in low-bit quantization of LLMs. SmoothQuant Xiao et al. (2023) redistributes activation and weight scales to reduce outlier concentration. GPTQ Frantar et al. (2023), AWQ Lin et al. (2024), QLoRA Dettmers et al. (2023), and SVDQuant Li et al. (2025) propose various strategies to absorb or compensate for large-magnitude components. QuaRot Ashkboos et al. (2024) and HALO Ashkboos et al. (2025) employ orthogonal rotations or Hadamard transforms to mitigate outliers. Outlier Suppression Wei et al. (2022) explicitly targets extreme activation dimensions. Outlier Suppression+ Wei et al. (2023) further addresses channel asymmetry and concentration through equivalent channel-wise shifting and scaling for post-training quantization (PTQ). HadaNorm Federici et al. (2025) combines dynamic, per-sequence channel-mean subtraction with scaling and Hadamard transforms for diffusion-transformer PTQ, restoring the mean contribution in the linear output. These methods establish closely related approaches to centering and transforming activations before quantization. Recent work on FP8 and FP4 training further highlights the sensitivity of low-precision regimes to activation dynamic range Micikevicius et al. (2022); Kuzmin et al. (2022); Perez et al. (2023); Cao et al. (2025). Although Averis shares the idea of mean extraction with prior methods, it is motivated by our observation of a shared mean across tokens and is applicable to pretraining.

6 Conclusion

FP4 training remains fragile because activation outliers inflate blockwise quantization scales and compress long-tail signals. We identify a coherent rank-one mean bias, aligned with the leading anisotropic spectral component, as a structural source of these outliers. Our method, *Averis*, isolates this mean before FP4 quantization of activations and output gradients using only reductions and elementwise subtractions. On dense and MoE Qwen3 models, Averis narrows BF16 loss and downstream accuracy gaps, complements Hadamard transformation, and incurs only 2.2% end-to-end overhead on Blackwell GPUs.

LIMITATIONS

Long-run experiments use FP4 simulation rather than end-to-end Blackwell execution. Validation across more scales, datasets, and architectures—and specialized backward-pass treatment, since mean bias is weaker in output gradients—remain future work.

AI USE STATEMENT

In this work, we used generative AI tools to prepare initial drafts of sections of the manuscript and to revise the writing for clarity, grammar, and style. We also used these tools to assist with manuscript formatting, reference formatting, and bibliographic verification. Generative AI also assisted in proving mathematical claims, including exploring proof strategies, deriving non-asymptotic bounds, and checking intermediate derivations. We additionally used AI for literature retrieval and discovery, including identifying related work on low-precision training and outlier suppression. The authors reviewed and revised all AI-assisted text and manually verified the cited references. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, 1964.

Eduardo Alvarez, Omri Almog, Eric Chung, Simon Layton, Dusan Stosic, Ronny Krashinsky, and Kyle Aubrey. Introducing NVFP4 for efficient and accurate low-precision inference. NVIDIA Technical Blog, June 2025. URL https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/.

Saleh Ashkboos, Mahdi Nikdan, Soroush Tabesh, Roberto L. Castro, Torsten Hoefler, and Dan Alistarh. Halo: Hadamard-assisted lower-precision optimization for llms. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2501.02625.

Saleh Ashkboos et al. Quarot: Outlier-free 4-bit inference in rotated llms. In *NeurIPS*, 2024.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Hengjie Cao, Mengyi Chen, Yifeng Yang, Ruijun Huang, Fang Dong, Jixian Zhou, Anrui Chen, Mingzhi Dong, Yujiang Wang, Jinlong Hou, Yuan Cheng, Fan Wu, Fan Yang, Tun Lu, Ning Gu, and Li Shang. Metis: Training llms with fp4 quantization. arXiv preprint arXiv:2509.00404, 2025. URL https://arxiv.org/abs/2509.00404.

Brian Chmiel, Maxim Fishman, Ron Banner, and Daniel Soudry. Fp4 all the way: Fully quantized training of llms. arXiv preprint arXiv:2505.19115, 2025. URL https://arxiv.org/abs/2505.19115.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.14314.

Kirthi Devleker and Farshad Ghodsian. NVFP4 trains with precision of 16-bit and speed and efficiency of 4-bit. NVIDIA Technical Blog, August 2025. URL https://developer.nvidia.com/blog/nvfp4-trains-with-precision-of-16-bit-and-speed-and-efficiency-of-4-bit/.

Marco Federici, Riccardo Del Chiaro, Boris van Breugel, Paul Whatmough, and Markus Nagel. HadaNorm: Diffusion transformer quantization through mean-centered transformations, 2025. URL https://arxiv.org/abs/2506.09932.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.17323.

Andrey Kuzmin, Mart Van Baalen, Yuwei Ren, Markus Nagel, Jorn Peters, and Tijmen Blankevoort. Fp8 quantization: The power of the exponent. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5e07476b6bd2497e1fbd11b8f0b2de3c-Abstract-Conference.html.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. arXiv preprint arXiv:1704.04683, 2017. URL https://arxiv.org/abs/1704.04683.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.05007.

Ji Lin, Jiaming Tang, et al. Awq: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of MLSys*, 2024. URL https://arxiv.org/abs/2306.00978.

Paulius Micikevicius et al. Fp8 formats for deep learning. arXiv preprint arXiv:2209.05433, 2022.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1525–1534, 2016. URL https://aclanthology.org/P16-1144/.

Sergio P. Perez, Yan Zhang, James Briggs, Charlie Blake, Josh Levy-Kramer, Paul Balanca, Carlo Luschi, Stephen Barlow, and Andrew William Fitzgibbon. Training and inference of large language models using 8-bit floating point. arXiv preprint arXiv:2309.17224, 2023. URL https://arxiv.org/abs/2309.17224.

Albert Tseng, Tao Yu, and Youngsuk Park. Training llms with mxfp4. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. URL https://proceedings.mlr.press/v258/tseng25a.html.

Ruizhe Wang, Yeyun Gong, Xiao Liu, Guoshuai Zhao, Ziyue Yang, Baining Guo, Zhengjun Zha, and Peng Cheng. Optimizing large language model training using fp4 quantization. arXiv preprint arXiv:2501.17116, 2025. URL https://arxiv.org/abs/2501.17116.

Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. Outlier suppression: Pushing the limit of low-bit transformer language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17402–17414, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f6db140de9c9f111b12ef8a216320a9-Abstract-Conference.html.

Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. Outlier Suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pp. 1648–1665. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.102. URL https://aclanthology.org/2023.emnlp-main.102/.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. URL https://aclanthology.org/P19-1472/.

A Additional Mean-Bias Visualizations Across Layers and Training Stages

To verify that the pattern shown in Figure 1 is not an isolated late-stage layer-27 artifact, we examine the same three-panel diagnostic across additional layers and checkpoints. Figures 7, 8, and 9 report three representative cases: layer 0 at 10k, layer 0 at 170k, and layer 27 at 170k. In all three cases, the same qualitative structure persists: the singular spectrum exhibits a leading spike, token-wise projections onto the mean direction remain predominantly one-sided, and the mean vector is most strongly aligned with the top right singular vector.

These supplementary examples highlight two aspects of generality. Across training, the phenomenon is already visible in early checkpoints and becomes stronger at later stages, as seen by comparing Figures 7 and 8. Across depth, the effect appears in both shallow and deep layers, while becoming especially sharp in the deep late-stage activation shown in Figure 9. Taken together, these observations show that mean bias is not a single-layer artifact but a robust structural feature of LLM activations.

