ANALYSIS OF MXFP4 QUANTIZATION ERROR FOR LLM REINFORCEMENT LEARNING: REDUCIBLE BIAS, RECOVERABLE DEADZONE, AND AN IRREDUCIBLE ERROR

Anonymous authors
Paper under double-blind review

ABSTRACT

MXFP4 arithmetic can dramatically accelerate reinforcement learning (RL) post-training of large language models (LLMs), yet the quantization error introduces severe accuracy degradation. Existing work treats the quantization error as a monolithic noise term, missing the distinct mechanisms upon interpreting how quantization error damages training. We prove that this error is not monolithic but admits an exact, *privileged* decomposition into three additive components: *scale bias* from power-of-two rounding, *deadzone truncation* from zeroing small values, and *grid noise* from rounding to the 4-bit grid. The deadzone component is exactly orthogonal to the other two, and, crucially, the grid component is *invariant to scale precision*: refining the block scale reduces scale bias without touching grid noise, driving the total error toward an irreducible floor set by the grid and deadzone alone. Each component dominates a distinct RL failure mode: scale bias accumulates multiplicatively through the backward pass, affecting gradient accuracy; deadzone truncation degrades rollout quality; and grid noise raises the policy's entropy. Our contribution is this diagnostic framework rather than a new correction: to validate it, we use three *existing*, mechanism-targeted techniques as controlled interventions (Macro-block scaling for scale bias, Outlier Fallback for deadzone, and Adaptive Quantization Noise for grid noise/entropy), each removing one component and recovering its predicted RL pathway. On Qwen2.5-3B (dense) this recovers BF16 accuracy to within 0.7 pp, on Qwen3-30B-A3B-Base (MoE) to parity, and on Llama-3.1-8B (a different family) to within 3.2 pp. The same component-to-pathway mapping holds across all three, and the results generalize from GSM8K to MATH500.

1 INTRODUCTION

RL post-training turns base LLMs into capable reasoners (Shao et al., 2024; Yu et al., 2025), but the compute cost is steep: rollout generation alone can dominate the training budget. The MXFP4 format (Rouhani et al., 2023) has 4-bit elements with a shared E8M0 scale per block of 32, offering up to $4\times$ throughput and $4\times$ memory reduction, with native support on NVIDIA Blackwell, AMD MI350, and other accelerators. However, simply replacing BF16 with MXFP4 during RL post-training produces a large accuracy gap. To restore the accuracy, quantization-aware training Liu et al. (2024) and post-training quantization Lin et al. (2024); Frantar et al. (2023) techniques are employed to mitigate this degradation. These works treat quantization error as a single noise source.

Different from how previous works view the quantization error, our novel perspective decomposes the MXFP4 quantization error into three additive error components induced by: (1) *Scale bias*: the E8M0 scale overshoots the ideal scale (E8M∞) by $\sim 44\%$ on average, a bias that accumulates multiplicatively across layers in backward pass, affecting the gradient accuracy. This error can be reduced by improving scale precision. (2) *Deadzone truncation*: absolute values below $\frac{0.5}{2} \div 6 = 1/24$ of the block absolute maximum are zeroed, pruning 9% of weights and degrading rollout quality. The weight values zeroed by the deadzone are recoverable by a known technique called Outlier Fallback (OF) Zhang et al. (2025). (3) *Grid noise*: the coarse E2M1 grid adds zero-mean rounding noise that

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.

raises the policy's effective temperature, increasing exploration. This component is *invariant to scale precision*, as it only depends on weights and the E2M1 grid, not the block scale. Therefore, once the precision format is given, this error is irreducible.

Contributions.

1. A privileged error basis. We show MXFP4 error decomposes into three additive components (scale bias, deadzone truncation, grid noise) that are structurally decoupled in a way an arbitrary split is not: the deadzone term is *exactly* orthogonal to the other two (Lemma 3.2), while scale bias and grid noise hold a fixed anti-correlation ($\cos \approx -0.66$) stable across 19,100 tensors and three models. The grid term is moreover *invariant to scale precision*, establishing an irreducible error floor that no amount of scale refinement can cross.
2. A *dominance analysis* mapping each error component to a specific RL failure mode: scale bias causes inaccurate gradient, deadzone causes rollout degradation in forward pass, and grid noise increases the policy entropy.
3. To validate the diagnosis, we use three *existing* techniques as mechanism-targeted controlled interventions, one per failure mode: Macro-block Scaling (MBS) (Chhugani et al., 2026), Outlier Fallback (OF) (Zhang et al., 2025), and Adaptive Quantization Noise (AQN) (Huang et al., 2026). These are prior, off-the-shelf techniques used here as probes; their role is to test that removing a component recovers its predicted RL pathway.
4. We confirm this across three models spanning two families and both dense and mixture-of-experts architectures (Qwen2.5-3B dense, Qwen3-30B-A3B-Base MoE, and Llama-3.1-8B) during RL post-training: MBS+AQN+OF recovers BF16 accuracy to within 0.7 percentage points on dense and to parity on MoE, the component-to-pathway mapping reproduces on the new family, and the results generalize from GSM8K to MATH500.

2 Related work

Post-training quantization (PTQ). Early LLM quantization focused on integer formats: LLM.int8() (Dettmers et al., 2022) and ZeroQuant (Yao et al., 2022) demonstrated INT8 inference with round-to-nearest over weight groups. GPTQ (Frantar et al., 2023) pushed to INT4 via second-order weight adjustments, and AWQ (Lin et al., 2024) introduced activation-aware scaling. SqueezeLLM (Kim et al., 2024) and SpQR (Dettmers et al., 2024) handle outliers with mixed dense-sparse representations. These methods assume uniform integer grids and relatively large group sizes; MXFP4's non-uniform E2M1 grid and small block size (32) change the quantization dynamics fundamentally, as our decomposition reveals.

Rotation-based and extreme compression. Incoherence processing and rotations reshape the weight/activation distribution *upstream* of quantization: QuIP/QuIP#/QTIP (Chee et al., 2023; Tseng et al., 2024a;b), SmoothQuant (Xiao et al., 2023), QuaRot (Ashkboos et al., 2024), SpinQuant (Liu et al., 2025), and more recent variants (Choi et al., 2025; Liang et al., 2025; Egiazarian et al., 2026). In our decomposition's terms these primarily reduce grid and deadzone error; our corrections operate *during or after* quantization and are complementary.

Quantization-aware training (QAT). LLM-QAT (Liu et al., 2024) performs data-free QAT by using the pre-trained model to generate training data. QLoRA (Dettmers et al., 2023) combines NF4 weight quantization with LoRA adapters for memory-efficient fine-tuning. QeRL (Huang et al., 2026) applies LoRA-based QAT specifically to RL post-training, demonstrating that quantization noise can enhance exploration. Our work differs from QeRL in two ways: we use full-parameter updates (no LoRA), providing cleaner causal evidence, and we ground the exploration benefit in our error decomposition. Specifically, the grid noise component acts as a natural entropy regularizer.

Low-precision FP4/FP8 training. A growing line of work targets training (not just inference) in 4–8 bit floating-point formats. On the FP8 side, Micikevicius et al. (2022) introduced the E4M3/E5M2 formats and Fishman et al. (2025); Hernández-Cano et al. (2025) demonstrated trillion-token-scale FP8 LLM pre-training, identifying activation-outlier instability as the central obstacle. On the FP4 side, beyond the original MXFP4 spec (Rouhani et al., 2023) and the MXFP4-RL pre-training of Tseng et al. (2025), recent work targets native FP4 training (Quartet (Castro et al., 2025), FP4 All The Way (Chmiel et al., 2025)), NVFP4 with finer scale precision (NVIDIA, 2025; Panferov

Table 1: Empirical error decomposition across architectures and model scales. All three tested models show the same structure: scale and grid are anti-correlated ($\cos \approx -0.66$); deadzone is exactly orthogonal to both; the grid component is $\sim 71\%$ of total MSE.

Model	$\frac{\|e^{\text{scale}}\|^2}{\|e\|^2}$	$\frac{\|e^{\text{DZ}}\|^2}{\|e\|^2}$	$\frac{\|e^{\text{grid}}\|^2}{\|e\|^2}$	$\frac{2\langle e^{\text{scale}}, e^{\text{grid}}\rangle}{\|e\|^2}$	cos	# tensors
Qwen2.5-3B	1.725	0.026	0.703	-1.454	-0.657	252
Llama-3.1-8B-Instruct	1.726	0.022	0.713	-1.461	-0.657	224
Qwen3-30B-A3B-Base	1.726	0.022	0.712	-1.460	-0.658	18 624

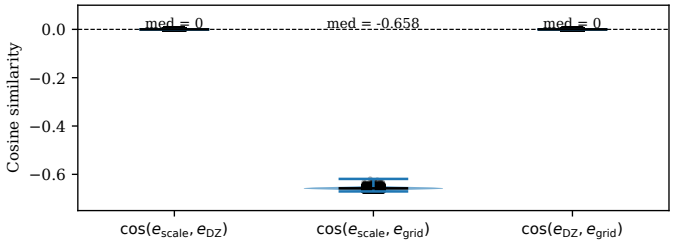


Figure 1: Pairwise error component cosine similarities across 18,624 weight tensors (Qwen3-30B-A3B-Base). DZ is exactly orthogonal to both scale and grid (point masses at 0); scale and grid are anti-correlated with $\cos \approx -0.66$ and minimal variance. This visualizes the structural relationships summarized in Table 1.

et al., 2026; Cook et al., 2025), MoE-specific FP4 training on Hopper (Zhang et al., 2026), and oscillation-reduced MXFP4 training (TetraJet) (Chen et al., 2025). Quantization-aware distillation recovers post-training accuracy under NVFP4 (Xin et al., 2026), and Outlier-Safe Pre-Training (Park et al., 2025) attacks the same problem from the pre-training side by preventing outliers from forming in the first place. Hao et al. (2025) provide a recent survey covering both fixed-point and floating-point low-precision training. Our contribution is orthogonal to these training-time format choices: we target RL post-training of an existing BF16 checkpoint and explain how the static MXFP4 error structure maps onto specific RL failure modes.

3 MXFP4 QUANTIZATION ERROR DECOMPOSITION

3.1 PRELIMINARIES AND DECOMPOSITION

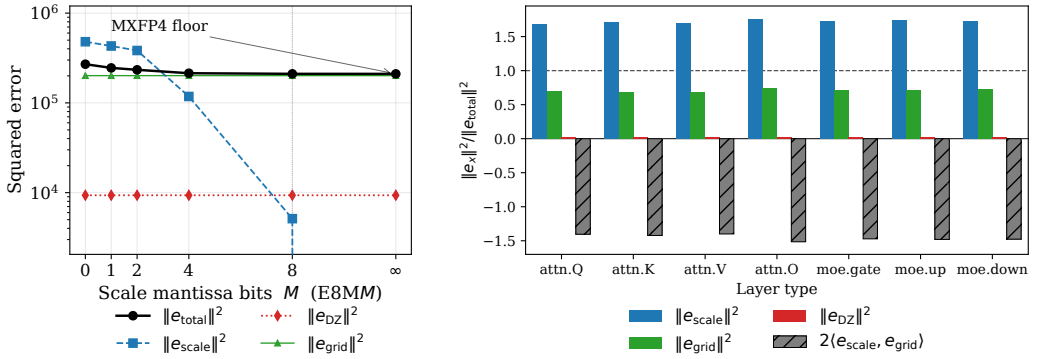
The MXFP4 format (OCP MX standard) quantizes a tensor $\boldsymbol{x} \in \mathbb{R}^n$ block-wise with block size B=32. Each block $\boldsymbol{x}_b = (x_{b,1}, \ldots, x_{b,B}) \in \mathbb{R}^B$ is quantized in two steps: **(i)** a shared E8M0 scale $s_b = 2^{\lceil \log_2(\max_i |x_{b,i}|/q_{\max}) \rceil}$ and **(ii)** per-element rounding to the E2M1 grid $\mathcal{G} = \{0, \pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 3, \pm 4, \pm 6\}$ ($q_{\max}$=6), giving a *dequantized* value $Q(x_{b,i}) = s_b \cdot \arg\min_{g \in \mathcal{G}} |x_{b,i}/s_b - g|$.

We introduce the *ideal-scale quantizer* Q^* that uses the unquantized scale $s_b^* = \max_i |x_{b,i}|/q_{\max}$ instead of the E8M0 s_b. This separates scale quantization from element-level quantization. The *deadzone* $\mathcal{D}_b$ is the set of elements satisfying $|x_{b,i}| < m_b/24$ (where $m_b = \max_j |x_{b,j}|$), which are rounded to zero; empirically 9% of values per block.

