WHEN CLEANER CACHES BREAK: TRAINING RECIPES SHAPE MLA CACHE QUANTIZABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantizing the latent cache of multi-head latent attention (MLA) changes both reconstruction error and final predictive quality. We study how these outcomes depend on the trained model and on which cache coordinates share a quantization scale. Our evaluation uses three production models and six matched AdamW/Muon checkpoint pairs across three sizes. A streaming temporal policy preserves causality by quantizing groups only when they close. Under streaming INT3, Muon has greater loss damage in every pair despite cleaner marginal statistics, while retaining better final quality. Under tokenwise INT3, Muon has worse final quality at 125M and 350M, but better quality at 600M. At INT4 Muon keeps better final quality in every pair under both policies, so the reversal is specific to the seven-level symmetric INT3 reference. The activation measurements motivate a geometry-specific interpretation: channel outliers can enlarge neighboring channels' steps under tokenwise grouping, but not under per-channel temporal grouping. This grouping argument does not order the recipes: Muon has the lower outlier ratios yet the larger tokenwise damage at 125M and 350M. Controls that hold the closure schedule fixed while regrouping channels, and that tile the latent width evenly, preserve key recipe-dependent differences; an asymmetric INT3 quantizer, which stores one extra FP16 value per group, reduces Muon's streaming damage by 51–77% across the six checkpoints. Achieved fit remains a possible mediator of the sensitivity differences. Token-local noise probes, full-prompt MMLU, FP8 controls, and independent engine tests through 32K complement the causal comparisons. We report all matched seeds and distinguish compression damage from final quality at the tested training budget.

1 INTRODUCTION

Selecting a model for low-bit cache deployment requires choosing a training recipe and a cache policy. We test whether BF16 loss and cache reconstruction error select the best final quantized model. Multi-head Latent Attention (MLA) (DeepSeek-AI, 2024a) caches a learned latent from which keys and values are reconstructed. Training shapes this representation; the cache policy sets which coordinates share precision and when they are quantized.

We study whether the quality advantage of a training recipe survives native MLA cache quantization. Six AdamW/Muon (Jordan et al., 2024; Liu et al., 2025) checkpoint pairs share initialization, data order, architecture, and a 1B-token training budget within each pair. Muon reaches lower BF16 loss in every pair. Under a causal Stream3 policy, it incurs greater compression damage but still has lower final NLL. Under Token3, the final-quality ordering reverses in all four 125M/350M pairs, while the two 600M pairs continue to favor Muon. Thus the same BF16 recipe ranking leads to different quantized-model choices under different cache policies (Section 4).

Marginal activation shape does not explain this pattern by itself. Muon has lower channel-outlier ratios, yet its Token3 damage can be much larger. We summarize channel outliers by $\max_c a_c / \operatorname{median}_c a_c$, where $a_c = \max_t |z_{tc}|$ (Appendix M.1). The relationship also persists at the same precision: on BF16 reference activations, Muon's relative streaming-INT3 reconstruction MSE is 21.8–33.6% lower across the six pairs despite greater measured damage. Raw MSE reverses direction at the larger sizes, making the distinction between normalized distortion and absolute perturbation essential.

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.

The three production checkpoints expose a second decision: which policy to use for a fixed model. With write timing and the BF16 open tail held fixed, Tail3 has higher mean error and damage than Stream3 for V2-Lite and Coder. Moonlight reverses the damage ordering: its matched-tail policy has 24.3% higher relative error and 23.8% higher raw error, but 68.1% less damage. The disagreement holds for all 47 evaluated articles in relative error and damage. Reconstruction error therefore cannot alone settle either every recipe comparison or every within-model policy choice in this panel (Section 5). Selecting Moonlight's lower-error streaming policy would cost an additional 0.07445 nats/token of mean NLL.

Contributions. (1) *Policy-dependent recipe selection:* fixed-budget native-MLA pairs show when a BF16 advantage survives INT3 or reverses. (2) *Error–damage disagreements:* same-precision cross-recipe and within-model comparisons show that lower reconstruction error need not mean less predictive damage. Matched-tail, even-tiling, and common-BF16-input controls test alternative explanations. (3) *Operating conditions:* three production models, calibrated quantizers, perturbation probes, task accuracy, FP8, and independent engine measurements test extensions and exceptions. We report all matched seeds, adverse and inconclusive results, damage and final quality, and the failed achieved-fit matching control. The findings concern named recipes at fixed budgets, not faster decoding or converged matched-recipe production models; achieved fit remains a possible mediator.

2 Related work

Training and quantizability. Vlassis et al. (2025) compare six optimizers under weight-and-activation PTQ and QAT, finding that outlier metrics fail to predict PTQ quality and full-precision optimizer rankings need not survive QAT. Catalan-Tatjer et al. (2025) connect learning-rate decay to worsening PTQ robustness despite improving validation loss. OSP (Park et al., 2025) combines Muon with normalization and embedding changes to suppress outliers and improve four-bit quantization. These studies establish that training affects quantizability, but none varies the cache policy, so a recipe's quantized ranking reads as a property of the recipe. Our added axis is exactly that: on fixed checkpoint pairs, a pair's *final NLL ordering* changes with the causal native-MLA policy. Learning rate and weight decay travel with the optimizer here, so equal-fit optimizer effects remain unresolved.

Training-time cache geometry. Sun et al. (2026) regularize standard K/V, changing preferred scaling axes and final-quality rankings across bit widths and quantizer grids. The advantage vanishes under their tested KIVI-style configuration, and evaluation uses cache-disabled full-sequence forwards. This is the closest prior design to ours; the difference is what varies. We hold the checkpoints and the seven-level symmetric INT3 quantizer fixed and change only the grouping and closure schedule of the causally quantized MLA latent, so a ranking reversal is attributable to the policy and is read off final NLL rather than a reconstruction proxy.

Cache representations and precision. KIVI (Liu et al., 2024) uses per-channel keys and per-token values; KVQuant (Hooper et al., 2024) adds sensitivity-aware nonuniform quantization and outlier handling. Transform-based methods (Li et al., 2025; Zandieh et al., 2026; Staniszewski & Łańcucki, 2026) expand these quantizer choices. Palu (Chang et al., 2025) creates and quantizes post-hoc low-rank cache representations; MLA (DeepSeek-AI, 2024a;b) learns its latent during training. This shared latent makes recipe and scale grouping a coupled choice. FP8 kernels (DeepSeek-AI, 2025) address a separate practical precision regime from our sub-8-bit reference policies.