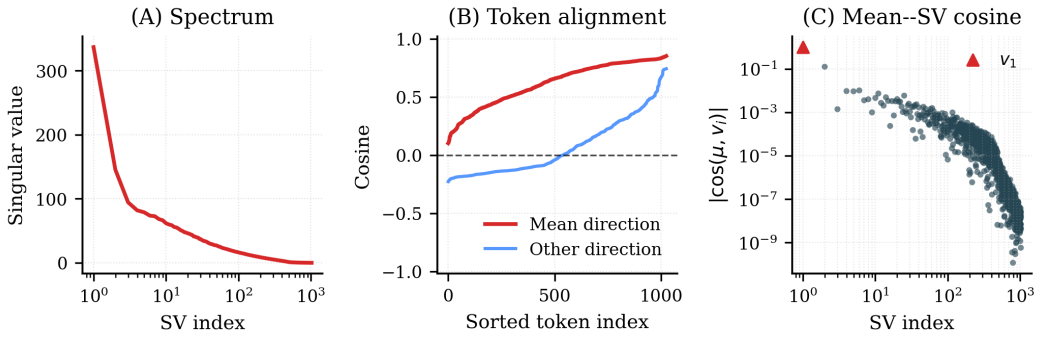


Figure 7: Additional mean-bias visualization for layer 0 at training step 10k. The leading spectral spike, predominantly one-sided token alignment to the mean direction, and strong mean–top-singular-vector alignment are already visible at an early checkpoint.

B Empirical Validation of the Diagonal Variance Approximation

Under uniform row sampling $I \sim \mathrm{Unif}(\{1, \ldots, l\})$, the residual variance admits the exact decomposition

$$\tau_j^2 = \mathrm{Var}_I(\eta_{Ij}) = \tau_{j,\mathrm{diag}}^2 + \tau_{j,\mathrm{cross}}^2, \qquad \tau_{j,\mathrm{diag}}^2 = \frac{1}{l}\sum_{k=1}^{r} \sigma_k^2(1-\beta_k^2)v_{kj}^2. \tag{11}$$

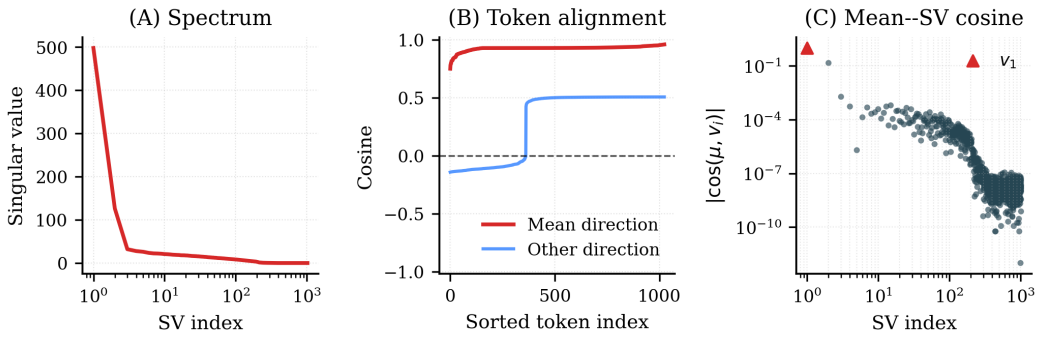


Figure 8: Additional mean-bias visualization for layer 0 at training step 170k. Compared with the early checkpoint, the same qualitative structure persists and the coherence of the mean component becomes stronger.

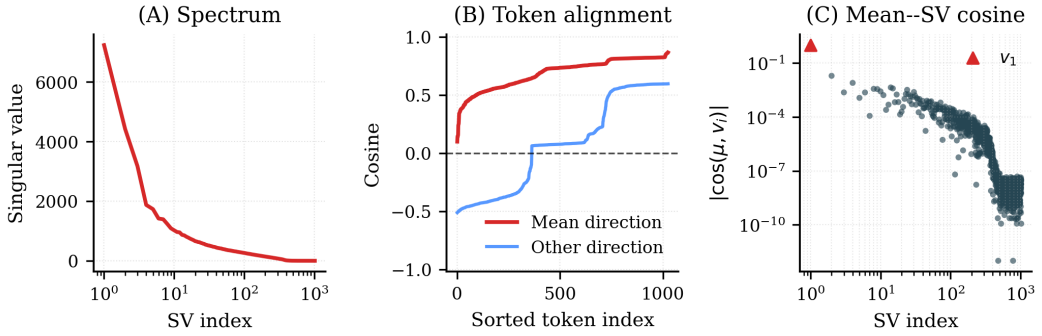


Figure 9: Additional mean-bias visualization for layer 27 at training step 170k. The phenomenon is especially pronounced in this deep late-stage activation, matching the strongest regime discussed in the main text.

For $k \neq p$, orthonormality of the left singular vectors gives

$$\mathrm{Cov}_I(u_{kI}, u_{pI}) = \frac{1}{l}\sum_{i=1}^{l} u_{ki}u_{pi} - \frac{\beta_k\beta_p}{l} = -\frac{\beta_k\beta_p}{l}. \tag{12}$$

Consequently, the cross-mode covariance contribution is

$$\begin{aligned}\tau^2_{j,\mathrm{cross}} &= -\frac{1}{l}\sum_{\substack{k,p=1\\k\neq p}}^{r} (\sigma_k\beta_k v_{kj})(\sigma_p\beta_p v_{pj}) \\ &= \frac{1}{l}\sum_{k=1}^{r} \sigma_k^2\beta_k^2 v_{kj}^2 - m_j^2.\end{aligned} \tag{13}$$

Here $\tau^2_{j,\mathrm{cross}}$ denotes a signed covariance contribution, not a nonnegative variance. Assumption 2 neglects this cross-mode contribution; its validity is supported by the relative cross-term magnitudes reported in Figure 10.

We empirically validate Assumption 2 by comparing the row-sampling residual variance of each column with its diagonal spectral approximation $\tau^2_{j,\mathrm{diag}}$. Figure 10 shows that the diagonal estimate closely tracks the empirical residual variance, while the relative contribution of cross-mode covariance terms remains small. Quantitatively, the cross-term contribution has median 0.006 and 95-th percentile 0.036, supporting the diagonal approximation used in the main-text tail model.

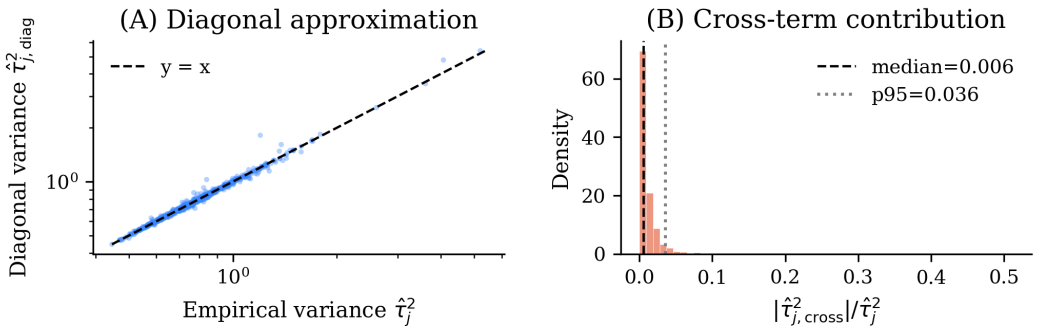
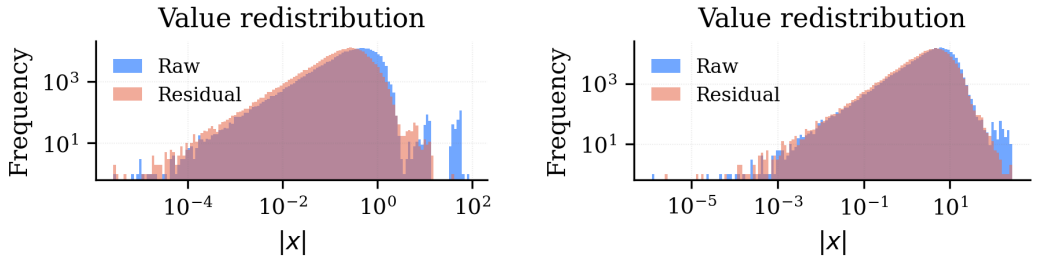


Figure 10: Empirical validation of the diagonal variance approximation. The diagonal spectral estimate closely matches the empirical residual variance, and cross-mode covariance terms contribute little to the total variance.