Definition 3.1 (Error components)**.** The total quantization error $e_{b,i} = Q(x_{b,i}) - x_{b,i}$ decomposes as

$$e_{b,i} = \underbrace{Q(x_{b,i}) - Q^*(x_{b,i})}_{e_{b,i}^{\text{scale}}} + \underbrace{[Q^*(x_{b,i}) - x_{b,i}]\, \mathbf{1}_{\mathcal{D}_b}(i)}_{e_{b,i}^{\text{DZ}}} + \underbrace{[Q^*(x_{b,i}) - x_{b,i}]\, \mathbf{1}_{\mathcal{D}_b^c}(i)}_{e_{b,i}^{\text{grid}}}, \tag{1}$$

where $\mathbf{1}_{\mathcal{D}_b}(i) = \mathbf{1}[|x_{b,i}/s_b^*| < q_{\min}/2]$ is the deadzone indicator ($q_{\min}$=0.5). Since $\mathbf{1}_{\mathcal{D}_b} + \mathbf{1}_{\mathcal{D}_b^c} = 1$, the decomposition is exact: $\boldsymbol{e} = \boldsymbol{e}^{\text{scale}} + \boldsymbol{e}^{\text{DZ}} + \boldsymbol{e}^{\text{grid}}$.



(a) Scale mantissa precision sweep.

(b) Per-layer-type error decomposition (Qwen3-30B-A3B-Base).

Figure 2: Error decomposition analysis. (a) Improving scale precision drives total error to the irreducible floor $\left\|\boldsymbol{e}^{\text{grid}}\right\|^2 + \left\|\boldsymbol{e}^{\text{DZ}}\right\|^2$ without affecting grid error. (b) All layer types share nearly same decomposition: $\left\|\boldsymbol{e}^{\text{scale}}\right\|^2 \approx 1.72\,\|\boldsymbol{e}\|^2$, $\left\|\boldsymbol{e}^{\text{grid}}\right\|^2 \approx 0.71\,\|\boldsymbol{e}\|^2$, $\left\|\boldsymbol{e}^{\text{DZ}}\right\|^2 \approx 0.02\,\|\boldsymbol{e}\|^2$.

3.2 Interactions among three error components

Lemma 3.2 (Exact orthogonality: DZ $\perp$ Scale and DZ $\perp$ Grid)**.** $\langle \boldsymbol{e}^{\text{DZ}}, \boldsymbol{e}^{\text{scale}}\rangle = \langle \boldsymbol{e}^{\text{DZ}}, \boldsymbol{e}^{\text{grid}}\rangle = 0$ *(pointwise, no assumptions needed).*

Proof. $\boldsymbol{e}^{\text{DZ}}$ and $\boldsymbol{e}^{\text{grid}}$ have disjoint support by construction ($\mathbf{1}_{\mathcal{D}_b} \cdot \mathbf{1}_{\mathcal{D}_b^c} = 0$). For $\boldsymbol{e}^{\text{DZ}} \perp \boldsymbol{e}^{\text{scale}}$: on deadzone elements, $|x_i/s_b^*| < q_{\min}/2$, so $Q^*(x_i) = 0$. Since $s_b \geq s_b^*$ (ceiling rounding), $|x_i/s_b| \leq |x_i/s_b^*| < q_{\min}/2$, so $Q(x_i) = 0$ as well. Therefore $e_i^{\text{scale}} = Q(x_i) - Q^*(x_i) = 0$ on every deadzone element. □

Expanding $\|\boldsymbol{e}\|^2 = \left\|\boldsymbol{e}^{\text{scale}} + \boldsymbol{e}^{\text{DZ}} + \boldsymbol{e}^{\text{grid}}\right\|^2$ and applying Lemma 3.2 to eliminate two of three cross terms yields the MSE identity:

$$\boxed{\|\boldsymbol{e}\|^2 = \left\|\boldsymbol{e}^{\text{scale}}\right\|^2 + \left\|\boldsymbol{e}^{\text{DZ}}\right\|^2 + \left\|\boldsymbol{e}^{\text{grid}}\right\|^2 + 2\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}}\rangle}, \tag{2}$$

with exactly one surviving cross term.

Remark 3.3 (Anti-correlation between scale and grid). We observe $\cos(\boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}}) \approx -0.66$ consistently across 19,100 weight tensors in three tested architectures (Table 1 and Figure 1). The negative sign is expected: since $s_b \geq s_b^*$ (ceiling rounding), the quantized scale overshoots the ideal scale, meaning Q uses a coarser grid than Q^*. Where Q^* rounds a value upward (positive e_i^{grid}), the coarser Q grid tends to round to the same or a lower point (negative e_i^{scale}), creating systematic sign opposition. The magnitude ≈0.66 is an empirical observation that we do not derive in closed form or claim as universal beyond the tested architectures. Note that component norms sum to $\approx 2.46\,\|\boldsymbol{e}\|^2$, exceeding the total; the cross term $2\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}}\rangle \approx -1.46\,\|\boldsymbol{e}\|^2$ accounts for the difference. This has a practical consequence: MBS reduces both $\left\|\boldsymbol{e}^{\text{scale}}\right\|^2$ (172% of total) *and* the negative cross term, reducing total MSE to $\left\|\boldsymbol{e}^{\text{grid}}\right\|^2 + \left\|\boldsymbol{e}^{\text{DZ}}\right\|^2 \approx 0.73\,\|\boldsymbol{e}\|^2$, a 27% reduction.

3.3 Grid noise is invariant to scale precision

The grid error e_i^{grid} on non-deadzone elements depends only on the weight values x_i and the ideal scale s_b^*, *not on the actual E8M0 scale* s_b. Changing the scale precision (e.g., from E8M0 to E8M8 via MBS) modifies $\boldsymbol{e}^{\text{scale}}$ but leaves $\boldsymbol{e}^{\text{grid}}$ completely unchanged. As scale precision improves:

$$\left\|\boldsymbol{e}^{\text{scale}}\right\| \to 0, \quad \langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}}\rangle \to 0, \quad \left\|\boldsymbol{e}^{\text{total}}\right\|^2 \to \underbrace{\left\|\boldsymbol{e}^{\text{grid}}\right\|^2 + \left\|\boldsymbol{e}^{\text{DZ}}\right\|^2}_{\text{irreducible floor}}. \tag{3}$$

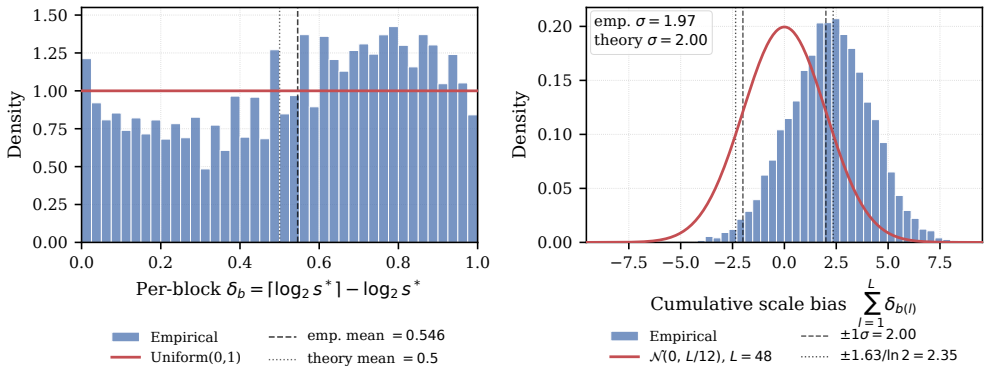


Figure 3: Scale bias from E8M0 scale rounding (Qwen3-30B-A3B-Base, L=48). (a) Per-block rounding residual δ_b is approximately Uniform(0,1) with mean 0.546. (b) Cumulative scale bias $\sum_l \delta_l$ across layers matches the theoretical standard deviation with empirical $\sigma = 1.97 \approx \sigma_{\text{theory}} = 2.0$, with empirical mean near 2.0 due to imperfect Uniform distribution, validating the gradient inaccuracy in Section 4.1.

This floor is the structural limit of MXFP4: to go below it, the E2M1 grid itself must change. Figure 2a visualizes this convergence.

4 Impact on RL Training

Each error component affects all three RL training pathways (policy gradient accuracy, rollout quality, and exploration–exploitation balance), but the mathematical structure of each creates a *dominant* effect on one pathway. We analyze these below; the full 3×3 analysis is in Appendix D.

4.1 Scale Bias: Dominant Effect on Policy Gradients

Scale bias is *multiplicative*: each layer's output is scaled by $\gamma_b = s_b / s_b^*$. In the forward pass, LayerNorm resets activation magnitudes at every layer, preventing accumulation, so the per-layer forward-pass error is $O(\gamma_b - 1) \approx O(0.44)$, calculation shown in text after Eq. (4) . In the backward pass, however, the straight-through estimator (STE) chain rule multiplies scale factors across all L layers without any normalization. Let $\hat{\nabla}$ denote the gradient computed under quantization and ∇_{true} the full-precision gradient. Their log magnitude ratio accumulates as:

$$\log \frac{\left\|\hat{\nabla}\right\|}{\|\nabla_{\text{true}}\|} \approx \sum_{l=1}^{L} \delta_{b(l)}\,, \quad \delta_b \sim \text{Uniform}[0,\,1)\,, \tag{4}$$

where $\delta_b = \lceil \log_2 s_b^* \rceil - \log_2 s_b^*$ is the E8M0 ceiling-rounding error in log-space, approximately uniform by a Benford's-law argument (Appendix A, Step 1). Under ceiling rounding, $\gamma_b = 2^{\delta_b}$ with $\delta_b \sim \text{Uniform}[0, 1)$, giving $\text{mean}(\gamma - 1) = 1/\ln 2 - 1 \approx 0.44$ and $\text{RMSE}(\gamma - 1) \approx 0.57$. By the CLT, the centred sum $\sum_l (\delta_{b(l)} - \frac{1}{2}) \sim \mathcal{N}(0, L/12)$. For Qwen2.5-3B with L=36 layers, $\text{std} = \sqrt{3} \approx 1.73$, so the gradient magnitude ratio $\left\|\hat{\nabla}\right\| / \|\nabla_{\text{true}}\|$ ranges from $\sim$0.18$\times$ to $\sim$5.6$\times$ within one standard deviation. This **exponential amplification in the backward pass**, absent in the forward pass, is what makes scale bias primarily a gradient problem. For the 48-layer MoE model (Qwen3-30B-A3B-Base), $\text{std}_{\text{theory}} = \sqrt{48/12} = 2.0$; we measure $\text{std}_{\text{emp}} = 1.97$ (Figure 3). MBS corrects this by adding an 8-bit mantissa to the block scale, reducing $\text{Var}(\gamma)$ by $\sim(256)^{-2}$.

Causal confirmation. This prediction is directly testable. Degrading *only* the backward pass to E8M0 while keeping the actor forward and rollout at standard MXFP4 drives GSM8K training to 37.7%; replacing just the backward scale with the ideal continuous Q^* (same E2M1 grid, forward and rollout untouched) recovers it to 72.8% (Llama-3.1-8B, single seed). This +35 pp swing, obtained without altering a single forward activation, isolates backward-pass scale bias as the dominant *reducible*

cause of the gradient pathology and confirms the amplification predicted above (per-step trajectory in Appendix H).

Why not the other pathways? Scale bias does perturb rollout logits and exploration at $O(\gamma-1) \approx 49\%$ per layer, but LayerNorm re-normalizes inputs between layers, preventing multiplicative compounding in the forward pass.

4.2 Deadzone: dominant effect on rollout distribution

Deadzone truncation is *pruning*: it affects RL training by deterministically setting weights to zeros in the forward pass, changing the rollout distribution. On Qwen3-30B-A3B-Base model, 9.0% of elements fall in this ideal deadzone and contribute 2.2% of $\|e_{\text{total}}\|^2$.

Why not gradients? The STE backward pass **ignores the deadzone**: it passes gradients through as if the full-precision value were present. So deadzone damage is fully visible in the forward pass (degraded rollouts) but largely invisible in the backward pass (intact gradients). This asymmetry makes deadzone primarily a rollout quality problem.

Why not exploration? Deadzone is **systematic and contractive** ($e_i^{\text{DZ}} = -x_i$, always toward zero), not stochastic. It removes capacity rather than adding randomness, producing blander outputs (lower reward mean) rather than more diverse ones. In our dense experiments, OF recovers deadzone values (DZ rate: 9% $\rightarrow$ 2%), yielding $+17.5$ pp.