Distortion and predictive sensitivity. HAWQ, AdaRound, and BRECQ (Dong et al., 2019; Nagel et al., 2020; Li et al., 2021) already account for loss sensitivity beyond unweighted tensor error; SmoothQuant and QuaRot (Xiao et al., 2023; Ashkboos et al., 2024) reshape activation geometry. Our contribution is a measured ranking disagreement at the same INT3 precision (Section 5): lower BF16-reference relative MSE accompanies greater cross-recipe damage, and Moonlight's matched-tail policy has higher *raw and relative* MSE but lower damage. Raw MSE and common-input controls delimit normalization and feedback explanations; finite noise probes do not identify a Hessian mechanism.

3 MODELS, CAUSAL POLICIES, AND COMPARISON UNITS

Controlled recipes and production models. The twelve dense DeepSeek-V3-style checkpoints comprise three paired seeds at 125M, one at 350M, and two at 600M. Each recipe trains on 1B FineWeb-Edu tokens with GPT-2 tokenization. Within a pair, initialization, data order, architecture, and schedule are matched. Selected peak learning rates are 10^{-3} for AdamW and 0.01 for Muon hidden matrices, with hidden-matrix weight decay 0.1 and 0.01; Muon uses AdamW for embeddings and norms. The three production models are DeepSeek-V2-Lite, DeepSeek-Coder-V2-Lite-Instruct, and Moonlight-16B-A3B-Instruct (DeepSeek-AI, 2024a; Liu et al., 2025). Production comparisons are observational; training data, tokenizer, and base/instruct status differ. Appendix E records training details and checkpoint revisions.

Immediate and closure-based quantization. Token3/4 (immediate tokenwise INT3/4, printed as *Token INT3/4* in the tables) quantizes 128-channel groups within each token. Stream3/4 instead assigns each channel a scale over 128 consecutive tokens. An incomplete temporal group remains BF16; when token 128 arrives, the first group is quantized *before* that token's attention. Earlier hidden states are not recomputed. This schedule repeats at each closure, using only arrived tokens. Tail3/4 uses the same closure schedule and BF16 open group as Stream3/4, but quantizes each closed token in 128-channel groups. It therefore isolates grouping from those two scheduling differences.

The reference uses symmetric round-to-nearest, FP16 scales, and seven INT3 levels $\{-3, -2, -1, 0, 1, 2, 3\}$ in a nominal three-bit payload. The decoupled RoPE key remains BF16. Asymmetric temporal INT3 (Asym3) and rotation plus calibrated codebook INT3 (RtC3) use eight levels and additional metadata, specified in Appendix H.1. The research implementation reconstructs expanded K/V to evaluate quality. Packed-engine measurements are a separate experiment.

Matched data and losses. A frozen WikiText-103 validation partition assigns four articles to calibration, eight to development, and 48 to confirmation. All 47 confirmation articles eligible for 1,024 tokens under every recorded tokenizer are evaluated, scoring 1,023 next-token predictions each. Article hashes separate the three partitions. Twins share identical scored GPT-2 sequences across recipes; production effects are interpreted within checkpoint because token units and scored character spans can differ.

For checkpoint m, policy p, precision b, and article i, damage is

$$d_{m,p,b,i} = L_{m,p,b,i} - L_{m,p,0,i}, \tag{1}$$

where $L_{m,p,0,i}$ is the policy's matched BF16 NLL. Token3/4 results use full-forward BF16; Stream3/4 and Tail3/4 use sequential BF16. We report both mean damage and final NLL. Full-forward/sequential differences are not charged to quantization. The first 127 predictions under the closure schedule see only BF16 entries; subsequent closures also quantize the initial tokens. Appendix A quantifies exposure and specifies the scale and padding arithmetic.

Reconstruction and uncertainty. Relative MSE divides pooled squared reconstruction error by latent energy; raw MSE divides it by the real element count. The realized-trajectory diagnostics measure each policy's own cache writes. A separate BF16-reference control applies Stream3 and Tail3 to identical unquantized tensors within each twin checkpoint, without feeding reconstructed values into the model. Both closure policies measure seven completed blocks, or 896 positions per layer. Equations and aggregation details are in Appendices J–K.

All displayed 95% intervals resample matched articles 10,000 times, conditional on the fixed checkpoints. Seeds are 20260911 for earlier panels, 20260913 for the added policy controls and Moonlight, and 20260915 for the INT3 diagnostics and BF16-reference analyses; each appendix specifies its article ordering. These are pointwise intervals, not multiplicity-adjusted or training-seed intervals. Post-run contrasts are descriptive, and an interval spanning zero is not evidence of equivalence. Seed dispersion is reported separately.

Execution validation. The production reference ran on H100 with BF16 weights, eager attention, grouped expert operations, PyTorch 2.13.0+cu126, and Transformers 5.14.1; the twin and production-control extensions ran on Apple MPS, retaining their own BF16 baselines. The evaluated single-token paths matched frozen references with zero measured drift on 257-token development probes. CPU tests check prefix invariance, closed-group reconstruction, open groups, and positional

Table 1: Existing corrected causal twin results, 47 articles per checkpoint. Stream4/Stream3 damage uses the incremental BF16 baseline. Final INT3 NLL shows why greater damage must not be conflated with worse absolute quality.

Size	Seed	Recipe	BF16 NLL	ΔINT4	ΔINT3	Final INT3
125M	0	AdamW	3.93102	+0.00165	+0.00755	3.93857
125M	0	Muon	3.83479	+0.00354	+0.02968	3.86446
125M	1	AdamW	3.92792	+0.00366	+0.00763	3.93555
125M	1	Muon	3.82472	+0.00409	+0.03478	3.85950
125M	2	AdamW	3.94432	+0.00455	+0.01139	3.95571
125M	2	Muon	3.84959	+0.00498	+0.02972	3.87930
350M	0	AdamW	3.74223	+0.00125	+0.00409	3.74632
350M	0	Muon	3.64772	+0.00387	+0.01890	3.66661
600M	0	AdamW	3.72623	+0.00116	+0.00681	3.73304
600M	0	Muon	3.58961	+0.00173	+0.01041	3.60002
600M	1	AdamW	3.69573	+0.00269	+0.00776	3.70349
600M	1	Muon	3.60652	+0.00173	+0.01140	3.61792

Table 2: Muon minus AdamW damage for every matched seed pair. INT3 excess is positive under streaming in all pairs, but negative under tokenwise quantization in both 600M pairs. CIs resample matched articles, conditional on the individual checkpoint pair; they do not measure training-run uncertainty.