C Residual Tail Contraction After Mean Removal

To support the tail-amplification mechanism in Theorem 1, we directly compare the empirical value distributions of raw activations and their mean-centered residuals. Figures 11a and 11b show representative late-stage examples from a shallow layer (layer 0) and a deep layer (layer 27), both at the 170k checkpoint. In both cases, subtracting the feature-wise mean visibly contracts the high-magnitude tail of the distribution: the residual histogram is more concentrated near zero and exhibits substantially fewer extreme values than the raw activation histogram.

This pattern is consistent with the main-text interpretation. A coherent mean shift contributes directly to extreme entrywise magnitudes by translating the columnwise distribution away from zero. Once this mean component is removed, the remaining residual fluctuations are much less heavy-tailed, especially in the far-tail regime most relevant to blockwise quantization scales. These examples therefore provide direct empirical support for the claim that mean removal reduces the outlier population that drives FP4 dynamic-range inflation.



(a) Layer 0 at training step 170k. After subtracting the feature-wise mean, the residual distribution is more concentrated and exhibits a substantially weaker high-magnitude tail.

(b) Layer 27 at training step 170k. The contraction of the tail after mean removal is even more pronounced in this deep late-stage activation, where mean bias is strongest.

Figure 11: Raw-versus-residual value distributions at the 170k checkpoint for a shallow layer and a deep layer. In both cases, subtracting the feature-wise mean contracts the high-magnitude tail, with the effect becoming especially pronounced in the deep late-stage activation.

D Output-Gradient Mean Centering

Although output gradients exhibit a much weaker mean-bias structure than activations, mean centering still provides a small but consistent quantization benefit. Figure 12 shows a representative three-panel diagnostic for an output-gradient matrix. Compared with the activation examples in the main text, the leading spectral spike is weaker, token-wise alignment to the mean direction is less one-sided, and

the mean vector is less sharply aligned with the top right singular vector. In this sense, mean bias is not as pronounced for output gradients as it is for activations.

Nevertheless, quantization results still show a slight benefit from mean centering. Under NVFP4 quantization, the relative quantization error decreases from 13.6% to 13.5% after centering the output-gradient matrix. This gain is modest, but it is directionally consistent with the Averis design: even when the coherent mean component is weak, isolating it can still slightly improve low-bit numerical fidelity.

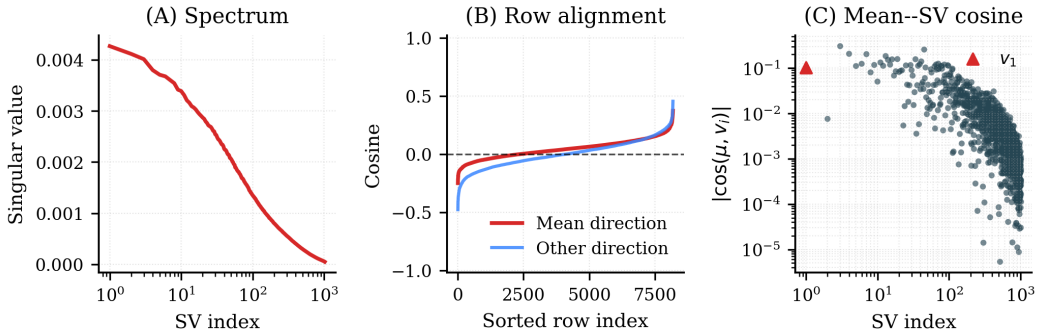


Figure 12: Three-panel visualization of a representative output-gradient matrix. Compared with activations, the mean-bias structure is weaker: spectral dominance is milder, token-wise alignment to the mean direction is less one-sided, and the mean vector is less sharply aligned with the top right singular vector. Even so, mean centering still yields a slight reduction in NVFP4 quantization error.

E Proof of Theorem 1

Spectral decomposition and residual variance

We first retain the spectral identities underlying the theorem. Throughout this derivation, probabilities are understood under the marginal row-coordinate model stated in the theorem, conditional on the spectral quantities entering m_j and τ_j. For notational simplicity, we keep the conditioning implicit when no ambiguity arises.

Step 1: Exact spectral expression for the column mean. By the singular value decomposition of $\mathbf{X}$,

$$X_{ij} = \sum_{k=1}^{r} \sigma_k u_{ki} v_{kj}.$$

Averaging over the row index i gives

$$m_j = \frac{1}{l} \sum_{i=1}^{l} X_{ij} = \frac{1}{l} \sum_{i=1}^{l} \sum_{k=1}^{r} \sigma_k u_{ki} v_{kj}.$$

Interchanging the summations yields

$$m_j = \sum_{k=1}^{r} \sigma_k v_{kj} \left(\frac{1}{l} \sum_{i=1}^{l} u_{ki} \right).$$

Since

$$\beta_k = \left\langle \mathbf{u}_k, \frac{\mathbf{1}}{\sqrt{l}} \right\rangle = \frac{1}{\sqrt{l}} \sum_{i=1}^{l} u_{ki},$$

we have

$$\frac{1}{l} \sum_{i=1}^{l} u_{ki} = \frac{\beta_k}{\sqrt{l}}.$$

Therefore,

$$m_j = \frac{1}{\sqrt{l}} \sum_{k=1}^{r} \sigma_k \beta_k v_{kj},$$

which is the column-mean expansion used in the theorem. The leading-mode contribution is correspondingly

$$m_j^{(1)} = \frac{\sigma_1 \beta_1}{\sqrt{l}} v_{1j}.$$

Step 2: Exact mean–fluctuation decomposition of the row-sampled entry. Let $I \sim \mathrm{Unif}(\{1, \dots, l\})$ and define

$$Y_j = X_{Ij}.$$

Using the SVD again,

$$Y_j = \sum_{k=1}^{r} \sigma_k u_{kI} v_{kj}.$$

For each singular mode, add and subtract its row-average contribution $\beta_k/\sqrt{l}$:

$$Y_j = \sum_{k=1}^{r} \sigma_k \left(\frac{\beta_k}{\sqrt{l}} + u_{kI} - \frac{\beta_k}{\sqrt{l}} \right) v_{kj}.$$

Splitting the two terms gives

$$Y_j = \underbrace{\frac{1}{\sqrt{l}} \sum_{k=1}^{r} \sigma_k \beta_k v_{kj}}_{m_j} + \underbrace{\sum_{k=1}^{r} \sigma_k \left(u_{kI} - \frac{\beta_k}{\sqrt{l}} \right) v_{kj}}_{\eta_{Ij}}.$$

Hence

$$Y_j = m_j + \eta_{Ij}.$$

Moreover, under the row-sampling distribution,

$$\mathbb{E}_I[\eta_{Ij}] = \mathbb{E}_I[Y_j] - m_j = \frac{1}{l} \sum_{i=1}^{l} X_{ij} - m_j = 0.$$

Thus η_{Ij} is a centered row-coordinate fluctuation.

Step 3: Residual variance scale and marginal Gaussian approximation. The theorem assumes a marginal Gaussian approximation for the row-sampled fluctuation:

$$\eta_{Ij} \dot{\sim} N(0, \tau_j^2).$$

This is a marginal assumption on the randomly sampled coordinate η_{Ij}. It does not require the row residuals $\{\eta_{ij}\}_{i=1}^{l}$ to be mutually independent.