4.3 Grid noise: dominant effect on exploration

Grid noise is the only component that is **dense and approximately zero-mean**. A zero-mean perturbation to logits does not shift the mode of the output distribution; it *widens* it, acting approximately as temperature scaling. Under additive noise $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \boldsymbol{I})$ on logits, the effective temperature is

$$T_{\text{eff}} \approx \sqrt{1 + 2\sigma_\eta^2 / \operatorname{Var}(\Delta\ell)} > 1\,, \tag{5}$$

where $\sigma_\eta^2 \propto \sum_l \sigma_{\delta,l}^2(\text{grid})$ accumulates over layers. The noise level is *constant* throughout training, providing no mechanism for annealing. Grid noise adds variance but no systematic bias to gradients (far milder than the $5.6\times$ scale-induced fluctuation), and its effect on reward quality averages out over full response sequences. Equation (5) treats the accumulated grid noise as approximately Gaussian logit noise; the true grid noise is bounded and structured, so this holds only in the aggregate-over-layers limit (Appendix D.4).

5 Mechanism-targeted interventions

The decomposition of Section 3 is a *diagnosis*; to test it we pair each component with one prior, off-the-shelf correction and verify that removing that component recovers its predicted RL pathway. Each of the three corrections below is **prior work, used here purely as a diagnostic probe**: we state each compactly (which component it targets and its defining operation) and defer full formulations and hyperparameters to Appendix I.

MBS (targets scale bias) (Chhugani et al., 2026). The E8M0 block scale is a pure power of two ($\gamma_b = s_b/s_b^*$ has $\sim$54% error). Macro Block Scaling restores an 8-bit mantissa at a coarser 128-element macro-block granularity, applied as a prescale–quantize–postscale wrap $\hat{x}_i = Q\big((1+m)x_i\big)/(1+m)$ with $m \in [0, 1)$ (adding scale mantissa bits is prior art; NVFP4 (NVIDIA, 2025) uses an FP8 micro-scale). This shrinks $\operatorname{Var}(\gamma)$ by $\sim(256)^{-2}$, so $\|\boldsymbol{e}^{\text{scale}}\| \rightarrow 0$; by grid-invariance (Section 3.3) $\boldsymbol{e}^{\text{grid}}$ is untouched, and total MSE converges to the irreducible floor $\|\boldsymbol{e}^{\text{grid}}\|^2 + \|\boldsymbol{e}^{\text{DZ}}\|^2$ at < 0.1 bit/element overhead.

OF (targets deadzone) (Zhang et al., 2025). Deadzone truncation is *information loss* no rescaling can undo. Outlier Fallback uses two-pass residual quantization (the same residual-fallback structure appears in ARCQuant (Meng et al., 2026) and MosaicQuant (Hu et al., 2026)):

$$\hat{\boldsymbol{x}}_1 = Q(\boldsymbol{x}), \quad \hat{\boldsymbol{x}}_2 = Q(\boldsymbol{x} - \hat{\boldsymbol{x}}_1), \quad \hat{\boldsymbol{x}}_{\text{OF}} = \hat{\boldsymbol{x}}_1 + \alpha\, \hat{\boldsymbol{x}}_2, \tag{6}$$

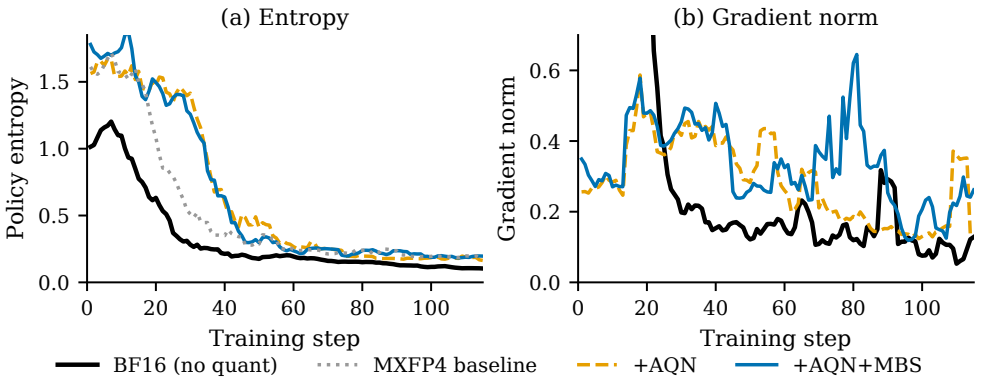


Figure 4: Training dynamics (MoE, GSM8K). (a) Policy entropy; (b) gradient norm. Under quantization, +AQN and +AQN+MBS sustain higher entropy than BF16 (extra noise drives exploration), and +AQN+MBS maintains the gradient norm above BF16 in late training, indicating continued learning.

where Q is the standard MXFP4 quantizer. Pass 1's outliers set the block scale and send small values to the deadzone; Pass 2's small dynamic range makes them representable. We use $\alpha = 0.5$: since $\hat{\boldsymbol{x}}_2$ is itself lossy, half-strength addition beats $\alpha = 1.0$ on GSM8K by $\sim$1 pp and minimizes rollout–training drift (Appendix D.2).

AQN (targets grid noise / entropy) (Huang et al., 2026). After MBS and OF, the residual is dominated by the grid floor. Adaptive Quantization Noise is a complementary RL *recipe* (not a static correction), in the lineage of NoisyNet (Fortunato et al., 2018) and gradient-noise injection (Neelakantan et al., 2015): before each rollout it perturbs LayerNorm weights with Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ whose σ decays exponentially from 1% to 0.1% over 10 stages, preventing the premature entropy collapse that constant grid noise induces (Section 4.3).

QDQ emulation. Experiments use simulated quantize/dequantize in PyTorch, the standard methodology for MXFP4 training research (Tseng et al., 2025; Chen et al., 2025), as native W4A4 tensor cores are not yet widely accessible. QDQ is numerically faithful (identical E2M1 grid, E8M0 block scale, FP32 accumulation). We emphasize that *no speedup is measured in this paper*: the $4\times$ figure is a property of the format spec, and native-accelerator latency is future work that does not affect the accuracy results here.

6 Experiments

6.1 Setup

Models and task. We evaluate on two architectures: Qwen2.5-3B (dense, 36 layers) and Qwen3-30B-A3B-Base (MoE, 48 layers, 3B active parameters). The task is GSM8K (Cobbe et al., 2021) math reasoning with verifiable rewards.

Training. All runs use GRPO (Shao et al., 2024) with batch size 64 and max response length 1024; the scaffolding builds on verl alongside related RL-LLM frameworks (Yu et al., 2025; Fu et al., 2025). For MoE: Megatron parallelism, 2 samples per prompt. For dense: FSDP2, 4 samples per prompt. Full hyperparameter list in Appendix I.

Quantization. MXFP4 W4A4 via QDQ emulation: both weights and activations quantized to E2M1 elements with E8M0 block scale (block size 32). Truncated Importance Sampling (TIS) Yao et al. (2025) clips the importance ratio ρ_i from above at $\rho_{\max}$ (i.e., $\min(\rho_i, \rho_{\max})$): we use TIS $= 2$ for MoE and TIS $= 5$ for dense. Controlled single-knob ablations (varying only TIS) confirm TIS $= 2$ is stable, not collapse-inducing: it changes final accuracy by only $\approx$2.35 pp, far smaller than the dense/MoE OF asymmetry, so the per-model TIS choice is not a confound (Appendix H).

Table 2: MoE results (Qwen3-30B-A3B-Base, GSM8K, W4A4). BF16 baseline: 91.51%.

Condition	AQN	MBS	OF	GSM8K (%)	Gap
MXFP4 baseline				88.8	−2.7
+MBS		✓		90.1	−1.4
+OF			✓	90.3	−1.2
+AQN(1%)	✓			89.2	−2.3
+AQN+MBS	✓	✓		90.5	−1.0
+MBS+OF		✓	✓	91.1	−0.4
+AQN+MBS+OF	✓	✓	✓	**92.49**	**+1.0**

Table 3: Dense results (Qwen2.5-3B, GSM8K, MXFP4 W4A4). BF16 baseline: 82.0%.

Condition	MBS	AQN	OF	GSM8K (%)	Gap
MXFP4 baseline				60.1	−21.9
+MBS	✓			75.2	−6.8
+OF			✓	77.6	−4.4
+MBS+OF	✓		✓	80.8	−1.2
+MBS+AQN(1%)+OF	✓	✓	✓	**81.3**	−0.7

6.2 Main results

Tables 2–3 present the main results (visualized in Figure 11, Appendix H). On MoE, AQN+MBS+OF ($\alpha = 0.5$) peaks at 92.49% (+1.0 pp over BF16); on dense, the naïve −21.9 pp gap recovers to within 0.7 pp of BF16 (81.3% vs. 82.0%).

Correction composition (per setting). Additivity is setting-dependent. On MoE the corrections are *near-additive* (MBS +1.3, OF +1.5, AQN +0.4 pp vs. +2.3 combined). On dense they are *strongly sub-additive*: MBS (+15.1) and OF (+17.5 pp) each exceed the +21.9 pp gap because both recover forward-pass capacity and overlap. Either way OF is the critical dense correction, as the decomposition predicts: dense models lack expert routing to mask deadzone pruning, so OF (+17.5 pp dense vs. +1.5 pp MoE) is essential.

Cross-family generalization, baselines, and per-component attribution (Llama-3.1-8B). Run over three seeds on a different family, the full stack reaches GSM8K 80.9 ± 0.7 / 83.7 ± 0.2 and transfers to MATH500 at 35.2 ± 2.4 (the tight $\sigma \leq 0.7$ pp spread indicates stable recovery, not seed noise), and it outperforms every prior-method baseline we ran (Table 4): block-rotation quantization (BRQ) (Shao et al., 2026), MBS combined with Overflow-Aware Scaling (OAS), both from Chhugani et al. (2026), FP4-All-the-Way (Chmiel et al., 2025), and QaRL (Gu et al., 2026) (weight-only W4A16 by design; run here as a W4A4 adaptation). Beyond ranking accuracy, the decomposition measures *which* error term each method reduces (Table 5, 224 Llama projections): each prior method moves a different term, while only the full stack drives all three toward the floor. Backward-pathway, TIS, and cross-task ablations are in Appendix H.

AQN σ sensitivity. On dense MXFP4 with MBS+OF, $\sigma_{\text{start}} = 0.1\%$ is too weak (71.2%), 2% too strong (80.1%), and 1% optimal (81.3%), matching the $\sim 1\%$ per-element grid-noise scale.

Training dynamics. Figure 4 shows +AQN and +AQN+MBS sustain higher policy entropy than BF16 throughout training and keep the gradient norm elevated late in training, both signatures of continued exploration and learning on the quantization noise floor.

Table 4: Prior-method baselines on Llama-3.1-8B (MXFP4 W4A4, GSM8K training; MATH500 as transfer eval). BF16 (upper bound), naïve MXFP4 (lower bound), and full stack are three-seed means ($\pm$ sample std, shown as subscripts). The single-run baselines (BRQ, MBS+OAS, FP4-All-the-Way, QaRL) are reported without variance: each trails the full stack by at least 4 pp on GSM8K final, a margin far larger than the three-seed spread, so seed noise cannot change their ranking. †QaRL targets weight-only W4A16; run here as a W4A4 adaptation.

Method	GSM8K final (%)	GSM8K peak (%)	MATH500 (%)
BF16 (upper bound)	$84.1_{\pm 0.9}$	$86.0_{\pm 0.6}$	$42.1_{\pm 0.4}$
Naïve MXFP4 (no correction)	$46.6_{\pm 11.7}$	$46.6_{\pm 11.7}$	$23.5_{\pm 2.1}$
Full stack	$\mathbf{80.9}_{\pm 0.7}$	$\mathbf{83.7}_{\pm 0.2}$	$\mathbf{35.2}_{\pm 2.4}$
BRQ (block rotation)	76.9	78.9	25.2
MBS+OAS	72.6	77.5	24.8
FP4-All-the-Way-style	58.5	58.5	24.0
QaRL-style W4A4†	22.9	31.4	11.6

Table 5: Forward-pass per-component error change vs. naïve MXFP4 (Llama-3.1-8B, 224 projections). Negative = reduction.