		Stream3		Token3	
Size	Seed	Excess	95% CI	Excess	95% CI
125M	0	+0.02213	[+0.01894, +0.02537]	+0.20377	[+0.19274, +0.21605]
125M	1	+0.02714	[+0.02333, +0.03118]	+0.17655	[+0.16373, +0.18917]
125M	2	+0.01832	[+0.01480, +0.02202]	+0.59308	[+0.54325, +0.64462]
350M	0	+0.01481	[+0.01148, +0.01793]	+0.43989	[+0.40647, +0.47580]
600M	0	+0.00360	[+0.00128, +0.00589]	−0.20750	[−0.21933, −0.19542]
600M	1	+0.00364	[+0.00142, +0.00589]	−0.12824	[−0.13780, −0.11898]

keys. The failed accelerated chunk path and historical future-dependent temporal NLL results are excluded. Appendix I documents the corrections and provenance.

4 Cache policy changes the final-quality ranking

Muon achieves lower BF16 loss in every controlled pair. Whether this advantage survives INT3 depends on the complete cache policy. Table 1 reports the streaming comparison; the complete Token3/4 final losses appear in Table 3. All twelve checkpoints are retained throughout.

Greater damage can coexist with better final quality. Under Stream3, Muon incurs greater mean damage in all six pairs, and all paired-article excess-damage intervals are positive (Table 2). Its BF16 advantage is nevertheless larger than that excess: final quantized NLL remains lower in every pair. Seed-mean Muon/AdamW damage ratios are $3.54\times$, $4.62\times$, and $1.50\times$ at 125M, 350M, and 600M. These sensitivity ratios alone would give the wrong final-quality ranking for the streaming policy.

Token3 changes which recipe wins. The four 125M/350M pairs favor AdamW after Token3, with Muon-minus-AdamW final-NLL gaps of approximately $+0.108$, $+0.073$, $+0.498$, and $+0.345$ nats/token. The two 600M gaps favor Muon (-0.344 and -0.218). All corresponding paired-article intervals exclude zero. At INT4, Muon retains lower final NLL under both policies in every pair. Thus neither the BF16 ranking nor the greater streaming-INT3 damage alone determines the model choice at another cache policy or precision.

The largest Token3 case is 125M seed 2: Muon's NLL increases from 3.84919 to 4.48561, while AdamW moves from 3.94444 to 3.98777. The final gap is $+0.49783$ with paired interval $[+0.45058, +0.54704]$. This checkpoint has an original channel-outlier ratio of 1.83 and kurtosis 3.02. Its failure is therefore not ordered by those aggregate marginal descriptors. All seeds and both

Table 3: Complete tokenwise final NLLs for all 12 controlled checkpoints. The baseline is full-forward BF16. Muon retains lower final NLL at tokenwise INT4 in every pair, but Token3 makes Muon worse at 125M/350M and better at 600M. All conditions use identical GPT-2 token sequences.

Size	Seed	Recipe	BF16 NLL	Token4	Token3
125M	0	AdamW	3.93081	3.93749	3.97111
125M	0	Muon	3.83480	3.86785	4.07887
125M	1	AdamW	3.92781	3.93574	3.96999
125M	1	Muon	3.82465	3.84979	4.04339
125M	2	AdamW	3.94444	3.95550	3.98777
125M	2	Muon	3.84919	3.88965	4.48561
350M	0	AdamW	3.74209	3.74875	3.81371
350M	0	Muon	3.64737	3.67036	4.15888
600M	0	AdamW	3.72630	3.74431	4.00641
600M	0	Muon	3.58966	3.60020	3.66227
600M	1	AdamW	3.69583	3.72729	3.90692
600M	1	Muon	3.60637	3.61508	3.68922

precisions, including reversals and inconclusive damage differences, are retained in Appendices D and M.3.

The effect sizes vary across training runs: at 125M, Token3 excess damage has sample SD 0.23303 across three pairs, compared with 0.00442 for Stream3. With 3/1/2 pairs, the observations identify the tested checkpoint outcomes, not a precise population interaction with parameter scale. Reporting every pair makes that variation visible. The 600M AdamW checkpoints are also atypical: channel kurtosis 47.53 and 72.49 against 8.50–15.47 at the smaller sizes, and BF16 gains over 350M of 0.016 and 0.047 nats where Muon gains 0.041 and 0.058 under the same reused learning rate (Table 15). We therefore do not claim the 600M reversal as a scale effect.

5 Reconstruction error and loss damage have different orderings

Cross-recipe comparison at the same precision. On each checkpoint's unquantized BF16 reference activations, Muon has 21.8–33.6% lower streaming-INT3 relative MSE in all six pairs. Each ratio interval is below one, while each saved damage-difference interval is above zero (Table 4). This comparison removes quantization feedback from the reconstruction measurement. The error ordering therefore persists beyond the original offline-INT4 versus causal-INT3 comparison and the realized-trajectory INT3 diagnostics, whose relative reduction is 22.0–33.4%.

Table 4: Streaming INT3 on captured BF16 reference tensors: all six Muon/AdamW pairs. R^{B} is pooled relative MSE and A^{B} raw MSE per element; M/A denote Muon/AdamW. Damage d is the previously saved causal evaluation in nats per shared twin token, with its original sequential BF16 baseline. Brackets are pointwise 95% paired-article intervals (47 articles, 10,000 resamples).