The diagonal approximation $\tau_{j,\mathrm{diag}}^2$ keeps the diagonal mode-wise variance contributions. For a single singular mode,

$$\mathbb{E}_I \left[\left(u_{kI} - \frac{\beta_k}{\sqrt{l}} \right)^2 \right] = \frac{1}{l} \sum_{i=1}^{l} u_{ki}^2 - \left(\frac{1}{l} \sum_{i=1}^{l} u_{ki} \right)^2.$$

Since $\|\mathbf{u}_k\|_2 = 1$ and

$$\frac{1}{l} \sum_{i=1}^{l} u_{ki} = \frac{\beta_k}{\sqrt{l}},$$

we obtain

$$\mathbb{E}_I \left[\left(u_{kI} - \frac{\beta_k}{\sqrt{l}} \right)^2 \right] = \frac{1}{l} - \frac{\beta_k^2}{l} = \frac{1 - \beta_k^2}{l}.$$

Therefore, the diagonal contribution of mode k to the row-sampling variance of η_{Ij} is

$$\frac{1}{l}\sigma_k^2(1-\beta_k^2)v_{kj}^2.$$

The exact residual variance also includes the cross-mode covariance contribution in Equations (11)–(13). Assumption 2 neglects this contribution, supported by the relative cross-term magnitudes in Figure 10, giving the approximation

$$\tau_j^2 \approx \tau_{j,\mathrm{diag}}^2 = \frac{1}{l}\sum_{k=1}^{r}\sigma_k^2(1-\beta_k^2)v_{kj}^2.$$

Under the marginal Gaussian row-coordinate approximation,

$$Y_j = m_j + \eta_{Ij} \dot{\sim} N(m_j, \tau_j^2).$$

Notation for the tail analysis. Throughout this subsection we fix a column j, write $Y_j = X_{Ij} = m_j + \eta_{Ij}$ with $I \sim \mathrm{Unif}(\{1,\dots,l\})$, and condition on the spectral quantities that determine m_j and τ_j. Let φ and Φ be the standard normal density and CDF, $Q(x) = 1-\Phi(x)$, and

$$R(x) \triangleq \frac{Q(x)}{\varphi(x)}, \qquad \lambda(x) \triangleq \frac{\varphi(x)}{Q(x)} = \frac{1}{R(x)}, \qquad z_p \triangleq Q^{-1}(p), \tag{14}$$

the Mills ratio, the Gaussian hazard function, and the upper-p standard normal quantile. For a threshold $t > 0$ we write $s \triangleq |m_j|$, $\tau \triangleq \tau_j > 0$, and

$$a \triangleq \frac{t-s}{\tau}, \quad b \triangleq \frac{t+s}{\tau}, \quad c \triangleq \frac{t}{\tau}, \quad \Delta \triangleq \frac{c^2-a^2}{2} = \frac{s(2t-s)}{2\tau^2}, \quad \Delta' \triangleq \frac{b^2-c^2}{2} = \frac{s(2t+s)}{2\tau^2}. \tag{15}$$

Let $Y_j^{(0)} \sim N(0,\tau^2)$ be the zero-mean, variance-matched baseline, and define the amplification ratio and the constants

$$A(t) \triangleq \frac{\mathbb{P}(|Y_j| > t)}{\mathbb{P}(|Y_j^{(0)}| > t)}, \tag{16}$$

$$K_+(a,c) \triangleq \frac{c+\sqrt{c^2+4}}{a+\sqrt{a^2+8/\pi}}, \qquad K_-(a,c) \triangleq \max\Big\{1,\ \frac{c+\sqrt{c^2+8/\pi}}{a+\sqrt{a^2+4}}\Big\},$$

$$L(c) \triangleq \log\Big(1+\sqrt{\tfrac{\pi}{2}}\,(1+c)\Big). \tag{17}$$

For $p \in (0,1)$, $q_p(|Y|)$ denotes the unique $q \geq 0$ with $\mathbb{P}(|Y| > q) = p$.

Theorem 2 (Mean-bias amplification of columnwise outliers, non-asymptotic form)**.** *Fix a column j and suppose Assumption 1 holds exactly, i.e., $\eta_{Ij} \sim N(0,\tau^2)$ with $\tau > 0$. Then for every threshold $t > 0$ the following hold; parts (i)–(v) require no asymptotic condition.*

(i) ***Exact two-sided tail and one-sided dominance.***

$$\mathbb{P}(|Y_j| > t) = Q(a) + Q(b), \qquad Q(a) \ \leq\ \mathbb{P}(|Y_j| > t) \ \leq\ \big(1 + e^{-2ts/\tau^2}\big)\,Q(a). \tag{18}$$

(ii) ***Monotonicity in the mean.*** *For fixed t, the map $s \mapsto \mathbb{P}(|Y_j| > t)$ is strictly increasing on $[0,\infty)$. In particular, $A(t) > 1$ whenever $m_j \neq 0$.*

(iii) ***Exact amplification identity.***

$$A(t) = \frac{1}{2}\,e^{\Delta}\,\frac{R(a)}{R(c)} + \frac{1}{2}\,e^{-\Delta'}\,\frac{R(b)}{R(c)}. \tag{19}$$

(iv) ***Uniform bounds.*** *If $t \geq s$, then*

$$\frac{1}{2}\,K_-(a,c)\,e^{\Delta} \ \leq\ A(t) \ \leq\ \frac{1}{2}\,K_+(a,c)\,e^{\Delta} + \frac{1}{2}, \qquad K_+(a,c) \leq \sqrt{\tfrac{\pi}{2}}\,(1+c). \tag{20}$$

Consequently,

$$0 \ \leq\ \log\big(2A(t)\big) - \Delta \ \leq\ L(c), \qquad \textit{and} \qquad A(t) \ \geq\ \frac{1}{2}\exp\Big(\frac{s^2}{2\tau^2}\Big). \tag{21}$$

If instead $0 < t < s$, then $\mathbb{P}(|Y_j| > t) > 1/2$ and $A(t) > 1/\big(4Q(c)\big)$.

(v) ***Quantile (magnitude) shift.*** *For every $p \in (0, 1/2]$,*

$$s + \tau z_p \;\le\; q_p(|Y_j|) \;\le\; s + \tau z_{p/2}, \qquad q_p(|Y_j^{(0)}|) = \tau z_{p/2}, \tag{22}$$

and hence

$$\begin{gathered} -\tau\,(\log 2)\,R(z_p) \;\le\; q_p(|Y_j|) - q_p(|Y_j^{(0)}|) - |m_j| \;\le\; 0, \\ \tau\,(\log 2)\,R(z_p) \le \sqrt{\tfrac{\pi}{2}}\,(\log 2)\,\tau < 0.87\,\tau. \end{gathered} \tag{23}$$

(vi) ***Far-tail asymptotics.*** *Along any sequence with $a = (t-s)/\tau \to \infty$ and $ts/\tau^2 \to \infty$,*

$$\mathbb{P}(|Y_j| > t) \sim Q\Big(\frac{t - |m_j|}{\tau_j}\Big), \qquad A(t) \sim \frac{t}{2(t-|m_j|)} \exp\Big(\frac{2t|m_j| - m_j^2}{2\tau_j^2}\Big). \tag{24}$$

Moreover $K_-(a,c)/(c/a) \to 1$ and $K_+(a,c)/(c/a) \to 1$, so (20) *reduces to* (24) *in this regime.*

Proposition 1 (Stability of the tail under perturbations of the mean and the scale)**.** *For $s \ge 0$, $\tau > 0$ and $t > 0$ let $P(s,\tau;t) \triangleq \mathbb{P}\big(|N(s,\tau^2)| > t\big)$, which by Theorem 2(i) depends on the mean only through its absolute value. With τ fixed, write $\Delta(s)$ and $A(t;s)$ for the quantities* (15)–(16) *evaluated at mean magnitude s.*

(i) ***Mean perturbation (probability level).*** *For all $s_1, s_2 \ge 0$,*

$$\big|\log P(s_1,\tau;t) - \log P(s_2,\tau;t)\big| \;\le\; \frac{|s_1 - s_2|}{\tau}\,\big(\bar{a}_+ + 1\big), \qquad \bar{a} \triangleq \frac{t - \min\{s_1,s_2\}}{\tau}. \tag{25}$$

Since the baseline does not depend on s, the same bound holds for $|\log A(t;s_1) - \log A(t;s_2)|$.

(ii) ***Scale perturbation.*** *For $0 \le s \le t$ and $0 < \tau_1 \le \tau_2$,*

$$\big|\log P(s,\tau_1;t) - \log P(s,\tau_2;t)\big| \;\le\; M\,\log\frac{\tau_2}{\tau_1}, \qquad M \triangleq a_1(a_1+1) + b_1(b_1+1)\,e^{-2ts/\tau_2^2}, \tag{26}$$

where $a_1 = (t-s)/\tau_1$ and $b_1 = (t+s)/\tau_1$.