Method	$\Delta \left\|\boldsymbol{e}^{\mathrm{scale}}\right\|^2$	$\Delta \left\|\boldsymbol{e}^{\mathrm{DZ}}\right\|^2$	$\Delta \left\|\boldsymbol{e}^{\mathrm{grid}}\right\|^2$	Forward-GEMM err. $\downarrow$
MBS	-54.4%	0.0%	0.0%	44.2%
BRQ (block rotation)	-21.7%	-88.3%	$+10.3\%$	35.6%
FP4-All-the-Way (fwd)	0.0%	0.0%	0.0%	0.0%
MBS + OF (full stack)	-99.3%	-98.8%	-99.0%	73.2%

7 Conclusion

We presented a three-way decomposition of MXFP4 quantization error into scale bias, deadzone truncation, and grid noise, anchored by two formal results: exact orthogonality of the deadzone to both others (Lemma 3.2) and invariance of grid noise to scale precision (Section 3.3), which establishes an irreducible floor intrinsic to the E2M1 grid. The component ratios and the $\cos \approx -0.66$ anti-correlation hold across 19,100 tensors in all three tested models (Table 1), a format-driven pattern rather than a universal constant. Using MBS, OF, and AQN as mechanism-targeted probes (each removing one component and recovering its predicted RL pathway) validates the diagnosis: MBS+AQN+OF recovers BF16 within 0.7 pp on dense and reaches parity on MoE. Practically, the decomposition doubles as a bit-allocation rule: because grid noise fixes a floor that scale refinement cannot cross (Section 3.3), spending precision on the block scale beyond the mantissa correction of MBS yields no further error reduction, so the marginal bit is better invested in deadzone recovery (OF) or, since the residual grid floor is irreducible in-format, in exploration control (AQN) that helps RL training tolerate it.

Limitations and future work. GSM8K uses short-to-medium responses; long responses are harder to recover due to autoregressive trajectory divergence (Appendix G), and coding/safety benchmarks remain to be validated. All results use numerically-faithful QDQ emulation (no speedup is measured), so native-accelerator throughput is future work; Qwen rows are single-seed (Llama full stack is three-seed), as is the Q^* backward-only control. Our interventions are existing corrections rather than surgical single-component knobs: MBS also shrinks the negative scale–grid cross term, and AQN perturbs LayerNorm weights rather than injecting grid noise directly, so the grid-noise-to-exploration link is established indirectly. Only the backward-only ideal-scale (Q^*) control (Section 4.1) isolates a single pathway cleanly; matching synthetic isolations that directly toggle the deadzone set and inject calibrated grid-magnitude noise would close the deadzone and grid legs with the same rigor and are the natural next step. The deadzone-to-rollout effect is itself architecture-dependent (OF adds 17.5 pp on dense but 1.5 pp on MoE, where expert routing compensates for pruned weights), so we scope that mapping to the settings tested rather than claiming it universally. Combining MBS/OF with upstream techniques (Tseng et al., 2025) and extending to activation quantization are natural next steps.

THE USE OF LARGE LANGUAGE MODELS

Large language models were used as general-purpose writing and coding assistants during this work. Specifically: (i) polishing prose and improving the clarity of the manuscript; (ii) assisting with plotting and data-aggregation scripts for the figures and tables; and (iii) assisting with routine boilerplate in the training and evaluation scaffolding. LLMs were *not* used to generate research ideas, theoretical results, or experimental findings: the error decomposition, all proofs (Lemma 3.2 and Appendix A), the failure-mode analysis, and all reported numbers were produced and verified by the authors. The authors take full responsibility for the entire content of the paper, including any text or code produced with LLM assistance.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27(3):446–472, 1948.

Roberto L. Castro, Andrei Panferov, Soroush Tabesh, Oliver Sieberling, Jiale Chen, Mahdi Nikdan, Saleh Ashkboos, and Dan Alistarh. Quartet: Native FP4 training can be optimal for large language models. *arXiv preprint arXiv:2505.14669*, 2025.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yuxiang Chen, Haocheng Xi, Jun Zhu, and Jianfei Chen. Oscillation-reduced MXFP4 training for vision transformers. In *International Conference on Machine Learning (ICML)*, 2025.

Jatin Chhugani, Geonhwa Jeong, Bor-Yiing Su, Yunjie Pan, Hanmei Yang, Aayush Ankit, Jiecao Yu, Summer Deng, Yunqing Chen, Nadathur Satish, and Changkyu Kim. Unveiling the potential of quantization with MXFP4: Strategies for quantization error reduction. *arXiv preprint arXiv:2603.08713*, 2026.

Brian Chmiel, Maxim Fishman, Ron Banner, and Daniel Soudry. FP4 all the way: Fully quantized training of LLMs. *arXiv preprint arXiv:2505.19115*, 2025.

Euntae Choi, Sumin Song, Woosang Lim, and Sungjoo Yoo. Grouped sequency-arranged rotation: Optimizing rotation transformation for quantization for free. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 165–172, 2025. doi: 10.18653/v1/2025.acl-srw.10.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jack Cook, Junxian Guo, Guangxuan Xiao, Yujun Lin, and Song Han. Four over six: More accurate NVFP4 quantization with adaptive block scaling. *arXiv preprint arXiv:2512.02010*, 2025.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. In *International Conference on Learning Representations*, 2024.

Vage Egiazarian, Roberto L. Castro, Denis Kuznedelev, Andrei Panferov, Eldar Kurtic, Shubhra Pandit, Alexandre Noll Marques, Mark Kurtz, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Bridging the gap between promise and performance for microscaling FP4 quantization. In *International Conference on Learning Representations*, 2026.

Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. Scaling FP8 training to trillion-token LLMs. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E1EHO0imOb.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Rémi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rywHCPkAW.

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*, 2023.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. AReaL: A large-scale asynchronous reinforcement learning system for language reasoning. *arXiv preprint arXiv:2505.24298*, 2025.

Hao Gu, Hao Wang, Jiacheng Liu, Lujun Li, Qiyuan Zhu, Bei Liu, Binxing Xu, Lei Wang, Xintong Yang, Sida Lin, Sirui Han, and Yike Guo. QaRL: Rollout-aligned quantization-aware RL for fast and stable training under training–inference mismatch. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, 2026.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pp. 1737–1746, 2015. URL http://proceedings.mlr.press/v37/gupta15.html.

Zhiwei Hao, Jianyuan Guo, Li Shen, Yong Luo, Han Hu, Guoxia Wang, Dianhai Yu, Yonggang Wen, and Dacheng Tao. Low-precision training of large language models: Methods, challenges, and opportunities. *arXiv preprint arXiv:2505.01043*, 2025.

Alejandro Hernández-Cano, Dhia Garbaya, Imanol Schlag, and Martin Jaggi. Towards fully FP8 GEMM LLM training at scale. *arXiv preprint arXiv:2505.20524*, 2025.

Yangjia Hu, Haodong Wang, Zicong Hong, Qianli Liu, Quanxin Shou, Jian Lin, Song Guo, Xiaowei Shen, Xiangjun Huang, Dian Wang, and Jian Yang. MosaicQuant: Inlier–outlier disaggregation for unified 4-bit LLM quantization. *arXiv preprint arXiv:2606.15652*, 2026.

Wei Huang, Yi Ge, Shuai Yang, Yicheng Xiao, Huizi Mao, Yujun Lin, Hanrong Ye, Sifei Liu, Ka Chun Cheung, Hongxu Yin, Yao Lu, Xiaojuan Qi, Song Han, and Yukang Chen. QeRL: Beyond efficiency—quantization-enhanced reinforcement learning for LLMs. In *International Conference on Learning Representations*, 2026.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yesheng Liang, Haisheng Chen, Song Han, and Zhijian Liu. ParoQuant: Pairwise rotation quantization for efficient reasoning LLM inference. *arXiv preprint arXiv:2511.10645*, 2025.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 467–484, 2024.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, 2025.

Haoqian Meng, Yilun Luo, Yafei Zhao, Wenyuan Liu, Peng Zhang, and Xindian Ma. ARCQuant: Boosting NVFP4 quantization with augmented residual channels for LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8609–8623, San Diego, California, United States, 2026. Association for Computational Linguistics.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi, Stuart F. Oberman, Mohammad Shoeybi, Michael Y. Siu, and Hao Wu. FP8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022.

Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

NVIDIA. Pretraining large language models with NVFP4. *arXiv preprint arXiv:2509.25149*, 2025.

Andrei Panferov, Erik Schultheis, Soroush Tabesh, and Dan Alistarh. Quartet II: Accurate LLM pre-training in NVFP4 by improved unbiased gradient estimation. *arXiv preprint arXiv:2601.22813*, 2026.

Jungwoo Park, Taewhoo Lee, Chanwoong Yoon, Hyeon Hwang, and Jaewoo Kang. Outlier-safe pre-training for robust 4-bit quantization of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12582–12600, 2025.

Bita Darvish Rouhani, Ritchie Zhao, Ankit More, Mathew Hall, Alireza Khodamoradi, Summer Deng, Dhruv Choudhary, Marius Cornea, Eric Dellinger, Kristof Denolf, et al. OCP microscaling formats (MX) specification. *Open Compute Project*, 2023.

Yuantian Shao, Peisong Wang, Yuanteng Chen, Chang Xu, Zhihui Wei, and Jian Cheng. Block rotation is all you need for MXFP4 quantization. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*. PMLR, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. QTIP: Quantization with trellises and incoherence processing. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Albert Tseng, Tao Yu, and Youngsuk Park. Training llms with mxfp4, 2025. URL `https://arxiv.org/abs/2502.20586`.

Bernard Widrow and István Kollár. *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge University Press, 2008. ISBN 978-0-521-88671-0.

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*, 2023.

Meng Xin, Sweta Priyadarshi, Jingyu Xin, Bilal Kartal, Aditya Vavre, et al. Quantization-aware distillation for NVFP4 inference accuracy recovery. *arXiv preprint arXiv:2601.20088*, 2026.

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient rl framework secretly brings you off-policy rl training, August 2025. URL `https://fengyao.notion.site/off-policy-rl`.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27168–27183, 2022.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Pengle Zhang, Jia Wei, Jintao Zhang, Jun Zhu, and Jianfei Chen. Accurate INT8 training through dynamic block-level fallback. *arXiv preprint arXiv:2503.08040*, 2025.

Wuyue Zhang, Chongdong Huang, Chunbo You, Cheng Gu, Fengjuan Wang, and Mou Sun. Practical FP4 training for large-scale MoE models on hopper GPUs. *arXiv preprint arXiv:2603.02731*, 2026.

A Analysis of the cross term $\langle \boldsymbol{e}^{\mathrm{scale}}, \boldsymbol{e}^{\mathrm{grid}} \rangle$

Note on scope. The main body establishes the MSE identity (Eq. 2) and reports the empirical anti-correlation $\cos(\boldsymbol{e}^{\mathrm{scale}}, \boldsymbol{e}^{\mathrm{grid}}) \approx -0.66$ (Table 1). This appendix analyzes the cross term under the following idealized assumption. Under this assumption the cross term is $\mathcal{O}(B^{-1/2})$, but in practice, the ceiling-based scale selection ($\gamma_b \geq 1$ always) creates a structural anti-correlation absent from the idealized model. The analysis below characterizes the idealized setting for completeness; the empirical measurements in Table 1 and Appendix F characterize the real setting.

Assumption A.1. Within each block the elements $(x_{b,1}, \ldots, x_{b,B})$ are i.i.d. from a continuous, symmetric distribution $\mathcal{P}$ with finite fourth moment. The quantity $\log_2 s_b^*$ is non-lattice.

Define the scale ratio $\gamma_b = s_b / s_b^* = 2^{\delta_b}$, normalized elements $u_{b,i} = q_{\max} \cdot x_{b,i}/m_b$, and let $r(\cdot)$ denote $\mathrm{round}_{\mathrm{E2M1}}$.

Step 1: Characterize γ_b. Under Assumption A.1, $\delta_b = \lceil \log_2 s_b^* \rceil - \log_2 s_b^*$ depends only on the fractional part of $\log_2 m_b$. For ceiling rounding, $\delta_b \in [0, 1)$. By the non-lattice condition, δ_b is approximately $\mathrm{Uniform}[0, 1)$, giving

$$\gamma_b \sim 2^U,\ U \sim \mathrm{Uniform}[0,\ 1)\,, \quad \mathbb{E}[\gamma_b] = \frac{1}{\ln 2} \approx 1.44\,, \quad \mathrm{RMSE}(\gamma_b - 1) \approx 0.57\,. \tag{7}$$

Empirically, we measure $\mathrm{RMSE}(\gamma_b - 1) = 0.57$ on Qwen3-30B-A3B-Base, consistent with the ceiling-rounding model.