Pair	Relative MSE ratio $R^{\mathrm{B}}_{\mathrm{M}}/R^{\mathrm{B}}_{\mathrm{A}}$	Raw MSE ratio $A^{\mathrm{B}}_{\mathrm{M}}/A^{\mathrm{B}}_{\mathrm{A}}$	Damage difference $d_{\mathrm{M}} - d_{\mathrm{A}}$
125M s0	$0.693\ [0.691, 0.694]$	$0.860\ [0.858, 0.861]$	$+0.02213\ [+0.01895, +0.02529]$
125M s1	$0.688\ [0.686, 0.689]$	$0.864\ [0.862, 0.865]$	$+0.02714\ [+0.02338, +0.03109]$
125M s2	$0.666\ [0.664, 0.667]$	$0.831\ [0.829, 0.832]$	$+0.01832\ [+0.01488, +0.02197]$
350M s0	$0.664\ [0.662, 0.666]$	$1.310\ [1.307, 1.313]$	$+0.01481\ [+0.01151, +0.01790]$
600M s0	$0.744\ [0.742, 0.745]$	$1.825\ [1.820, 1.831]$	$+0.00360\ [+0.00132, +0.00594]$
600M s1	$0.782\ [0.780, 0.783]$	$2.166\ [2.160, 2.172]$	$+0.00364\ [+0.00141, +0.00579]$

The normalization is consequential. Muon's raw streaming MSE is lower at 125M, but is $1.31\times$ AdamW's at 350M and $1.83/2.17\times$ at 600M. Lower relative error is therefore not lower absolute perturbation at every scale. The common-reference control fixes Stream3/Tail3 inputs *within* each

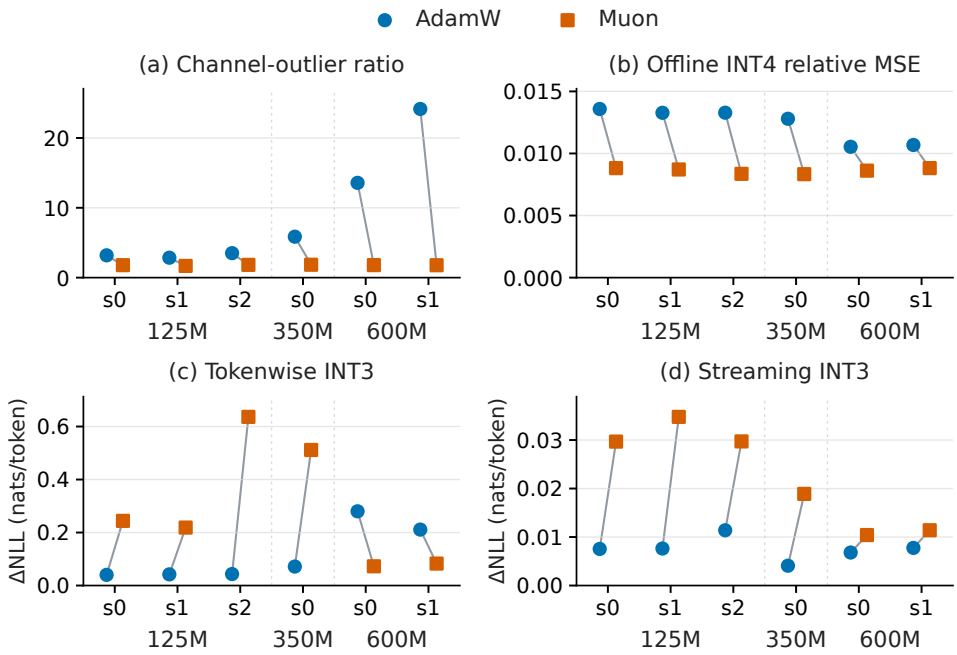

Figure 1: All six matched pairs. Muon has lower channel-outlier ratios and offline INT4 MSE, but greater Token3 damage at 125M/350M and less at 600M; Stream3 damage is greater in every pair. Points are checkpoint means; axes are independent. Panels (a,b) are descriptive offline measurements from four original 1K chunks; (c,d) show causal INT3 damage on 47 confirmation articles (paired intervals: Table 2). Same-precision BF16-reference comparisons appear in Table 4.

checkpoint; it does not equate learned activations across recipes. Loss damage remains the saved causal intervention with its sequential BF16 baseline. These distinctions support the observed ranking disagreement without treating a damage/MSE ratio as an identified sensitivity coefficient.

Within-checkpoint policy comparison often agrees with error. For the twelve twins plus V2-Lite and Coder, matched-tail INT3 has 1.38–$2.69\times$ the realized-trajectory relative MSE of streaming INT3, and higher mean damage in every checkpoint. Ten damage intervals are positive; four include zero. On common BF16 inputs, all twelve twins also retain higher Tail3 error. This agreement is part of the result: reconstruction error is informative in many of these policy comparisons, even though it does not order the cross-recipe effects.

Moonlight reverses the within-model damage ordering. At identical closure timing and open-tail precision, Tail3 damage is $+0.03481$ versus Stream3's $+0.10926$ nats/token, a 68.1% reduction. The paired difference is -0.07445 $[-0.08317, -0.06618]$. Yet Tail3 has $1.24264\times$ the relative MSE and $1.23777\times$ the raw MSE (Table 5; ratio intervals in Appendix L.2). All 47 articles have higher Tail3 relative error and lower damage; all 27 layers have higher pooled Tail3 relative MSE. Equal-article and equal-layer weighting preserve the error direction.

Selecting Stream3 over Tail3 solely for its smaller measured error would thus incur 0.07445 more mean NLL in this Moonlight comparison. That is a consequence for this fixed pair of policies, not an evaluation of a learned policy selector. Both policies measure 896 closed writes per layer and score 1,023 predictions. Their diagnostics follow their own realized trajectories; production BF16-reference captures remain an open control. Because raw and normalized error both reverse the damage ordering, normalization alone does not account for this example.

6 Controls identify which explanations survive

Write schedule and open-tail precision. Token3/4 and Stream3/4 differ in both grouping and exposure. Tail3 holds closure timing and the BF16 tail fixed while switching the closed blocks

Table 5: Moonlight INT3 reconstruction diagnostics on each policy's realized trajectory. R, A, and E are pooled relative MSE, raw MSE per element, and latent energy per element from Equation 2; all 27 layers and 47 articles are included. Token INT3 measures 1,024 positions per layer; the closure-based policies measure 896. These are not identical BF16 reference inputs.