(iii) ***Mean perturbation (logarithmic scale).*** *Let $s_1 > 0$, $s \ge 0$, $t \ge \max\{s, s_1\}$ and $\rho \triangleq |s - s_1|/s_1$. Then*

$$\big|\Delta(s) - \Delta(s_1)\big| \le \rho\,\Delta(s_1), \qquad \big|\log(2A(t;s)) - \log(2A(t;s_1))\big| \le \rho\,\Delta(s_1) + L(c). \tag{27}$$

Consequently, $\log A(t;s)/\log A(t;s_1) \to 1$ along any sequence with $\rho \to 0$, $\Delta(s_1) \to \infty$ and $L(c)/\Delta(s_1) \to 0$.

Corollary 1 (Attribution to the rank-one mean-bias component)**.** *Fix j, let $\eta \sim N(0, \tau_j^2)$, and compare $Y_j = m_j + \eta$ with the leading-mode surrogate $Y_j^{(1)} = m_j^{(1)} + \eta$.*

(a) If Assumption 3(b) holds at t, then

$$e^{-\varepsilon} \;\le\; \frac{\mathbb{P}(|Y_j| > t)}{\mathbb{P}(|Y_j^{(1)}| > t)} \;\le\; e^{\varepsilon}, \tag{28}$$

and the same bounds hold for the ratio of the corresponding amplification factors.

(b) If only $\rho_j \triangleq |m_j - m_j^{(1)}|/|m_j^{(1)}|$ is small and $t \ge \max\{|m_j|, |m_j^{(1)}|\}$, then the amplification exponents satisfy $|\Delta_j - \Delta_j^{(1)}| \le \rho_j \Delta_j^{(1)}$: the exponent is approximated with relative error at most ρ_j, while the probability ratio in (28) *is bounded above by $\exp\big(\rho_j \Delta_j^{(1)} + L(c)\big)$.*

Across rows, the leading-mode mean contribution is the rank-one matrix $\frac{\sigma_1 \beta_1}{\sqrt{l}} \mathbf{1} v_1^\top$.

Interpretation of the perturbation bounds. A small relative mean error alone does not guarantee probability-level accuracy; Assumption 3(b) supplies a sufficient condition for the multiplicative bound. The scale-perturbation bound also quantifies the effect of the diagonal variance approximation: if $\tau_{j,\mathrm{diag}}^2/\tau_j^2 \in [1-\kappa, 1+\kappa]$ with $0 \le \kappa < 1$, then $|\log(\tau_{j,\mathrm{diag}}/\tau_j)| \le -\frac{1}{2}\log(1-\kappa)$. Together with Proposition 1(ii), this gives an upper bound on the log-probability error for $t \ge |m_j|$, rather than an estimate of the actual error.

Finite samples and quantization. With l entries in a column, the empirical exceedance frequency has resolution $1/l$; probabilities much smaller than $1/l$ cannot be reliably estimated from that column alone. A large amplification factor can therefore arise when the baseline probability is too small to estimate reliably. The quantile bound provides a complementary magnitude-level description relevant to quantization dynamic ranges; it does not by itself determine blockwise maxima or scales. For $0 < t < |m_j|$, the exceedance probability within the column is larger than $1/2$, although such magnitudes may still be large relative to other columns.

Lemma 1 (Elementary Gaussian tail facts)**.**

(a) $Q(x) < \varphi(x)/x$ *for all* $x > 0$.

(b) $R'(x) = xR(x) - 1 < 0$ *for all* $x \in \mathbb{R}$*; hence* R *is strictly decreasing on* $\mathbb{R}$.

(c) For all $x \geq 0$,

$$\frac{2}{x + \sqrt{x^2 + 4}} \;<\; R(x) \;\leq\; \frac{2}{x + \sqrt{x^2 + 8/\pi}}, \tag{29}$$

with equality in the upper bound at $x = 0$*, where* $R(0) = \sqrt{\pi/2}$.

(d) $\lambda(x) < x_+ + 1$ *for all* $x \in \mathbb{R}$.

(e) For $p \in (0, 1/2]$, $0 \leq z_{p/2} - z_p \leq (\log 2)\, R(z_p) \leq \sqrt{\pi/2}\, \log 2 < 0.87$.

Proof. (a) Since $\varphi'(u) = -u\varphi(u)$ and $u/x > 1$ for $u > x > 0$, $Q(x) = \int_x^\infty \varphi(u)\, du < \int_x^\infty \frac{u}{x}\varphi(u)\, du = \frac{\varphi(x)}{x}$.

(b) Using $Q' = -\varphi$ and $\varphi'(x) = -x\varphi(x)$, $R'(x) = \frac{-\varphi(x)^2 + x\varphi(x)Q(x)}{\varphi(x)^2} = xR(x) - 1$. For $x \leq 0$ we have $xR(x) \leq 0 < 1$; for $x > 0$, part (a) gives $xR(x) < 1$.

(c) Abramowitz & Stegun (1964, Eq. 7.1.13) state that for $y \geq 0$,

$$\frac{1}{y + \sqrt{y^2 + 2}} < e^{y^2} \int_y^\infty e^{-u^2} du \leq \frac{1}{y + \sqrt{y^2 + 4/\pi}}.$$

The substitution $v = \sqrt{2}\, u$ gives $Q(x) = \pi^{-1/2} \int_{x/\sqrt{2}}^\infty e^{-u^2} du$, and $e^{x^2/2} = \big(\sqrt{2\pi}\, \varphi(x)\big)^{-1}$. Hence, with $y = x/\sqrt{2}$, $e^{y^2} \int_y^\infty e^{-u^2} du = R(x)/\sqrt{2}$. Multiplying the displayed inequality by $\sqrt{2}$ and simplifying yields (29). At $x = 0$ the upper bound equals $2/\sqrt{8/\pi} = \sqrt{\pi/2} = R(0)$.

(d) For $x \geq 0$, part (c) gives $\lambda(x) = 1/R(x) < \frac{1}{2}\big(x + \sqrt{x^2 + 4}\big) \leq \frac{1}{2}(x + x + 2) = x + 1$. For $x < 0$, $Q(x) > 1/2$ and $\varphi(x) < \varphi(0) = (2\pi)^{-1/2}$, so $\lambda(x) < \sqrt{2/\pi} < 1$.

(e) Q is a continuous strictly decreasing bijection $\mathbb{R} \to (0, 1)$, so $\zeta(u) \triangleq Q^{-1}(e^{-u})$ is well defined and differentiable for $u > 0$. Differentiating $Q(\zeta(u)) = e^{-u}$ gives $\zeta'(u) = e^{-u}/\varphi(\zeta(u)) = R(\zeta(u)) > 0$. With $u_p = \log(1/p)$ we have $z_p = \zeta(u_p) \geq 0$ for $p \leq 1/2$, and

$$z_{p/2} - z_p = \int_{u_p}^{u_p + \log 2} R(\zeta(u))\, du \leq (\log 2)\, R(z_p),$$

because $\zeta(u) \geq z_p$ on the interval of integration and R is decreasing by (b). Finally, $R(z_p) \leq R(0) = \sqrt{\pi/2}$. □

PROOF OF THEOREM 2

Step 1: part (i). Under Assumption 1, $Y_j \sim N(m_j, \tau^2)$. Hence $\mathbb{P}(Y_j > t) = Q\big((t - m_j)/\tau\big)$ and $\mathbb{P}(Y_j < -t) = \Phi\big((-t - m_j)/\tau\big) = Q\big((t + m_j)/\tau\big)$. The sum is invariant under $m_j \mapsto -m_j$, so it equals $Q(a) + Q(b)$ with $s = |m_j|$. This is an exact identity, not an approximation. For the bounds, $Q(b) \geq 0$ gives the lower bound. Writing $Q = \varphi R$,

$$\frac{Q(b)}{Q(a)} = \exp\Big(-\frac{b^2 - a^2}{2}\Big) \frac{R(b)}{R(a)} \leq \exp\Big(-\frac{2ts}{\tau^2}\Big),$$

since $b^2 - a^2 = 4ts/\tau^2$ and $b \geq a$ implies $R(b) \leq R(a)$ by Lemma 1(b). This gives the upper bound in (18).