Step 2: Conditional independence structure. Write the cross term for block b:

$$\langle \boldsymbol{e}_b^{\mathrm{scale}},\, \boldsymbol{e}_b^{\mathrm{elem}} \rangle = \sum_{i=1}^{B} e_{b,i}^{\mathrm{scale}} \cdot e_{b,i}^{\mathrm{elem}}\,, \tag{8}$$

where $e^{\mathrm{elem}} = e^{\mathrm{DZ}} + e^{\mathrm{grid}}$. Conditioning on m_b, the scale ratio γ_b becomes deterministic:

$$e_{b,i}^{\mathrm{scale}} = s_b \cdot r(x_{b,i}/s_b) - s_b^* \cdot r(x_{b,i}/s_b^*) = h(\gamma_b,\, u_{b,i})\,, \tag{9}$$

$$e_{b,i}^{\mathrm{elem}} = s_b^* [r(u_{b,i}) - u_{b,i}] = s_b^*\, g(u_{b,i})\,, \tag{10}$$

with $h(\gamma_b, u) = 0$ when $\gamma_b = 1$.

Step 3: Conditional distribution of non-max elements. Since the B elements are i.i.d. from $\mathcal{P}$, conditioning on the block maximum $m_b = \max_j |x_{b,j}|$ only tells us that each non-max element satisfies $|x_{b,i}| \le m_b$. By Bayes' rule, the conditional distribution of each non-max element is simply $\mathcal{P}$ truncated to $[-m_b, m_b]$:

$$P(x_{b,i} \in A \mid m_b) = \frac{P(x_{b,i} \in A \cap [-m_b, m_b])}{F(m_b) - F(-m_b)}\,, \tag{11}$$

and the $B-1$ non-max elements remain mutually independent (conditioning on the max restricts their range but does not introduce dependence). For symmetric $\mathcal{P}$ with smooth density f, the conditional expectation of any function of the normalized element $u_{b,i} = q_{\max} \cdot x_{b,i}/m_b$ is:

$$\mathbb{E}[g(u_{b,i}) \mid m_b] = \int_{-q_{\max}}^{q_{\max}} g(u)\, \frac{f(u \cdot m_b/q_{\max})}{F(m_b) - F(-m_b)} \cdot \frac{m_b}{q_{\max}}\, du\,. \tag{12}$$

For γ_b near 1, Taylor expansion gives $h(\gamma_b, u) \approx (\gamma_b - 1) \cdot \phi(u)$, so:

$$\mathbb{E}[h(\gamma_b, u_{b,i}) \cdot g(u_{b,i}) \mid m_b] \approx (\gamma_b - 1)\, \mathbb{E}[\phi(u_{b,i})\, g(u_{b,i}) \mid m_b]\,. \tag{13}$$

Step 4: CLT averaging. Taking expectation over m_b:

$$\mathbb{E}[e_{b,i}^{\mathrm{scale}} \cdot e_{b,i}^{\mathrm{elem}}] = \mathbb{E}_{m_b}[(\gamma_b - 1)\, \psi(m_b)]\,, \tag{14}$$

where $\psi(m_b) = \mathbb{E}[\phi(u)\, g(u) \mid m_b]$. The block-summed cross term has B conditionally i.i.d. terms, so by the CLT:

$$\frac{1}{B} \sum_{i=1}^{B} e_{b,i}^{\mathrm{scale}} \cdot e_{b,i}^{\mathrm{elem}} = (\gamma_b - 1)\, \psi(m_b) + \mathcal{O}_P(B^{-1/2})\,. \tag{15}$$

The leading term $\mathbb{E}[(\gamma - 1)\psi]$ is small: ψ involves the product $\phi(u) \cdot g(u)$, which has alternating sign across E2M1 grid intervals. Under the smoothness of f, these cancel to $\mathcal{O}(\mathrm{Var}(\gamma)/\sqrt{B})$. Summing over N blocks and normalizing by $\mathbb{E}[\|\boldsymbol{e}\|^2] = \Theta(NB \cdot \mathrm{Var}(\gamma))$ yields the $\mathcal{O}(B^{-1/2})$ relative bound. ■

B INDIVIDUAL BOUNDS FOR SCALE ⊥ DZ AND SCALE ⊥ GRID

This appendix illustrates the orthogonality $\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{DZ}} \rangle = 0$ (Lemma 3.2) and the anti-correlation $\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}} \rangle < 0$ (Remark 3.3) with a worked numerical example.

Because $\boldsymbol{e}^{\text{DZ}}$ and $\boldsymbol{e}^{\text{grid}}$ have disjoint support (Lemma 3.2), the single surviving cross term partitions element-wise:

$$\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{DZ}} \rangle = \sum_b \sum_{i \in \mathcal{D}_b} e^{\text{scale}}_{b,i} \cdot e^{\text{DZ}}_{b,i} \,, \tag{16}$$

$$\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}} \rangle = \sum_b \sum_{i \notin \mathcal{D}_b} e^{\text{scale}}_{b,i} \cdot e^{\text{grid}}_{b,i} \,. \tag{17}$$

These are sums over **disjoint index sets** within each block. Two consequences follow.

Why $\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{DZ}} \rangle = 0$ exactly. For ceiling-based exponent selection ($s_b = 2^{\lceil \log_2 s^*_b \rceil}$, the standard convention to avoid overflow), $\gamma_b \geq 1$, so the deadzone under s_b is *wider* than under s^*_b. Every element in $\mathcal{D}_b$ maps to zero under both scales, giving $e^{\text{scale}}_i = Q(x_i) - Q^*(x_i) = 0$ for all $i \in \mathcal{D}_b$. The only surviving cross term in the MSE identity (Eq. 2) is therefore $\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}} \rangle$, which acts exclusively on non-deadzone elements.

Numerical example. Consider a block $\boldsymbol{x} = [0.03,\ 0.1,\ 0.3,\ 0.5,\ 0.9,\ 1.5,\ 2.0,\ 4.0]$ with $q_{\max}$=6. The ideal scale is $s^* = 4.0/6 = 0.667$; the E8M0 scale is $s = 2^0 = 1.0$; $\gamma = s/s^* = 1.5$. The deadzone threshold is $s^* \times 0.25 = 0.167$, placing x_1=0.03 and x_2=0.1 in $\mathcal{D}_b$.

i	x_i	DZ?	$Q^*(x_i)$	$Q(x_i)$	e^{scale}_i	e^{DZ}_i	e^{grid}_i
1	0.03	✓	0	0	**0**	−0.03	0
2	0.10	✓	0	0	**0**	−0.10	0
3	0.30		0.333	0.50	+0.167	0	+0.033
4	0.50		0.667	0.50	−0.167	0	+0.167
5	0.90		1.000	1.00	0	0	+0.100
6	1.50		1.333	1.50	+0.167	0	−0.167
7	2.00		2.000	2.00	0	0	0
8	4.00		4.000	4.00	0	0	0

Since $\gamma \geq 1$, both deadzone elements have $e^{\text{scale}}_i = 0$ (bold), giving $\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{DZ}} \rangle = 0$ exactly. The only nonzero cross term is $\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}} \rangle = (+0.167)(+0.033) + (-0.167)(+0.167) + (+0.167)(-0.167) = -0.050$, which has mixed-sign terms that partially cancel within themselves. There is no opposing $\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{DZ}} \rangle$ to mask this value.

The numerical example also illustrates the anti-correlation: at $i = 3$, $e^{\text{scale}} = +0.167$ and $e^{\text{grid}} = +0.033$ (same sign, weak); at $i = 4$, $e^{\text{scale}} = -0.167$ and $e^{\text{grid}} = +0.167$ (opposite sign, strong); at $i = 6$, $e^{\text{scale}} = +0.167$ and $e^{\text{grid}} = -0.167$ (opposite sign). The net inner product is negative (-0.050), consistent with the $\cos \approx -0.66$ observed at scale across all weight tensors (Table 1).

C GEMM OUTPUT ERROR PROPAGATION

For a linear layer $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ with quantized weights $Q(\boldsymbol{W}) = \boldsymbol{W} + \boldsymbol{E}$:

$$\hat{\boldsymbol{y}} = Q(\boldsymbol{W})\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{E}\boldsymbol{x} \,, \quad \boldsymbol{E} = \boldsymbol{E}^{\text{scale}} + \boldsymbol{E}^{\text{DZ}} + \boldsymbol{E}^{\text{grid}} \,. \tag{18}$$

Corollary C.1 (GEMM-level decomposition)**.** *If $\boldsymbol{E}$ and $\boldsymbol{x}$ are independent, then $\mathbb{E}[\|\boldsymbol{E}\boldsymbol{x}\|^2] = \text{tr}(\mathbb{E}[\boldsymbol{E}^\top \boldsymbol{E}]\, \boldsymbol{\Sigma}_x)$. By Eq. 2 and Lemma 3.2, $\mathbb{E}[\boldsymbol{E}^\top \boldsymbol{E}]$ decomposes with one cross term (scale×grid) and exact orthogonality of the DZ component:*

$$\mathbb{E}[\|\boldsymbol{\delta}\|^2] \approx \underbrace{\text{tr}\big(\mathbb{E}[(\boldsymbol{E}^{\text{scale}})^\top \boldsymbol{E}^{\text{scale}}]\, \boldsymbol{\Sigma}_x\big)}_{\sigma^2_\delta(\text{scale})} + \underbrace{\text{tr}\big(\mathbb{E}[(\boldsymbol{E}^{\text{DZ}})^\top \boldsymbol{E}^{\text{DZ}}]\, \boldsymbol{\Sigma}_x\big)}_{\sigma^2_\delta(\text{DZ})} + \underbrace{\text{tr}\big(\mathbb{E}[(\boldsymbol{E}^{\text{grid}})^\top \boldsymbol{E}^{\text{grid}}]\, \boldsymbol{\Sigma}_x\big)}_{\sigma^2_\delta(\text{grid})} \,. \tag{19}$$

The approximation drops the scale×grid cross term $2\,\mathrm{tr}(\mathbb{E}[(\boldsymbol{E}^{\mathrm{scale}})^\top \boldsymbol{E}^{\mathrm{grid}}]\,\boldsymbol{\Sigma}_x)$; under MBS ($\left\|\boldsymbol{E}^{\mathrm{scale}}\right\| \to 0$) this term vanishes and the equality becomes exact.

For W4A4 (both weights and activations quantized), the activation error $\boldsymbol{e}_A$ adds an independent term $\|\boldsymbol{W}\boldsymbol{e}_A\|^2$ (the cross term $\boldsymbol{E}_W\boldsymbol{e}_A$ is second-order). The same three-way decomposition applies to $\boldsymbol{e}_A$.

D Detailed RL pathway analysis

This appendix provides (i) the full 3×3 analysis of how each error component interacts with each RL pathway, and (ii) extended derivations for each dominant effect.

D.1 Full 3×3 interaction analysis

Scale bias on reward signal (secondary). Scale error perturbs each layer's output by a multiplicative factor γ_b, shifting rollout logits. However, LayerNorm resets activation magnitudes at each layer boundary, limiting the per-layer forward-pass distortion to $O(\gamma_b{-}1) \approx O(0.5)$. This is significant but does not compound across layers the way it does in the backward pass.

Scale bias on exploration (secondary). The multiplicative logit shift also changes the effective entropy, but by the same $O(\gamma_b{-}1)$ per-layer amount. Since it has nonzero mean ($\mathbb{E}[\gamma{-}1] \approx 0.44$ under ceiling rounding), the effect is more bias-like than noise-like, contributing to mode shift rather than distribution widening.

Deadzone on gradients (secondary). The STE backward pass treats Q as identity: gradients flow through deadzone elements as if they retained their original values. The deadzone therefore has minimal direct impact on gradient computation, though the degraded forwardpass indirectly changes the rollout distribution.

Deadzone on exploration (secondary). Deadzone is contractive ($e_i^{\mathrm{DZ}} = -x_i$, always toward zero), not stochastic. It reduces the model's representational capacity without adding randomness. The entropy may increase slightly (a less capable model produces more uniform predictions), but this is a side effect of capacity loss, not a controlled exploration mechanism.

Grid noise on reward signal (secondary). Grid noise adds zero-mean variance to each token's logit vector. Over a full response sequence of hundreds of tokens, these i.i.d. perturbations average out in their effect on the overall response quality. The reward is determined by the complete sequence, not individual tokens, so the impact on reward signal quality is attenuated by $O(1/\sqrt{T})$ where T is the sequence length.

Grid noise on gradients (secondary). Zero-mean forward-pass noise produces zero-mean gradient noise. This adds variance to the gradient estimate but no systematic bias. In contrast, scale error introduces a multiplicative bias that grows exponentially with depth. The gradient variance from grid noise is comparable to the inherent stochasticity of mini-batch sampling and does not qualitatively change training dynamics.