Policy	R	A	E	ΔNLL
Stream INT3	0.068759	0.011511	0.167418	$+0.10926$
Tail INT3	0.085443	0.014248	0.166754	$+0.03481$
Token INT3	0.085452	0.014267	0.166961	$+0.06006$
RtC INT3	0.027546	0.004596	0.166867	$+0.05001$
Asym INT3	0.037067	0.006186	0.166892	$+0.06030$

to tokenwise groups. The twelve-checkpoint extension preserves greater Muon Tail3 damage at 125M/350M and lower damage at 600M. Final Tail3 quality favors AdamW for 125M seed 2 and 350M seed 0, and Muon for the other four pairs. The open-tail schedule attenuates the Token3/4 losses but does not remove the recipe dependence (Tables 18–19).

The 600M remainder is not necessary for the ordering. Latent widths are 256/384/448. Grouping width 448 by 128 leaves 64 real channels in the last group; zero padding does not change its absmax. Groups of 112 tile that width evenly while retaining four scales per token. Groups of 64 also tile it evenly, using seven scales. Muon retains lower Token3 damage under both controls in both 600M seeds. Group 64 narrows the gap substantially; its seed-1 paired difference is -0.00533 $[-0.01060, -0.00030]$, a narrowly negative pointwise interval. Tables 6 and 23 give the complete results. These controls exclude the remainder as a necessary explanation, while leaving width, architecture, and relative training progress coupled across sizes.

Table 6: Direct paired Muon-minus-AdamW Token3 damage for evenly tiled groups at 600M. Each interval uses 47 matched articles, 10,000 bootstrap resamples, seed 20260913, and confirmation-manifest article order. Intervals are pointwise and conditional on each checkpoint pair.

Group	Seed	Paired difference	95% CI
64	0	-0.02544	$[-0.03157, -0.01917]$
64	1	-0.00533	$[-0.01060, -0.00030]$
112	0	-0.15851	$[-0.16880, -0.14841]$
112	1	-0.10083	$[-0.10867, -0.09330]$

Quantizer choice can mitigate, but does not uniformly fix, damage. Asym3 reduces symmetric-streaming damage on all six Muon twins by 51.2–76.7%. RtC3 improves all three 125M Muon checkpoints but does not improve consistently elsewhere. Both use eight levels and two FP16 metadata values per 128 elements, compared with seven levels and one scale for the symmetric reference: ideal closed-block costs are 3.25 versus 3.125 bits per element, excluding the RtC codebook, BF16 tail, and positional branch. These named-policy comparisons do not isolate an offset or rotation effect at equal storage.

Equal-fit matching fails at this checkpoint cadence. The recipe intervention produces both better Muon BF16 loss and altered compression damage, so we searched the saved 600M stages for a Muon checkpoint matching the same-seed final AdamW loss on eight held-out development articles. We report the outcome as a negative result: the AdamW target does fall between two consecutive saved Muon stages in both seeds, but those stages are roughly 0.14 and 0.13 nats apart, so the nearest candidate still misses by 0.01783 and 0.04214 against a 0.01-nat tolerance (Table 24). We neither relaxed the threshold nor interpolated weights. Matching achieved fit therefore needs a denser checkpoint schedule or a run targeted at the AdamW loss, not post-hoc stage selection; the direct optimizer effect at equal fit stays unresolved.

7 Production and complementary evidence

Production policy rankings are checkpoint dependent. The complete three-model MPS extensions evaluate 47 articles and ten conditions each (1,410 records). V2-Lite and Coder favor Stream3

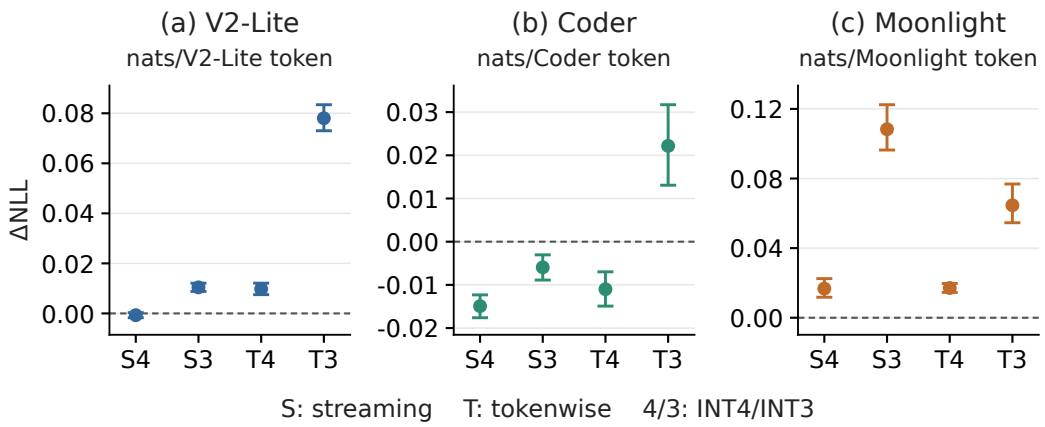

Figure 2: Causal production damage with pointwise paired-article 95% bootstrap intervals. Negative values indicate improvement. Streaming and tokenwise policies differ in grouping, scheduling, and open-group precision. Each panel has its own model-specific token unit and vertical scale; heights across panels are not a cross-model ranking. S4/S3 denote Stream4/Stream3 and T4/T3 Token4/Token3. Bars are grouped by policy, then ordered INT4 before INT3, so each policy pair is adjacent and the order is not monotone in bit width.

over Tail3 in mean damage by 0.00712 and 0.01481, respectively, with positive Tail3-minus-Stream3 paired intervals. Moonlight favors Tail3 by 0.07445. The grouping direction that reduces damage is therefore not shared across these three checkpoints. All comparisons remain within checkpoint and token unit.

RtC3 and Asym3 reduce Moonlight mean Stream3 damage by 54.2% and 44.8%, with negative paired intervals. RtC3 instead worsens V2-Lite and Coder; Asym3 worsens Coder and is inconclusive for V2-Lite. The production extension consequently supplies both mitigation cases and counterexamples to a uniform improvement. Full policy means, final NLLs, intervals, and reconstruction diagnostics appear in Appendices H.5, H.6, and L.