Step 2: part (ii). Since $\partial a/\partial s = -1/\tau$, $\partial b/\partial s = 1/\tau$ and $Q' = -\varphi$,

$$\frac{\partial}{\partial s}\big[Q(a) + Q(b)\big] = \frac{\varphi(a) - \varphi(b)}{\tau}.$$

For $s > 0$ we have $b - a = 2s/\tau > 0$ and $b + a = 2t/\tau > 0$, hence $b > |a|$. Because φ is even and strictly decreasing on $[0, \infty)$, $\varphi(a) > \varphi(b)$. The derivative is therefore strictly positive on $(0, \infty)$, and by continuity at $s = 0$ the map is strictly increasing on $[0, \infty)$. At $s = 0$ it equals the baseline $2Q(c)$, so $A(t) > 1$ for $s > 0$.

Step 3: part (iii). Divide (18) by $\mathbb{P}(|Y_j^{(0)}| > t) = 2Q(c)$ and use $Q = \varphi R$ together with $\varphi(a)/\varphi(c) = e^{(c^2-a^2)/2} = e^{\Delta}$ and $\varphi(b)/\varphi(c) = e^{-(b^2-c^2)/2} = e^{-\Delta'}$.

Step 4: part (iv). Let $t \geq s$, so that $0 \leq a \leq c \leq b$ and $\Delta, \Delta' \geq 0$.

Lower bound. The second term of (19) is non-negative. Since $a \leq c$, Lemma 1(b) gives $R(a)/R(c) \geq 1$. Using the lower bound of (29) for $R(a)$ and the upper bound for $R(c)$ gives $R(a)/R(c) > (c + \sqrt{c^2 + 8/\pi})/(a + \sqrt{a^2 + 4})$. Hence $R(a)/R(c) \geq K_-(a, c)$ and $A(t) \geq \frac{1}{2} K_- e^{\Delta}$.

Upper bound. Using the upper bound of (29) for $R(a)$ and the lower bound for $R(c)$ gives $R(a)/R(c) \leq K_+(a, c)$. In the second term, $e^{-\Delta'} \leq 1$ and $R(b)/R(c) \leq 1$ because $b \geq c$. Hence $A(t) \leq \frac{1}{2} K_+ e^{\Delta} + \frac{1}{2}$. Moreover, since $a \geq 0$ and $\sqrt{c^2 + 4} \leq c + 2$,

$$K_+(a, c) \leq \frac{c + \sqrt{c^2 + 4}}{\sqrt{8/\pi}} \leq \frac{2(1 + c)}{\sqrt{8/\pi}} = \sqrt{\tfrac{\pi}{2}}\,(1 + c).$$

Logarithmic form. From the lower bound, $2A(t) \geq K_- e^{\Delta} \geq e^{\Delta}$. From the upper bound and $\Delta \geq 0$, $2A(t) \leq e^{\Delta}(K_+ + e^{-\Delta}) \leq e^{\Delta}(1 + K_+) \leq e^{\Delta}\big(1 + \sqrt{\pi/2}\,(1 + c)\big)$. Taking logarithms gives (21). Finally, $2t - s \geq s$ implies $\Delta \geq s^2/(2\tau^2)$, which gives $A(t) \geq \frac{1}{2} e^{s^2/(2\tau^2)}$.

Case $0 < t < s$. Then $a < 0$, so $\mathbb{P}(|Y_j| > t) \geq Q(a) > 1/2$ and $A(t) > 1/(4Q(c))$.

Step 5: part (v). Let $F(q) \triangleq \mathbb{P}(|Y_j| > q) = Q\big(\frac{q-s}{\tau}\big) + Q\big(\frac{q+s}{\tau}\big)$ for $q \geq 0$. F is continuous with $F'(q) = -\tau^{-1}\big[\varphi(\frac{q-s}{\tau}) + \varphi(\frac{q+s}{\tau})\big] < 0$, $F(0) = 1$ and $F(q) \to 0$ as $q \to \infty$, so $q_p = F^{-1}(p)$ is unique for $p \in (0, 1)$. Fix $p \in (0, 1/2]$, so that $z_p \geq 0$ and $z_{p/2} > 0$.

Lower bound. Put $q^- = s + \tau z_p \geq 0$. Then $F(q^-) \geq Q\big((q^- - s)/\tau\big) = Q(z_p) = p = F(q_p)$. Since F is decreasing, $q_p \geq q^-$.

Upper bound. Put $q^+ = s + \tau z_{p/2} > 0$. Since $s \geq 0$ we have $(q^+ + s)/\tau \geq (q^+ - s)/\tau = z_{p/2}$, and therefore $F(q^+) \leq 2Q(z_{p/2}) = p = F(q_p)$. Hence $q_p \leq q^+$.

Baseline. For $s = 0$, $F^{(0)}(q) = 2Q(q/\tau)$, so $q_p(|Y_j^{(0)}|) = \tau z_{p/2}$.

Subtracting the baseline from (22) gives $s - \tau(z_{p/2} - z_p) \leq q_p(|Y_j|) - q_p(|Y_j^{(0)}|) \leq s$, and Lemma 1(e) yields (23).

Step 6: part (vi). By (18), $1 \leq \mathbb{P}(|Y_j| > t)/Q(a) \leq 1 + e^{-2ts/\tau^2} \to 1$ when $ts/\tau^2 \to \infty$, which proves the first relation in (24). For the second, (18) and $Q = \varphi R$ give

$$A(t) = \frac{Q(a)}{2Q(c)}\Big(1 + \frac{Q(b)}{Q(a)}\Big) = \frac{1}{2}\, e^{\Delta}\, \frac{R(a)}{R(c)}\Big(1 + \frac{Q(b)}{Q(a)}\Big).$$

The last factor tends to 1 by Step 1. By (29), $xR(x) \to 1$ as $x \to \infty$. Since $c \geq a \to \infty$,

$$\frac{R(a)}{R(c)} = \frac{c}{a} \cdot \frac{aR(a)}{cR(c)} \sim \frac{c}{a} = \frac{t}{t - s}.$$

Substituting $\Delta = (2ts - s^2)/(2\tau^2)$ proves (24). The same computation shows that both $K_+(a, c)$ and $(c + \sqrt{c^2 + 8/\pi})/(a + \sqrt{a^2 + 4})$ are asymptotic to c/a; since $c/a \geq 1$, it follows that $K_-(a, c)/(c/a) \to 1$ as well. Note that for $a \to 0$ the exact prefactor stays bounded by $\sqrt{\pi/2}(1 + c)$ by Step 4, whereas $c/a \to \infty$; the divergence of (24) near $t = |m_j|$ is therefore an artifact of the asymptotic approximation. □

PROOF OF PROPOSITION 1

Part (i). Let $G(s) \triangleq \log P(s, \tau; t)$ for $s \geq 0$, with $a(s) = (t - s)/\tau$ and $b(s) = (t + s)/\tau$. By Step 2 of the proof of Theorem 2,

$$G'(s) = \frac{\varphi(a) - \varphi(b)}{\tau\,[Q(a) + Q(b)]}.$$

The numerator is non-negative (Step 2). Since $Q(a) + Q(b) \geq Q(a)$ and $\varphi(b) \geq 0$,

$$0 \leq G'(s) \leq \frac{\varphi(a)}{\tau\, Q(a)} = \frac{\lambda(a(s))}{\tau} < \frac{a(s)_+ + 1}{\tau},$$

using Lemma 1(d). For s between s_1 and s_2 we have $a(s) \leq (t - \min\{s_1, s_2\})/\tau = \bar{a}$, hence $a(s)_+ \leq \bar{a}_+$. The mean value theorem gives (25). The baseline $2Q(t/\tau)$ does not depend on s, so $\log A(t; s_1) - \log A(t; s_2) = G(s_1) - G(s_2)$.