D.2 Dominant effect 1: deadzone and reward signal degradation

Effective rank reduction. Define $r_{\mathrm{eff}}(\boldsymbol{W}) = \|\boldsymbol{W}\|_*^2 / \|\boldsymbol{W}\|_F^2$. The deadzone applies a soft-thresholding operator $\mathcal{S}_t(\boldsymbol{W})$ that zeroes elements below threshold $t = m_b/24$. Since small singular values are disproportionately affected by element-wise thresholding (by the Eckart–Young theorem):

$$r_{\mathrm{eff}}(\mathcal{S}_t(\boldsymbol{W})) < r_{\mathrm{eff}}(\boldsymbol{W})\,. \tag{20}$$

Advantage variance increase. Lower effective rank leads to rollout responses with less nuanced reasoning. The reward signal r_i has lower mean and higher variance, making the group-normalized advantage $\hat{A}_i = (r_i - \bar{r})/\sigma_r$ noisier:

$$\mathrm{Var}(\hat{A}_i^Q) > \mathrm{Var}(\hat{A}_i)\,. \tag{21}$$

Why OF helps. OF's two-pass residual QDQ recovers deadzone values: Pass 1 quantizes normally (outliers are accurate, small values enter the deadzone); Pass 2 quantizes the residual (whose dynamic range is small enough for previously dead values to be represented). The DZ rate drops from 9% to 2%, directly restoring effective rank and rollout quality. In our dense experiments, OF alone provides +17.5 pp recovery (60.1% → 77.6%).

OF residual-blend coefficient α. The OF reconstruction $\hat{\boldsymbol{x}}_{\mathrm{OF}} = \hat{\boldsymbol{x}}_1 + \alpha\,\hat{\boldsymbol{x}}_2$ (Eq. 6) includes a blend coefficient α. The naive choice $\alpha = 1$ adds the full residual back, but $\hat{\boldsymbol{x}}_2$ is itself an MXFP4-quantized approximation of the residual and inherits its own grid + deadzone error. Adding it at half strength avoids over-correction and improves rollout quality on GSM8K (Table 6).

Table 6: OF residual-blend coefficient α sensitivity (Qwen3-30B-A3B-Base, GSM8K, W4A4 with MBS+AQN). $\alpha = 0.5$ (used throughout the paper) outperforms $\alpha = 1.0$ by $\sim$1 pp.

α	GSM8K final (%)	GSM8K peak (%)
0.5 (default)	**91.58**	**92.49**
1.0	90.45	90.45

Why $\alpha = 0.5$: rollout–training drift minimization. The static argument above explains why $\alpha = 1$ is suboptimal, but not why the optimum sits near one-half. We find a sharper, training-specific reason: $\alpha = 0.5$ minimizes the rollout-vs-training numerical drift that RL is sensitive to. In a controlled single-step isolation run (Qwen3-30B-A3B, OF applied identically to weights and activations, MBS/AQN disabled, learning rate 0, 32 fixed DAPO-MATH prompts so α is the only varying quantity), the divergence between the vLLM rollout policy and the training-side recomputed policy is U-shaped in α with a clean minimum at $\alpha = 0.5$ (Figure 5); both extremes incur roughly 1.5–2× the drift of $\alpha = 0.5$. The interpretation is that a half-strength residual best aligns the rollout fake-quantization grid with the training grid: $\alpha < 0.5$ under-recovers the deadzone, while $\alpha > 0.5$ over-corrects per-channel and amplifies the mismatch between the rollout and training kernels. Since RL post-training is drift-sensitive, the α that minimizes this mismatch is also the α that maximizes downstream accuracy, consistent with the GSM8K gap in Table 6.

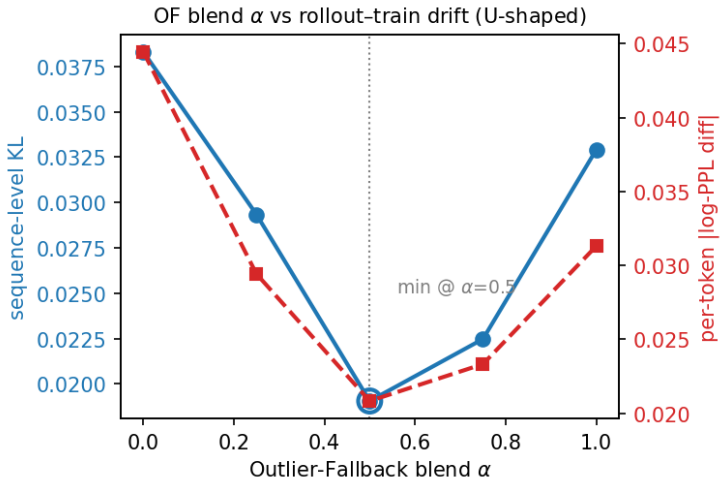


Figure 5: Outlier-Fallback blend α versus rollout–training drift (Qwen3-30B-A3B, controlled single-step isolation: OF only, MBS/AQN/attention-quant disabled, learning rate 0, 32 fixed DAPO-MATH prompts). Sequence-level KL between rollout and training log-probabilities, and per-token absolute log-perplexity difference, are both U-shaped in α with a clean minimum at $\alpha = 0.5$. The training-dynamics counterpart of Table 6.

Dense vs. MoE asymmetry. In MoE architectures, expert routing acts as a natural error-correcting code: even if one expert's small weights are pruned, other experts can compensate. Dense models have no such redundancy, so deadzone damage is fully expressed. This explains why OF is critical for dense (+17.5 pp) but only modestly helpful for MoE (+1.5 pp). Figure 6 shows the training curves.

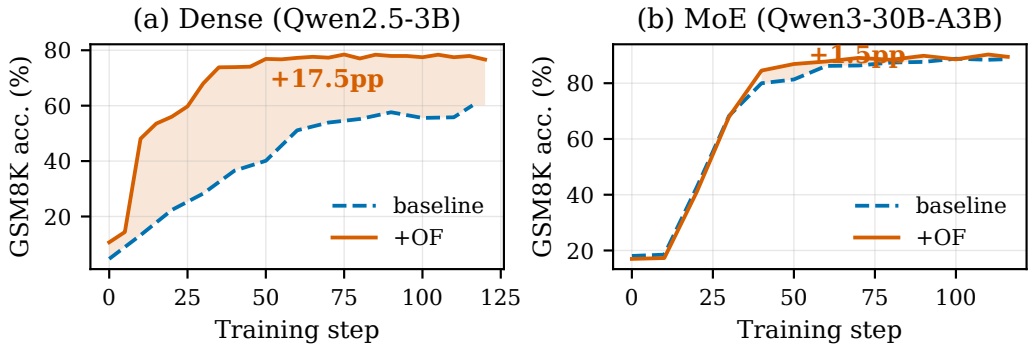


Figure 6: OF sensitivity by architecture. (a) Dense: OF provides +17.5 pp. (b) MoE: OF provides only +1.5 pp. MoE expert routing compensates for deadzone pruning.

D.3 Pathway 2 details: policy gradient bias

Gradient scale derivation. For a single linear layer with quantized weight, the forward pass output is $\hat{\boldsymbol{y}} = \gamma_b \cdot \boldsymbol{W}_b \boldsymbol{x} + \tilde{\boldsymbol{e}}_b \boldsymbol{x}$, where $\tilde{\boldsymbol{e}}_b$ contains element-level errors. The STE backward pass treats Q as identity, computing $\partial \mathcal{L}/\partial \boldsymbol{W}|_{\text{STE}} = (\partial \mathcal{L}/\partial \hat{\boldsymbol{y}}) \cdot \boldsymbol{x}^\top$ at a point shifted by γ_b. Through the chain rule across L layers:

$$\hat{\nabla}_\theta \mathcal{L} = \prod_{l=1}^{L} \gamma_{b(l)} \cdot \nabla_\theta \mathcal{L}\big|_{\text{true}} + \text{higher-order}\,. \tag{22}$$

Since $\delta_b \sim \text{Uniform}[0, 1)$ approximately i.i.d., by the CLT:

$$\sum_{l=1}^{L} \delta_{b(l)} \sim \mathcal{N}(0,\ L/12)\,. \tag{23}$$

For Qwen2.5-3B with L=36, std $= \sqrt{36/12} = \sqrt{3} \approx 1.73$, so the gradient magnitude ratio ranges from $\sim$0.18$\times$ to $\sim$5.6$\times$ within one standard deviation.

Why MBS helps. MBS replaces $s_b = 2^e$ with $s_b^{\text{MBS}} = 2^e \cdot (1 + c/256)$, where c is an 8-bit mantissa correction at macro-block granularity (128 elements). This reduces $\text{Var}(\gamma)$ by $\sim(256)^{-2}$, effectively eliminating scale-induced gradient bias.

Connection to TIS. TIS (Section 6.1) clips the importance ratio ρ_i to $\rho_{\max}$. We originally selected a larger TIS $= 5$ for dense because quantization-induced logit perturbations increase ρ_i variance and a tighter clip can amplify the resulting gradient bias, which is consistent with scale bias being a gradient-pathway problem. Controlled single-knob ablations (Appendix H) refine this picture: varying only the TIS threshold, TIS $= 2$ does not collapse training (it is within $\approx$2.35 pp of TIS $= 5$ on both a naïve dense run and a full-stack run), so the per-model TIS choice is not what drives the dense/MoE dichotomy; that asymmetry follows from expert-routing redundancy against deadzone pruning (Section 6.2).

D.4 Pathway 3 details: exploration–exploitation perturbation

Grid noise distribution. The grid error for non-deadzone elements has variance depending on the local E2M1 step size:

$$e_{b,i}^{\text{grid}} \sim \text{Uniform}[-\Delta(u_{b,i})/2,\ \Delta(u_{b,i})/2]\,, \tag{24}$$

where $\Delta = 0.5$ for $|u| \in [0, 2]$, $\Delta = 1$ for $|u| \in [2, 4]$, and $\Delta = 2$ for $|u| \in [4, 6]$.

Effective temperature. This noise propagates to logits as $\boldsymbol{\ell}^Q = \boldsymbol{\ell} + \boldsymbol{\eta}$, $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \boldsymbol{I})$ approximately by CLT, where $\sigma_\eta^2 \propto \sum_l \sigma_{\delta,l}^2(\text{grid})$ accumulates over layers. Matching the signal-to-noise ratio of logit differences under temperature scaling versus additive noise (derivation below) gives $T_{\text{eff}} \approx \sqrt{1 + 2\sigma_\eta^2 / \operatorname{Var}(\Delta\ell)} > 1$, directly increasing policy entropy:

$$\mathcal{H}[\pi_{Q(\boldsymbol{\theta})}] > \mathcal{H}[\pi_{\boldsymbol{\theta}}]. \tag{25}$$

This matches the experimental observation: baseline entropy drops to 0.35 by step 50, while AQN+MBS entropy stays at 0.61.

Derivation of T_{eff} (Eq. 5). The square-root form arises from matching the softmax-induced action distribution under additive logit noise to the same distribution under deterministic temperature scaling. Three observations drive the derivation.

(i) Logit differences are all that matter. The softmax policy depends only on pairwise differences: for actions a, b,

$$\log \frac{\pi(a \mid s)}{\pi(b \mid s)} \;=\; \ell_a - \ell_b \;=:\; \Delta\ell_{ab}. \tag{26}$$

Temperature scaling at T shrinks these differences uniformly to $\Delta\ell_{ab}/T$.

(ii) Per-pair noise has doubled variance. With η_a, η_b i.i.d. $\sim \mathcal{N}(0, \sigma_\eta^2)$, the pairwise difference is

$$\Delta\eta_{ab} \;:=\; \eta_a - \eta_b \;\sim\; \mathcal{N}(0,\, 2\sigma_\eta^2), \tag{27}$$

which is the source of the factor 2 inside the square root of Eq. 5.

(iii) Probit–logit Gaussian-marginalization identity. Using the classical approximation $\sigma(z) \approx \Phi(z/\kappa)$ with $\kappa = \sqrt{8/\pi}$, and the fact that the convolution of a probit with a Gaussian is another probit with summed variances,

$$\mathbb{E}_{\Delta\eta_{ab}}\big[\sigma(\Delta\ell_{ab} + \Delta\eta_{ab})\big] \;\approx\; \sigma\left(\frac{\Delta\ell_{ab}}{\sqrt{1 + \operatorname{Var}(\Delta\eta_{ab})/\kappa^2}}\right), \tag{28}$$

where $\sigma(\cdot)$ is the logistic function and $\Phi(\cdot)$ the standard-normal CDF. The right-hand side is exactly a temperature-scaled deterministic softmax at $T = \sqrt{1 + 2\sigma_\eta^2/\kappa^2}$, recovering the square-root form.