The independent H100 production panel retains its own baselines and all 846 records. Figure 2 plots damage; the complete final NLLs and intervals are in Table 41 (Appendix M, p. 35). Coder's Stream3/4 damage is negative on WikiText, whereas its Token3 damage is positive. This is a checkpoint-policy effect on predictive loss; corrected code-domain quality is not yet measured. MPS/H100 differences are documented rather than pooled or attributed solely to hardware. All recorded repeated-condition token identities agree across those execution panels.

Activation and perturbation evidence. For temporal group g and channel c, the INT3 step is $s_{gc} = \max_{t \in g} |z_{tc}|/3$; tokenwise group h instead uses $s_{th} = \max_{c \in h} |z_{tc}|/3$, ignoring metadata rounding. A high-amplitude channel can enlarge neighboring channels' steps in the latter, but not across the former's per-channel scales. The original latent, expanded-K/V, and GQA reconstruction measurements characterize this geometry (Table 40; Figure 1). AdamW's increasing outlier ratios accompany greater Token3 damage across the tested sizes; Muon's smaller-scale counterexamples prevent extending that association across recipes.

Token-local Gaussian probes add $\epsilon_t = \alpha \, \mathrm{RMS}(z_t)\xi_t$ at each latent hook, using only the current token to set amplitude. Muon has larger damage in all six pairs at both $\alpha = 0.114$ and 0.26. Its uncentered second-moment participation ratio is lower in every pair. These finite perturbation results extend sensitivity beyond RTN scale selection; they do not equate absolute noise norms across models or identify covariance rank, residual direction, or curvature mediation. Complete summaries and their available uncertainty are retained in Appendices M.4 and G.

Task, precision, and engine boundaries. On the separate 399-question full-prompt MMLU policy, Moonlight INT3 accuracy falls from 67.42% to 62.16%; exact paired McNemar $p = 0.00460$, or 0.0368 after adjustment over eight original comparisons. All model outcomes and discordances are reported in Table 7. This policy scores answer-letter logits after the supplied prompt; it is not the streaming-cache intervention.

Table 7: Recovered MMLU results, checked against all 399 per-question outcomes. Change is in percentage points; discordances are baseline-only correct: quantized-only correct. Exact two-sided McNemar p values are unadjusted.

Model	Prompt policy	BF16 correct	Quant. correct	Change	Discord.	p
V2-Lite	INT4	233/399	231/399	-0.50	12:10	0.8318
V2-Lite	INT3	233/399	230/399	-0.75	28:25	0.7838
Coder	INT4	249/399	251/399	+0.50	11:13	0.8388
Coder	INT3	249/399	249/399	+0.00	23:23	1.0000
Moonlight	INT4	269/399	261/399	-2.01	21:13	0.2295
Moonlight	INT3	269/399	248/399	-5.26	36:15	0.0046

Across 30 FP8 configurations, dynamic FP8 damage is only $+0.00014/+0.00036/+0.00037$ for V2-Lite/Coder/Moonlight. Independent llama.cpp (Gerganov et al., 2023) Metal tests with V2-Lite Q8_0 weights and q4_0 cache give $+0.01553/+0.01642/+0.01733$ NLL damage at 4K/16K/32K. These retained results establish their own task and precision regimes; different weights, grouping, and RoPE precision prevent treating the engine sweep as the research policy's long-context or serving-speed evaluation.

8 Discussion and conclusion

The tested recipe, cache policy, and quality criterion must be considered together. Muon's lower BF16 loss survives Stream3 in every pair, but Token3 reverses final quality at 125M/350M. Relative reconstruction error disagrees with streaming damage across all six recipe pairs, even on BF16 reference captures. Within-checkpoint error and damage often agree, yet Moonlight provides a same-schedule counterexample in both raw and normalized error. These observations make final quantized quality a necessary measurement for selecting among the tested recipe-policy combinations.

Scope and remaining questions. The causal evaluation covers 47 WikiText articles at 1K and 3/1/2 training pairs; article intervals do not estimate training-run uncertainty. Fixed 1B-token budgets give approximately 8.0/2.86/1.67 tokens per nominal parameter at 125M/350M/600M. Convergence is unestablished, particularly at 600M, and size, width, and training progress vary together. These are fixed-budget counterexamples, not evidence of persistence under longer training. Testing persistence requires longer matched-recipe runs, denser checkpoints, and separate comparisons at matched token budgets and BF16 loss. Production models extend within-checkpoint policy evidence to larger scale but cannot replace these controls.

Optimizer, selected peak learning rate, and hidden-matrix weight decay vary together; achieved BF16 loss is unmatched. Thus the comparisons concern whole recipes, not optimizer-only effects, whose isolation requires unperformed interventions varying one factor at a time. Production training histories and tokenizers differ, and BF16-reference captures remain incomplete. Population-level scaling and transfer to code, generation, and longer contexts remain open, as does the direct optimizer mechanism: accounts based on update geometry (Bernstein & Newhouse, 2024; Wang et al., 2026; Beneventano et al., 2026; Fan et al., 2025), learned representations (Ruan et al., 2026; Wang et al., 2025), or Muon-specific quantization (Sander et al., 2026) do not predict the measured sign.

Quality evidence is not a latency benchmark. The expanded-K/V causal reference measures quality, not packed Stream3/Tail3 NVIDIA decode latency or throughput. Neither H100 numerical checks nor separate Metal quality tests establish speedup. Ideal payload savings exclude the BF16 open tail and packing, closure, dequantization, and attention costs. Deployment claims require a packed implementation (e.g., Triton) benchmarked against BF16 with matched hardware, weights, batch sizes, context and generation lengths: synchronized repeated timings, peak memory, per-token latency, throughput, and separate prefill measurements. Faster decoding remains untested.

Evaluation-partition separation does not establish absence of pretraining overlap, and quantizer metadata differs. We report inconclusive and reversed results alongside supporting evidence. Cleaner marginal statistics, smaller reconstruction error, and better BF16 quality each leave the tested low-bit MLA selection problem unresolved on their own.

REPRODUCIBILITY STATEMENT

The evaluation partitions and causal schedule are specified in Appendices A–B; training configurations and artifact identifiers appear in Appendix E. The added controls include execution-time file hashes, per-article records, calibration data, and a table-generation script (Appendix H). The retained and excluded historical experiments are distinguished in Appendix I. Additional architecture and recipe specifications appear in Appendix N. Anonymous public artifact hosting remains pending.