Part (ii). Let $H(u) \triangleq \log P(s, e^u; t)$. With $\tau = e^u$ we have $da/du = -a$ and $db/du = -b$, so

$$H'(u) = \frac{a\varphi(a) + b\varphi(b)}{Q(a) + Q(b)}.$$

Since $t \geq s$, we have $a \geq 0$ and $b > 0$, so $H'(u) \geq 0$. For the upper bound, bound the two terms separately: $a\varphi(a)/[Q(a) + Q(b)] \leq a\lambda(a) \leq a(a + 1)$, and, by Step 1,

$$\frac{b\varphi(b)}{Q(a) + Q(b)} \leq \frac{b\varphi(b)}{Q(a)} = b\,\lambda(b)\,\frac{Q(b)}{Q(a)} \leq b(b + 1)\, e^{-2ts/\tau^2}.$$

On $\tau \in [\tau_1, \tau_2]$, both $a = (t - s)/\tau$ and $b = (t + s)/\tau$ are maximized at τ_1, while e^{-2ts/τ^2} is maximized at τ_2. Hence $0 \leq H'(u) \leq M$ on $[\log \tau_1, \log \tau_2]$, and the mean value theorem gives (26).

Part (iii). From (15),

$$\Delta(s) - \Delta(s_1) = \frac{s(2t - s) - s_1(2t - s_1)}{2\tau^2} = \frac{(s - s_1)(2t - s - s_1)}{2\tau^2}, \qquad \Delta(s_1) = \frac{s_1(2t - s_1)}{2\tau^2}.$$

Because $t \geq \max\{s, s_1\}$ and $s \geq 0$, we have $0 \leq 2t - s - s_1 \leq 2t - s_1$ and $2t - s_1 > 0$, so

$$\frac{|\Delta(s) - \Delta(s_1)|}{\Delta(s_1)} = \frac{|s - s_1|}{s_1} \cdot \frac{2t - s - s_1}{2t - s_1} \leq \rho.$$

For the second inequality in (27), write $\log(2A(t; \cdot)) = \Delta(\cdot) + E(\cdot)$. By (21), applicable since $t \geq \max\{s, s_1\}$, both $E(s)$ and $E(s_1)$ lie in $[0, L(c)]$, so $|E(s) - E(s_1)| \leq L(c)$. Finally,

$$\frac{\log A(t; s)}{\log A(t; s_1)} = \frac{\Delta(s) + E(s) - \log 2}{\Delta(s_1) + E(s_1) - \log 2} \longrightarrow 1$$

whenever $\rho \to 0$, $\Delta(s_1) \to \infty$ and $L(c)/\Delta(s_1) \to 0$, because then $\Delta(s)/\Delta(s_1) \to 1$ and $E(\cdot)$ and $\log 2$ are $o(\Delta(s_1))$. □

PROOF OF COROLLARY 1

Both Y_j and $Y_j^{(1)}$ are Gaussian with variance τ_j^2 and mean magnitudes $s = |m_j|$ and $s_1 = |m_j^{(1)}|$. Since $\big||m_j| - |m_j^{(1)}|\big| \leq |m_j - m_j^{(1)}|$, Proposition 1(i) together with (4) gives $|\log P(s, \tau_j; t) - \log P(s_1, \tau_j; t)| \leq \varepsilon_j(t) \leq \varepsilon$, which is (28); the statement for amplification factors follows because both share the baseline $2Q(t/\tau_j)$. Part (b) is Proposition 1(iii) with $\rho \leq \rho_j$. □

F Model Architectures and Training Configurations

Tables 4 and 5 give the model and optimization configurations. Within each model setting, BF16 and FP4 variants use the same architecture, data order, optimizer schedule, and training hyperparameters; only the numerical format and the stated quantization preprocessing differ. Both models are trained on DCLM to 100B tokens. The downstream scores in the main table are evaluated at the 100B-token checkpoint for both models.

Table 4: Model architecture configurations. Qwen3-7B-A1.5B is the MoE configuration used in our experiments.

Configuration	Qwen3-0.6B Dense	Qwen3-7B-A1.5B MoE
Transformer layers	28	40
Hidden size	1,024	1,536
Dense FFN hidden size	3,072	1,024
Attention heads	16	16
KV heads/query groups	8	8
Head/KV channel dimension	128	128
Vocabulary size	151,936	151,936
Normalization	RMSNorm	RMSNorm
RMSNorm epsilon	10^{-6}	10^{-6}
Activation	SiLU-gated FFN (SwiGLU)	SwiGLU
Attention dropout	0.0	0.0
Position encoding	RoPE	RoPE
RoPE base	10^6	10^6
Tied input/output embeddings	Yes	No
Number of experts	–	32
Activated experts per token	–	3
Expert FFN hidden size	–	1,024
Router	–	FP32 softmax, pre-softmax Top-3
Router auxiliary-loss coefficient	–	0.03

Precision notation and execution. W4A4G4 specifies FP4 GeMM operands: W4 denotes weights, A4 denotes forward activations, and G4 denotes the gradient-path operands in DGRAD and WGRAD, rather than a quantization group size of four. NVFP4 shares an 8-bit E4M3 scale across 16 E2M1 values and uses a per-tensor FP32 scale. For a tensor with N elements, this representation requires $4 + 8/16 + 32/N = 4.5 + 32/N$ bits per element, including scale metadata; this is not the memory cost of the complete training state. Long-run training-quality experiments use FP4 simulation on Hopper GPUs. The preprocessing and end-to-end step latencies are measured on NVIDIA GB300 GPUs using native Blackwell FP4 hardware.

G Mean Extraction in MoE Layers

For MoE expert layers, Averis computes the mean independently for each expert after token dispatch. Specifically, let $\mathbf{X}_e \in \mathbb{R}^{l_e \times m}$ contain the $l_e > 0$ tokens dispatched to expert e for a given linear operation, and let $\mathbf{D}_e \in \mathbb{R}^{l_e \times n}$ be the corresponding derivative with respect to that operation's output. Using the row-vector mean convention of Section 3, define

$$\mu_{\mathbf{X},e} = \frac{1}{l_e}\mathbf{1}_{l_e}^\top \mathbf{X}_e, \qquad \mu_{\mathbf{D},e} = \frac{1}{l_e}\mathbf{1}_{l_e}^\top \mathbf{D}_e.$$

The expert residuals are $\mathbf{X}_{\mathrm{R},e} = \mathbf{X}_e - \mathbf{1}_{l_e}\mu_{\mathbf{X},e}$ and $\mathbf{D}_{\mathrm{R},e} = \mathbf{D}_e - \mathbf{1}_{l_e}\mu_{\mathbf{D},e}$. Their column sums vanish over the same expert-local token set, giving the exact identity before quantization

$$\mathbf{X}_e^\top \mathbf{D}_e = \mathbf{X}_{\mathrm{R},e}^\top \mathbf{D}_{\mathrm{R},e} + l_e \mu_{\mathbf{X},e}^\top \mu_{\mathbf{D},e}.$$

Table 5: Training hyperparameters and parallelism configurations. Gradient clipping is disabled for Dense and uses threshold 1.0 for MoE.

Hyperparameter	Qwen3-0.6B Dense	Qwen3-7B-A1.5B MoE
Dataset	DCLM	DCLM
Training tokens	100B	100B
Sequence length	1,024	2,048
Local/microbatch size	16 per process	1 per process
Global batch size	256 warmup, then 512	512
Batch-size warmup	2,000 steps	None
Optimizer	AdamW	AdamW
Optimizer distribution	Replicated under DDP	Distributed optimizer
Adam $(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.95, 10^{-8})$	$(0.9, 0.95, 10^{-8})$
Peak learning rate	1×10^{-4}	1×10^{-4}
LR schedule	Cosine decay	Cosine decay
Minimum LR	0	1×10^{-5}
LR warmup	2,000 steps	1,000 steps
Weight decay	0.01	0.1
Gradient clipping	None	1.0
Data-parallel world size	8	16 (2 nodes $\times$ 8 GPUs)
Tensor-parallel size	1	1
Pipeline-parallel size	1	1
Expert-parallel size	1	4
Low-bit format	E2M1 NVFP4 for W/A/G	E2M1 NVFP4 for W/A/G
Backward rounding	Stochastic rounding for output gradients	Stochastic rounding for output gradients

Averis quantizes these components independently and evaluates

$$\widehat{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_e}} = \bar{\mathbf{X}}_{\mathrm{R},e}^{\top} \bar{\mathbf{D}}_{\mathrm{R},e} + l_e \bar{\mu}_{\mathbf{X},e}^{\top} \bar{\mu}_{\mathbf{D},e}.$$

The forward and input-gradient formulas apply separately to each expert with the same expert-local means. Thus, each expert uses the mean of its own GeMM inputs, rather than a mean computed across tokens before routing; l_e denotes that expert's token count, not the global training-batch size. The exact cancellation above concerns the unquantized centered residuals, while the quantized expression is an approximation to that two-term identity.