From κ^2 to $\operatorname{Var}(\Delta\ell)$. Eq. 5 substitutes the empirical logit-difference variance $\operatorname{Var}(\Delta\ell)$ for the constant κ^2. This is a signal-to-noise argument: κ is the characteristic logit scale over which the sigmoid transitions, while $\operatorname{Var}(\Delta\ell)^{1/2}$ is the policy's *own* characteristic logit scale at the current training step. The substitution makes T_{eff} adaptive to the current policy sharpness: (a) when the policy is confident, $\operatorname{Var}(\Delta\ell) \gg \sigma_\eta^2$, so $T_{\text{eff}} \to 1$ and noise is negligible; (b) when the policy is flat, $\operatorname{Var}(\Delta\ell) \ll \sigma_\eta^2$, so $T_{\text{eff}} \to \sqrt{2\sigma_\eta^2 / \operatorname{Var}(\Delta\ell)}$ and noise dominates. The "1+" clamps $T_{\text{eff}} \geq 1$: zero-mean grid noise can only *widen*, never sharpen, the policy. In the noise-free limit $\sigma_\eta \to 0$, $T_{\text{eff}} = 1$ recovers the unperturbed policy.

AQN as controlled exploration. AQN injects scheduled noise $\sigma_{\text{AQN}}(t) : \sigma_{\text{start}} \to \sigma_{\text{end}}$, giving total noise $\sigma_{\text{total}}(t) = \sqrt{\sigma_{\text{grid}}^2 + \sigma_{\text{AQN}}^2(t)}$. Early in training, $\sigma_{\text{AQN}} \gg \sigma_{\text{grid}}$ provides controlled broad exploration; late, $\sigma_{\text{AQN}} \to 0$ allows exploitation.

AQN+MBS synergy. Without MBS, the noise floor includes scale error ($\sim 54\%$ RMSE in the scale ratio, block-correlated). This structured noise contaminates AQN's controlled exploration. After MBS reduces σ_{scale} to $\sim 0.1\%$, only the i.i.d.-like σ_{grid} remains, which AQN can effectively "train through". This explains the superadditive result: MBS alone $+0.9$ pp, AQN alone $+0.4$ pp, combined $+2.0$ pp. Figures 7 and 3 provide supporting evidence.

E Complementarity with upstream techniques

Our two error corrections (MBS, OF) operate during quantization, while AQN operates on the training dynamics. An alternative strategy is to reshape the input distribution *before* quantization so that

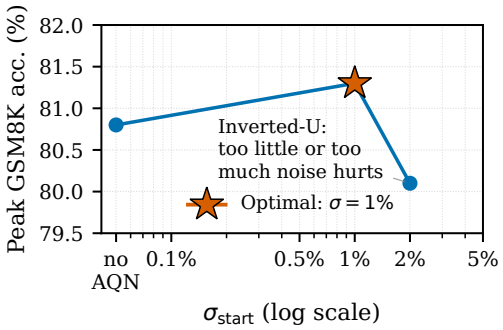


Figure 7: AQN σ_{start} sensitivity (Dense, MBS+OF). $\sigma = 1\%$ is optimal; 2% overshoots and degrades below no-AQN baseline.

the format's limitations bite less. Stochastic rounding (SR), originally introduced for low-precision neural training by Gupta et al. (2015), and random Hadamard transforms (RHT), as used by Tseng et al. (2025) for MXFP4 pre-training, both follow this upstream approach.

In terms of our decomposition, SR primarily targets $\boldsymbol{e}^{\text{grid}}$: it makes the rounding error exactly zero-mean at every point (not just on average; classical quantization-noise theory (Bennett, 1948; Widrow & Kollár, 2008) already gives the additive-uniform model that SR enforces unconditionally), eliminating the local bias within each quantization bin. SR also softens the deadzone (a value below the threshold has a nonzero probability of rounding to the nearest nonzero grid point) but does not eliminate it. RHT primarily targets $\boldsymbol{e}^{\text{DZ}}$: by spreading outliers across elements via a random orthogonal rotation, it equalizes magnitudes within each block, reducing the fraction of values that fall below the deadzone threshold $m_b/24$.

Crucially, neither SR nor RHT addresses $\boldsymbol{e}^{\text{scale}}$ directly: the E8M0 power-of-two scale remains after any input rotation. MBS is the only technique in our toolkit (or theirs) that corrects this component. The two families of techniques are therefore **complementary**: SR/RHT reshape the distribution to reduce deadzone and grid error at the source, while MBS/OF/AQN correct whatever error remains after quantization. Exploring their combination is a promising direction for future work.

F Empirical validation of decomposition structure

We measure the pairwise cosine similarities and the cross term on real weight tensors to validate the MSE identity (Eq. 2) and the anti-correlation (Remark 3.3). For each linear layer's weight tensor, we compute the three error components (Definition 3.1) and report:

1. **Identity verification**: confirm that $\left\|\boldsymbol{e}^{\text{scale}}\right\|^2 + \left\|\boldsymbol{e}^{\text{DZ}}\right\|^2 + \left\|\boldsymbol{e}^{\text{grid}}\right\|^2 + 2\langle \boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}} \rangle = \left\|\boldsymbol{e}\right\|^2$ holds exactly (up to floating-point precision) for each tensor.
2. **Exact orthogonality**: verify $\cos(\boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{DZ}}) = 0$ and $\cos(\boldsymbol{e}^{\text{DZ}}, \boldsymbol{e}^{\text{grid}}) = 0$ exactly, confirming Lemma 3.2.
3. **Anti-correlation**: report $\cos(\boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}})$ per layer; expected ≈ -0.66 consistently (Remark 3.3).

The main-body Table 1 reports aggregated results across three tested architectures, and the accompanying main-body Figure 1 visualizes the distribution of pairwise cosine similarities across all weight tensors, confirming: $\cos(\boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{DZ}}) = 0$ and $\cos(\boldsymbol{e}^{\text{DZ}}, \boldsymbol{e}^{\text{grid}}) = 0$ exactly (Lemma 3.2), while $\cos(\boldsymbol{e}^{\text{scale}}, \boldsymbol{e}^{\text{grid}}) \approx -0.66$ with remarkably tight spread.

G Long chain-of-thought: scope and trajectory divergence

Section 4.3 argues that grid noise, being zero-mean, raises the policy's effective temperature but "averages out over full response sequences." This holds for the short-to-medium reasoning of GSM8K

but reaches its limit as chains lengthen. We document the boundary as a scope limitation, not a revision of the GSM8K result.

Per-step drift is task-independent; outcomes are not. Holding the recipe family fixed (MBS+OF / AQN+MBS+OF), the rollout–actor probability Pearson correlation is essentially identical on a short task (GSM8K, median ≈ 280-token responses) and a long one (DAPO-MATH, ≈ 795 tokens): 0.999 in BF16 versus ≈ 0.97 in MXFP4 in both (0.975 / 0.971, Figure 8). The MXFP4 per-step drift is intrinsic to W4A4, not task-specific. Yet outcomes diverge: the short task *exceeds* BF16 by $+1.0$ pp (Table 2), while on the DAPO-MATH long-response validation set the same recipe under-recovers substantially. The only variable separating recovery from failure is sequence length: the floor does not cancel but *compounds* under longer autoregressive generation.

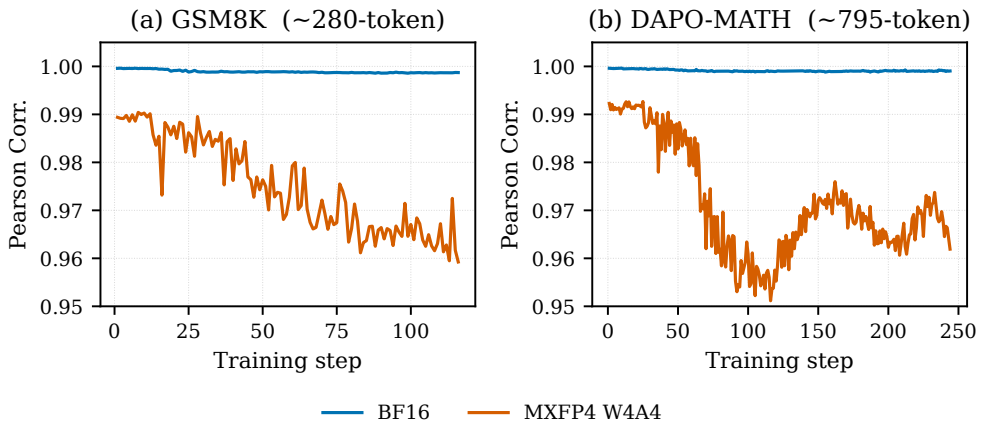


Figure 8: Per-step rollout–actor probability Pearson correlation across training, on a short task (a, GSM8K) and a long one (b, DAPO-MATH). BF16 stays at ≈ 1.000 on both tasks; MXFP4 W4A4 drifts to a mean of ≈ 0.97 on both (0.975 on GSM8K, 0.971 on DAPO-MATH). The MXFP4 per-step drift is therefore intrinsic to W4A4 rather than task-specific, yet end-task outcomes diverge sharply with response length (Table 2 for the short task; under-recovery on the long task). The same per-step number gives recovery in one regime and failure in the other.

Mechanism: greedy trajectory fork. Under deterministic greedy decoding (no sampling noise, so any divergence is purely the quantization floor), BF16 and MXFP4 policies starting from the same base checkpoint produce near-disjoint trajectories (Figure 9). A high teacher-forced per-token correlation looks benign because it is the quantity importance-sampling sees, but free generation occasionally flips an early arg-max, and that single flip compounds into a different solution path. After RL training the fork becomes systematically harmful: greedy disagreements are dominated by BF16-correct/MXFP4-wrong, with MXFP4 also producing many more unparseable or non-terminating responses. The long-CoT gap is therefore driven by trajectory divergence and non-termination, not by the static error budget our corrections minimize; importance-sampling-side fixes cannot close it because reweighting tokens cannot undo the fact that the explored *trajectories* differ.

Quality vs. response length. Under a strict 4-gram repetition detector that excludes unformatted-but-valid outputs and length-capped truncation, W4A4 quality degrades earlier than BF16 with length, and degeneracy dominates the long tail (Figure 10). The recovery the corrections achieve on short CoT does not extend to the long tail.

Where the gap is *not*. In our setup, rollout and training forward passes already share the same QDQ path (identical E8M0 block scale, E2M1 grid, OF residual, rounding and block size), so the residual divergence is intrinsic to free-running autoregressive decoding rather than a rollout–training *kernel* mismatch. The most promising next steps for long-CoT fidelity are therefore higher activation precision (e.g. W4A8) rather than rollout-kernel alignment.

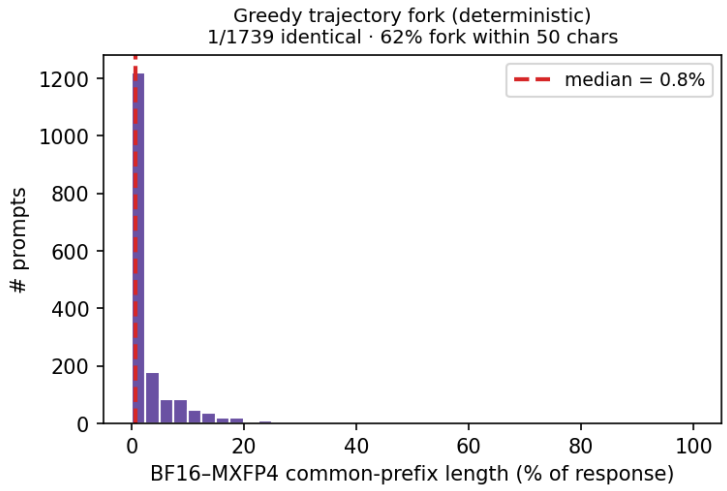


Figure 9: Greedy trajectory fork (deterministic decoding, no sampling noise; 1,739 DAPO-test prompts, shared base checkpoint). Distribution of BF16-vs-MXFP4 common-prefix length: trajectories diverge early and are near-disjoint. Because greedy decoding is deterministic, every divergence is attributable to the quantization floor alone. On *trained* checkpoints the fork becomes systematically harmful: greedy disagreements split asymmetrically toward BF16-correct/MXFP4-wrong, with MXFP4 also producing several-fold more unparseable / non-terminating outputs.