AI USAGE DISCLOSURE

Claude (Anthropic) and Codex (OpenAI) assisted with hypotheses, experiment design, implementation, data processing, literature searches, numerical analysis and interpretation, figures, translation of draft material, and manuscript drafting and revision under human direction. The evaluations reported here use public-corpus text, not AI-generated data. The research goals, compute decisions, and critical review were supplied by the authors, who are responsible for the final claims and artifacts.

Implementation corrections and the distinction between original and corrected evaluations are documented in Appendix I.

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. *NeurIPS*, 2024.

Pierfrancesco Beneventano, Mahmoud Abdelmoneum, and Tomaso Poggio. The spectral dynamics and noise geometry of muon. *arXiv preprint arXiv:2606.08388*, 2026.

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

Albert Catalan-Tatjer, Niccolò Ajroldi, and Jonas Geiping. Training dynamics impact post-training quantization robustness. *arXiv preprint arXiv:2510.06213*, 2025.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S. Abdelfattah, and Kai-Chiang Wu. Palu: KV-cache compression with low-rank projection. In *ICLR*, 2025.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient Mixture-of-Experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024b.

DeepSeek-AI. FlashMLA: Efficient MLA decoding kernels. `https://github.com/deepseek-ai/FlashMLA`, 2025.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian aware quantization of neural networks with mixed-precision. In *ICCV*, 2019.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. In *NeurIPS*, 2025.

Georgi Gerganov et al. llama.cpp: LLM inference in C/C++. `https://github.com/ggml-org/llama.cpp`, 2023.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *NeurIPS*, 2024.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for the hidden layers of neural networks. https://github.com/KellerJordan/Muon, 2024.

Junyan Li, Yang Zhang, Muhammad Yusuf Hassan, Talha Chafekar, Tianle Cai, Zhile Ren, Pengsheng Guo, Foroozan Karimzadeh, Colorado Reed, Chong Wang, and Chuang Gan. CommVQ: Commutative vector quantization for KV cache compression. In *ICML*, 2025.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *ICLR*, 2021.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *ICML*, 2024.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *ICML*, 2020.

Jungwoo Park, Taewhoo Lee, Chanwoong Yoon, Hyeon Hwang, and Jaewoo Kang. Outlier-safe pre-training for robust 4-bit quantization of large language models. *ACL*, 2025.

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

Jacob Sander, Brian Jalaian, and Venkat R. Dasari. Advancing model refinement: Muon-optimized distillation and quantization for LLM deployment. *arXiv preprint arXiv:2601.09865*, 2026.

Konrad Staniszewski and Adrian Łańcucki. KVTC: KV cache transform coding for compact storage in LLM inference. *ICLR*, 2026.

Libo Sun, Po-Wei Harn, Zewei Zhang, Peixiong He, and Xiao Qin. Regularize or localize: When training-time KV-cache geometry pays under quantization. *arXiv preprint arXiv:2607.17019*, 2026.

Georgios Vlassis, Saleh Ashkboos, Alexandra Volkova, Torsten Hoefler, and Dan Alistarh. Beyond outliers: A study of optimizers under quantization. *arXiv preprint arXiv:2509.23500*, 2025.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why muon outperforms adam: A curvature perspective. *arXiv preprint arXiv:2606.04662*, 2026.

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 *ICML*, 2023.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. TurboQuant: Online vector quantization with near-optimal distortion rate. *ICLR*, 2026.

A CACHE SCHEDULE AND LOSS BASELINES

For a temporal group of 128 positions, streaming uses a separate absmax scale per latent channel. The current incomplete group remains BF16. At position 128, the first group closes and is quantized before attention for that position. Positions 129–255 append a new BF16 group while the completed first group remains quantized; position 256 closes the second group. Earlier hidden states and predictions are not revised. At each closure all values used to determine the scale have already arrived. This preserves causal information flow, even though precision depends on group age.

Exposure at the scored length. The first 127 of 1,023 predictions see only BF16 entries. Prediction 128 is evaluated after the first group closes, so 896/1,023 predictions (87.59%) attend to at least one quantized entry. At scored query t, the open group has $t \bmod 128$ entries out of t. Its fraction averaged equally over queries is 24.46%; weighting each cache-entry/query pair equally gives 12.41%. These are schedule fractions, not attention-mass fractions or estimates of loss dilution.

Removing the initial 127 structurally unaffected predictions rescales paired streaming damage by $1023/896 = 1.14174$, assuming the identical pre-closure BF16 execution specified by the implementation. Thus streaming-INT3 damage becomes $+0.01190$, -0.00679, and $+0.12365$ for V2-Lite, Coder, and Moonlight. Signs and streaming model ordering are unchanged. This is an algebraic restriction of the measured damage, not a rerun or an always-quantized-cache counterfactual. Removing exactly 128 predictions cannot be recovered from the saved article sums, because the 128th prediction may already be affected.

Tokenwise quantization computes an absmax over 128 channels within a single token. With $q_{\max} = 2^{b-1} - 1$, the nominal symmetric codepoints span $[-q_{\max}, q_{\max}]$; the FP16-rounded scale is used in quantization and reconstruction. The implementation handles a zero scale by a positive floor. The RoPE branch is not quantized. The frozen implementation is the definitive specification of rounding and tensor layout. At latent width 448, channel groups contain 128, 128, 128, and 64 real channels. The last group is padded with 64 zeros for computation and has its own scale; padding is removed after reconstruction. Zero padding does not change the absmax, so the final scale is determined by its 64 real channels. This differs from having four full 128-channel groups.

Tokenwise scores use the full-forward BF16 baseline because channel-group scales have no temporal dependency. Streaming scores use the incremental BF16 baseline (`stream0`). Small full-forward/incremental differences must not be counted as quantization damage. Equal prediction counts make the reported mean article NLL equal to the token-weighted mean for these runs.