H Sensitivity of Mean Statistics to Input Composition

We examine microbatch size, sequence length, padding, and sequence packing using the layer-27 FFN-input activations of BF16 Qwen3-0.6B. For setting s, let $\boldsymbol{\mu}_s$ denote the feature-wise mean over tokens and $\mathbf{v}_{1,s}$ the leading right singular vector of the corresponding uncentered activation matrix. We report the mean-reference cosine $\cos(\boldsymbol{\mu}_s, \boldsymbol{\mu}_{\mathrm{ref}})$ and the mean–leading-direction cosine $|\cos(\boldsymbol{\mu}_s, \mathbf{v}_{1,s})|$. These measure the stability of the mean direction and its alignment with the dominant spectral direction, respectively; the reference is specified separately for each comparison.

Microbatch size. Starting from 16 sequences of length 1,024, we take the first k sequences, for $k = 1, \ldots, 16$, and compute the mean and leading right singular vector over the resulting $1{,}024k$ tokens. Table 6 reports the cosine between consecutive mean estimates, $\cos(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k-1})$, and their spectral alignment. The consecutive-mean cosine exceeds 0.95 for every $k \geq 2$ and reaches 0.9993 at $k = 16$. Alignment with the leading direction remains above 0.997 throughout, including for a single sequence. The direction is thus already apparent in small microbatches and becomes more stable as additional sequences are included.

Sequence length. Each shorter setting uses a strict prefix of the same 1,024-token sequences, with the length-1,024 setting as reference. As shown in Table 7, even at length 128 the mean-reference

Table 6: Mean-direction sensitivity to microbatch size. Each sequence contains 1,024 tokens. The first cosine compares consecutive microbatch prefixes, rather than a fixed reference; it is undefined for $k = 1$.

Sequences k	$\cos(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k-1})$	$\lvert\cos(\boldsymbol{\mu}_k, \mathbf{v}_{1,k})\rvert$
1	–	0.9995
2	0.9525	0.9978
3	0.9708	0.9986
4	0.9925	0.9976
5	0.9829	0.9983
6	0.9939	0.9986
7	0.9983	0.9987
8	0.9975	0.9986
9	0.9961	0.9986
10	0.9980	0.9982
11	0.9986	0.9983
12	0.9985	0.9982
13	0.9990	0.9983
14	0.9993	0.9983
15	0.9993	0.9984
16	0.9993	0.9985

cosine is 0.9898 and the mean–leading-direction cosine is 0.9987. Both remain close to one across the evaluated lengths.

Padding and valid-token masking. We fix 512 valid tokens per sequence and vary the padded length from 512 to 1,024, giving padding ratios from 0% to 50%. In the masked condition, padded positions are excluded from the mean reduction and the denominator counts only valid tokens. If $a_i \in \{0, 1\}$ is the valid-token indicator and $\mathbf{x}_i$ is a token's feature vector, the statistic is

$$N_{\text{valid}} = \sum_i a_i, \qquad \boldsymbol{\mu}_{\text{valid}} = \frac{1}{N_{\text{valid}}} \sum_i a_i \mathbf{x}_i, \qquad N_{\text{valid}} > 0.$$

The no-padding setting with the same valid tokens is the reference. Masking yields a mean-reference cosine of 1.0000 at the reported precision across the tested padding ratios. For comparison, the unmasked condition includes padded positions in both the sum and the token count; its mean-reference cosine decreases to 0.9608 at 50% padding. We therefore use valid-token masking when computing mean statistics on padded inputs, rather than treating padding as additional observations. A valid-token mask in this reduction is distinct from a causal or sequence-boundary attention mask.

Sequence packing. We compare four independent 256-token sequences with the same tokens concatenated into one 1,024-token sequence, using the unpacked setting as reference. The mean-reference cosine is 0.9996, and the mean–leading-direction cosine is 0.9992 in both settings. This comparison changes sequence boundaries and attention context; it shows that the mean direction remains similar under this particular packing construction.

These diagnostics support stability of the mean direction and its leading-mode alignment in the evaluated setting. They do not by themselves establish invariance of mean magnitude, quantization error, or end-to-end training quality across input compositions, layers, or model architectures.

Table 7: Sensitivity to sequence length, padding, and packing. The two metrics are mean-reference cosine and absolute cosine with the leading right singular vector. Masked padding statistics exclude padded positions; unmasked statistics include them.

Factor	Setting	Mean-reference	Mean–$\mathbf{v}_1$
Sequence length	128	0.9898	0.9987
	256	0.9941	0.9989
	512	0.9971	0.9992
	1,024 (reference)	1.0000	0.9990
Padding, masked	0–50%	1.0000	0.9992
Padding, unmasked	20%	0.9935	0.9990
	33%	0.9825	0.9987
	50%	0.9608	0.9988
Packing	Unpacked 4×256 (reference)	1.0000	0.9992
	Packed $1 \times 1{,}024$	0.9996	0.9992

I Quantization-Induced Cross Terms in the Weight Gradient

The cancellation in Section 3 applies to the unquantized centered residuals. Averis quantizes the two nonzero terms of this exact identity; it does not require the quantized residuals to remain exactly zero-mean. To evaluate the effect of restoring the cross terms introduced by quantization, we compare the two-term estimate with the complete four-term expansion of the quantized reconstructions. Using the row-vector mean convention of Section 3, define

$$\begin{aligned} \mathbf{R} &= \bar{\mathbf{X}}_{\mathrm{R}}^{\top}\bar{\mathbf{D}}_{\mathrm{R}}, \qquad \mathbf{M} = l\bar{\mu}_{\mathbf{X}}^{\top}\bar{\mu}_{\mathbf{D}}, \\ \mathbf{C}_1 &= \bar{\mathbf{X}}_{\mathrm{R}}^{\top}\mathbf{1}_l\bar{\mu}_{\mathbf{D}}, \quad \mathbf{C}_2 = \bar{\mu}_{\mathbf{X}}^{\top}\mathbf{1}_l^{\top}\bar{\mathbf{D}}_{\mathrm{R}}. \end{aligned} \tag{30}$$

The two estimates are

$$\widehat{\mathbf{G}}_{\mathrm{two}} = \mathbf{R} + \mathbf{M}, \qquad \widehat{\mathbf{G}}_{\mathrm{four}} = \mathbf{R} + \mathbf{C}_1 + \mathbf{C}_2 + \mathbf{M}.$$

Here $\mathbf{C}_1$ and $\mathbf{C}_2$ can be nonzero because quantization does not preserve exact centering. Both estimates are compared with the unquantized reference $\mathbf{G} = \mathbf{X}^{\top}\mathbf{D}$ using the relative Frobenius error

$$E(\widehat{\mathbf{G}}) = \frac{\|\widehat{\mathbf{G}} - \mathbf{G}\|_F}{\|\mathbf{G}\|_F}.$$

Experimental setup. We use layer-27 FFN activations and output gradients from Qwen3-0.6B at the 150K training-step checkpoint, evaluating the gate, up, and down projections. Activation operands use NVFP4 without stochastic rounding, while output-gradient operands use NVFP4 with stochastic rounding. Results are averaged over ten stochastic-rounding trials on these tensors, rather than independent training runs.

Table 8: Two-term versus four-term weight-gradient approximation for the layer-27 FFN of Qwen3-0.6B at the 150K checkpoint. Relative Frobenius errors are reported as percentages ($100E$), averaged over ten stochastic-rounding trials. Lower is better.

FFN projection	Four-term error (%)	Two-term error (%)
Gate	18.91	**17.74**
Up	19.18	**17.87**
Down	18.28	**16.64**

Table 8 shows that, for all three tested projections, adding the quantization-induced cross terms increases the weight-gradient approximation error. This comparison supports the two-term computation used by Averis without requiring exact centering after quantization; it does not imply that the two-term estimate must be more accurate for every tensor or quantization realization.