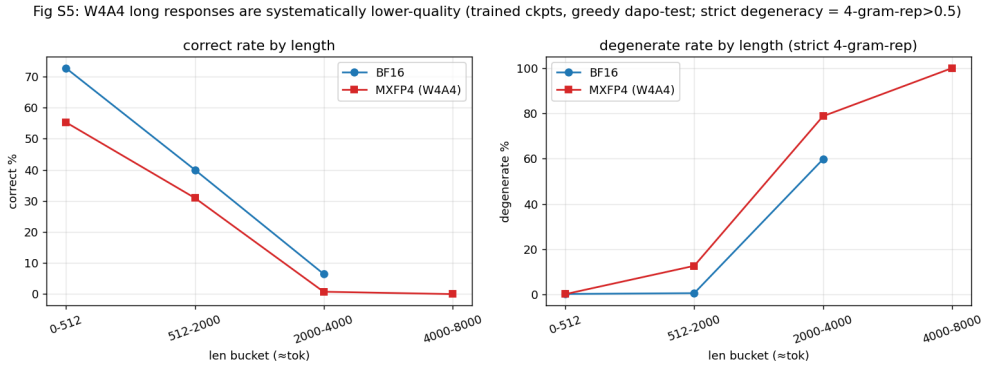


Figure 10: Quality vs. response length on the DAPO long-CoT validation set (trained checkpoints, greedy). *Left:* correct rate by length bucket. *Right:* degenerate rate under a strict 4-gram repetition detector (rep > 0.5; excludes unformatted-but-valid outputs and length-capped truncation, which would otherwise inflate the count). W4A4 degenerates earlier and more, and dominates the long tail.

H Additional evaluation and controlled ablations

This appendix reports additional experiments that (i) test the diagnosis on a new model family and task, (ii) compare against prior low-precision methods, and (iii) isolate the confounds discussed in the main text. Unless noted, the setting is MXFP4 W4A4 QDQ, GSM8K training, single seed; Llama full-stack results are three seeds.

Cross-family generalization: static decomposition. The component ratios and anti-correlation are near-identical on a third architecture (Llama-3.1-8B-Instruct), extending the pattern in Table 1 beyond the Qwen family: $\left\|e^{\text{scale}}\right\|^2 / \|e\|^2 = 1.726$, $\left\|e^{\text{DZ}}\right\|^2 / \|e\|^2 = 0.022$, $\left\|e^{\text{grid}}\right\|^2 / \|e\|^2 = 0.713$, and $\cos(e^{\text{scale}}, e^{\text{grid}}) = -0.657$ over 224 projection tensors.

The prior-method baseline comparison (Table 4) and the per-component forward-pass attribution (Table 5) are reported in the main body (Section 6.2). Figure 11 visualizes the main ablation results of Tables 2–3.

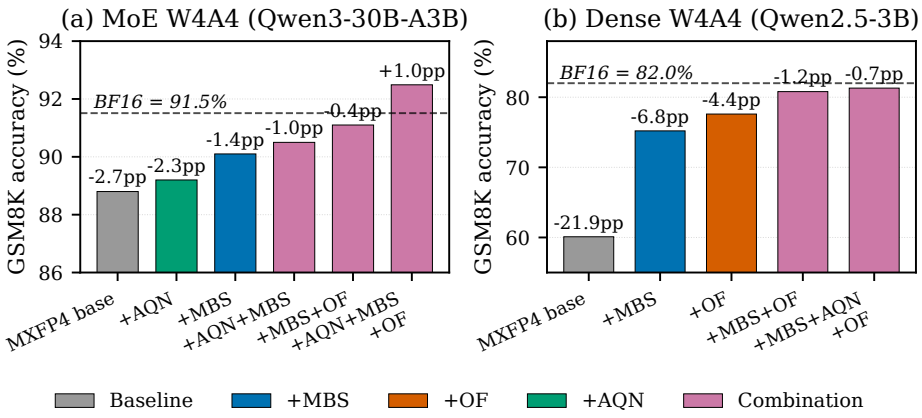


Figure 11: Ablation results on GSM8K. (a) MoE: corrections stack incrementally; AQN+MBS+OF reaches $+1.0$ pp (peak). (b) Dense: OF is critical ($+17.5$ pp alone); the full stack recovers to -0.7 pp. Dashed line: BF16 baseline.

Backward pathway: OF and AQN do not touch gradients. Measuring the gradient error (dX NMSE) confirms the off-diagonal of the 3×3 mapping: full-stack backward equals MBS-alone (-11.58% vs. naïve), because OF is forward-only and AQN is rollout-only (BRQ -1.35%). A stochastic-rounding baseline (FP4-All-the-Way) lowers gradient *bias* but its single-sample *variance* makes total dX MSE 51% *higher* than naïve, consistent with the decomposition's treatment of zero-mean vs. coherent error.

Causal isolation of the scale-bias→gradient pathway (per-step trajectory). This is the per-step counterpart of the $37.68\% \rightarrow 72.78\%$ backward-only swing reported in Section 4.1. The recovery is not an early-training artifact: the ideal-Q^* backward run actually *trails* naive E8M0 MXFP4 through step 30 (14.33 vs. 19.11), overtakes it at step 40 (39.58 vs. 22.14), and separates thereafter to finish at 72.78 vs. 37.68. The crossover is the signature of the multiplicative backward amplification of Section 4.1: the accumulated scale bias compounds over training rather than at initialization, and removing it from the backward pass alone unblocks the gradient pathway. Single seed; forward and rollout are held at standard MXFP4 throughout.

TIS is not the confound. Varying only the TIS threshold (same environment) shows TIS $= 2$ is stable, not collapse-inducing (Table 7); the $\approx$2.35 pp TIS effect is far smaller than the $\approx$16 pp OF asymmetry (OF $+17.5$ dense vs. $+1.5$ MoE), so the dense/MoE dichotomy is not a clipping artifact. A same-Qwen dense full-stack TIS=2 cell remains future work.

Table 7: TIS isolation (final / peak GSM8K %): only the TIS threshold changes.

Setting	TIS=5	TIS=2
Qwen2.5-3B, naïve W4A4	60.1 / 60.1	62.5 / 62.6
Llama-3.1-8B, full stack	80.9 / 83.5	83.2 / 83.7

Second RL task with a different reward (DAPO-17K). On Qwen3-30B-A3B-Base trained with DAPO-17K (reward normalized as $(\text{raw}+1)/2$), the recipe recovers a meaningful fraction of the MXFP4→BF16 gap (Table 8): 56.2% of the reward gap and 27.3% of the AIME24 gap. We report this as cross-task benefit, not full BF16 recovery. AIME24/25 are used only as out-of-distribution diagnostics for these GSM8K/DAPO-trained checkpoints, and we do not draw quantitative conclusions from the exact pass@k magnitudes.

Table 8: DAPO-17K cross-task result (Qwen3-30B-A3B-Base).

Configuration	Normalized DAPO reward	AIME24 (reported)
BF16	0.563	27.08
Pure MXFP4	0.410	14.48
Ours	**0.496**	**17.92**

Train–inference mismatch dynamics. Per-step trajectories of importance-ratio dispersion, KL, and policy-gradient clip fraction quantify how quantization drives rollout–training divergence and how corrections pull it back (Table 9, final-step values). MBS+OAS roughly halves KL relative to naïve MXFP4 but does not restore BF16-like clip rates. Matching full-stack trajectories are not yet in the parsed set, so this is diagnostic evidence rather than a claim that the full stack removes the mismatch.

Table 9: Final-step training dynamics (Llama-3.1-8B).

Method	Entropy	IS-ratio std	KL	PG clip frac.
BF16	0.055	0.025	0.00034	0.00036
Naïve MXFP4	3.902	0.297	0.04894	0.02104
MBS+OAS	3.651	0.269	0.02361	0.02898
FP4-All-the-Way	7.960	0.258	0.03331	0.02957

I TRAINING HYPERPARAMETERS (MOE)

We list every hyperparameter explicitly set for our MoE training runs. All other knobs use the default values of the underlying RL framework.

GRPO

- Advantage estimator: `grpo`.
- `rollout.n` (rollouts per prompt): 2.
- `train_batch_size` (prompts): 64.
- `ppo_mini_batch_size` (prompts): 32.
- `ppo_micro_batch_size_per_gpu`: 1 (training and inference).
- `use_dynamic_bsz`: `True`; `ppo_max_token_len_per_gpu` = 2048.
- `max_prompt_length`: 1024; `max_response_length`: 1024.
- KL: `algorithm.use_kl_in_reward=False`, `algorithm.kl_ctrl.kl_coef=0.0`; `actor.use_kl_loss=False`.
- PPO clip (token-level, asymmetric): ϵ_{low}=0.2, ϵ_{high}=0.28 (DAPO-style clip-higher), dual-clip c=10.0; `loss_type=ppo_clip`.
- Token-level Truncated Importance Sampling (TIS) (Yao et al., 2025): `rollout_is=token`, `rollout_is_threshold=2.0`, `tis_strategy=fixed`, `bypass_mode=false`.
- `loss_agg_mode=token-mean`; `entropy_coeff=0`.
- Reward: rule-based GSM8K grader (verl built-in, `reward_model.enable=False` since rewards are computed from the dataset's ground-truth field).
- Optimizer: `lr=1e-6`, `lr_warmup_steps=10`, `lr_decay_style=constant`, `weight_decay=0.1`, `clip_grad=1.0`.
- `trainer.total_epochs=1`; `trainer.val_before_train=True`; `trainer.test_freq=10`; `trainer.save_freq=300`.
- Single seed per configuration; statistical significance limitation acknowledged in Section 7.

ROLLOUT

- Engine: vLLM (`rollout.name=vllm`); `enforce_eager=True`; `free_cache_engine=True`; `enable_chunked_prefill=True`; `max_num_batched_tokens=8192`.
- `tensor_model_parallel_size=8` for inference.
- `gpu_memory_utilization=0.75`.
- Training rollout: `temperature=1.0`, `top_p=1.0`, `top_k=-1`.
- Validation rollout (`val_kwargs`): `temperature=1.0`, `top_p=1.0`, `top_k=-1`, `do_sample=True`, `n=1`.
- `calculate_log_probs=True`; `enable_rollout_routing_replay=True` (MoE router cached from rollout to actor).

QUANTIZATION (MXFP4 W4A4)

- Format: MXFP4 (E2M1 grid + E8M0 power-of-two block scale) (Rouhani et al., 2023); W4A4, both weight and activation tensors entering each matmul are quantized; gradients remain BF16.
- Block size: 32 along the input axis of each Linear.
- Quantizer path: native CUDA kernel (`actor.force_py=False`); inference uses the same kernel via `rollout.prequant_format=mxfp4`.

CORRECTIONS

Macro Block Scaling (MBS) (Chhugani et al., 2026). Macro-block size 128 ($= 4$ contiguous MXFP4 sub-blocks of 32); 8-bit mantissa correction (E0M8, 256 levels) per macro-block; applied identically to weights on both OF passes when MBS+OF is enabled.

Outlier Fallback (OF) (Zhang et al., 2025). Two-pass residual MXFP4 QDQ with blend $\alpha = 0.5$: $q_1 = \mathrm{MXFP4}(\boldsymbol{w})$, $q_2 = \mathrm{MXFP4}(\boldsymbol{w} - q_1)$, output $q_1 + \alpha\, q_2$.

Adaptive Quantization Noise (AQN) Huang et al. (2026). Per-rollout Gaussian additive noise injection on weights with exponentially decaying magnitude:

- `trainer.aqn.enabled=True`.
- `sigma_start=0.01`.
- `sigma_end=0.001`.
- `num_stages=10`.
- `target_patterns=['layernorm']`.
- `target_multipliers.post_attention_layernorm=1.414` (i.e. $\sqrt{2}$, compensating for the magnitude doubling at the residual-stream branch).

SYSTEM / PARALLELISM

- Model: `Qwen3-30B-A3B-Base`.
- Backend: Megatron-Core (PPO/GRPO Megatron trainer). Attention matmuls and grouped MoE GEMMs are both quantized to MXFP4 via custom quantization layers.
- Parallelism (training): `TP=1`, `PP=1`, `VPP=null`, `CP=1`, `EP=8`, `ETP=1`.
- Megatron MoE knobs: `moe_grouped_gemm=True`, `moe_token_dispatcher_type=flex`, `moe_router_dtype=fp32`, `moe_enable_deepep=True`, `moe_permute_fusion=False`.
- Memory: `param_offload=True`, `grad_offload=True`, `optimizer_offload=True` (`optimizer_offload_fraction=1.0`); `use_precision_aware_optimizer=True`; gradient checkpointing enabled.
- Checkpointing: `use_mbridge=True`, `use_dist_checkpointing=False`.

- Distributed setup: `trainer.n_gpus_per_node=8`, `trainer.nnodes=1`.
- Router replay: `router_replay.mode=R3` (replay rollout routing decisions during the actor update).
- LoRA: disabled (`lora.rank=0`; full-parameter training).

DATA

- Training: GSM8K `train.parquet` (7,473 problems).
- Validation: GSM8K `test.parquet` (1,319 problems).
- Preprocessing: parquet via `verl.data_preprocess.gsm8k`; prompt key `prompt`; truncation mode `error` (longer-than-budget prompts abort the job rather than silently truncate).