B EVALUATION PARTITIONS AND TRAINING-DATA MEMBERSHIP

Production and controlled-twin cache evaluations share the frozen WikiText article partition, with model-specific tokenization for production and GPT-2 for twins. The audit checked exact article text hashes across calibration, development, and confirmation partitions and token hashes across conditions within a checkpoint; the twins also share token hashes across checkpoints. These checks establish matching and exact partition separation only. They are not semantic deduplication or training-corpus membership checks.

Production pretraining corpora are not reconstructed here, so overlap with WikiText is unknown. The twins were trained on FineWeb-Edu, but different corpus names are not proof of non-overlap. No saved training-stream membership index establishes that the evaluation articles are absent from twin training. This is an evidentiary limitation, not a finding of contamination.

FineWeb-Edu pretraining-validation selection is a separate issue from the WikiText cache experiment. Earlier training validation used a stream-selection procedure affected by shuffle/skip ordering. Later deep-row validation was intended to reduce overlap, but a raw row offset together with an approximate row budget does not prove exclusion from a shuffled stream. We therefore report neither a disjointness guarantee nor corrected FineWeb validation NLLs as new evidence in this paper. The selected learning rates define the studied recipes; their global optimality is not established.

Table 8: Post-run direct production contrasts: Moonlight damage minus V2-Lite damage. Interaction rows subtract the streaming contrast from the tokenwise contrast at the same precision. All intervals pair the same 47 articles and are pointwise 95% CIs in model-specific nats/token.

Contrast	Mean	95% CI
Stream INT4	$+0.01744$	$[+0.01267, +0.02281]$
Stream INT3	$+0.09788$	$[+0.08597, +0.11151]$
Token INT4	$+0.00732$	$[+0.00374, +0.01089]$
Token INT3	-0.01340	$[-0.02299, -0.00266]$
Token − stream, INT4	-0.01012	$[-0.01637, -0.00388]$
Token − stream, INT3	-0.11129	$[-0.12055, -0.10239]$

C Remaining production contrasts

The production contrasts in Table 8 subtract changes expressed in each model's own token unit. They are retained to make the previous analysis fully traceable, rather than to rank production robustness or identify a cross-model policy effect. A matched-character-span evaluation would be needed for a common text-normalized interpretation.

Table 9 completes the 18-comparison family: three model pairs, four low-bit conditions, and two token-minus-stream interactions per pair. Resampling occurs within article, including all baseline terms. Pointwise intervals support description of these fixed checkpoints; no familywise discovery claim is made.

Table 9: Remaining post-run production contrasts. For each model pair, damage of the second model is subtracted from that of the first. Interaction rows subtract streaming from tokenwise contrast at the same bit width.

Model pair	Contrast	Mean	95% CI
Coder − V2-Lite	Stream INT4	-0.01423	$[-0.01680, -0.01164]$
Coder − V2-Lite	Stream INT3	-0.01637	$[-0.01947, -0.01330]$
Coder − V2-Lite	Token INT4	-0.02078	$[-0.02527, -0.01628]$
Coder − V2-Lite	Token INT3	-0.05585	$[-0.06520, -0.04643]$
Coder − V2-Lite	Token − stream, INT4	-0.00655	$[-0.01131, -0.00181]$
Coder − V2-Lite	Token − stream, INT3	-0.03948	$[-0.04761, -0.03137]$
Moonlight − Coder	Stream INT4	$+0.03167$	$[+0.02603, +0.03783]$
Moonlight − Coder	Stream INT3	$+0.11425$	$[+0.10176, +0.12845]$
Moonlight − Coder	Token INT4	$+0.02810$	$[+0.02346, +0.03264]$
Moonlight − Coder	Token INT3	$+0.04244$	$[+0.02836, +0.05730]$
Moonlight − Coder	Token − stream, INT4	-0.00357	$[-0.01108, +0.00340]$
Moonlight − Coder	Token − stream, INT3	-0.07181	$[-0.08308, -0.06088]$

D Remaining controlled-twin contrasts

Table 10 includes all six INT4 seed pairs, including near-zero and reversed effects. Table 11 reports the policy interaction within each pair. Positive interaction means the Muon-minus-AdamW damage difference is larger under tokenwise quantization. Each interval conditions on two trained checkpoints and 47 articles. We do not pool article-level observations across seeds as independent training replicates.

E Controlled training and reproducibility

The dense DeepSeek-V3-style twins have nominal sizes 125M/350M/600M, 12/24/28 layers, latent widths 256/384/448, and a GPT-2 tokenizer. Each recipe uses 1B FineWeb-Edu tokens, a cosine schedule with warmup, and gradient clipping at 1.0. Muon uses five-step Newton–Schulz updates for hidden matrices and AdamW for embeddings and norms. Selected peak learning rates are 10^{-3}

Table 10: INT4 excess damage (Muon minus AdamW), with pointwise paired-article 95% CIs for every seed pair.

		Streaming INT4		Tokenwise INT4	
Size	Seed	Excess	95% CI	Excess	95% CI
125M	0	$+0.00189$	$[+0.00034, +0.00351]$	$+0.02636$	$[+0.02298, +0.03011]$
125M	1	$+0.00043$	$[-0.00137, +0.00232]$	$+0.01720$	$[+0.01352, +0.02080]$
125M	2	$+0.00042$	$[-0.00117, +0.00191]$	$+0.02941$	$[+0.02518, +0.03359]$
350M	0	$+0.00262$	$[+0.00107, +0.00413]$	$+0.01632$	$[+0.01332, +0.01939]$
600M	0	$+0.00058$	$[-0.00034, +0.00153]$	-0.00747	$[-0.01015, -0.00484]$
600M	1	-0.00096	$[-0.00222, +0.00040]$	-0.02274	$[-0.02620, -0.01939]$

Table 11: Post-run controlled-twin policy interaction: tokenwise excess damage minus streaming excess damage, where excess means Muon minus AdamW. All intervals are pointwise, with articles paired across the four damage terms.

Size	Seed	Precision	Interaction	95% CI
125M	0	INT4	$+0.02447$	$[+0.02082, +0.02845]$
125M	0	INT3	$+0.18164$	$[+0.17066, +0.19340]$
125M	1	INT4	$+0.01677$	$[+0.01270, +0.02087]$
125M	1	INT3	$+0.14941$	$[+0.13722, +0.16152]$
125M	2	INT4	$+0.02898$	$[+0.02437, +0.03371]$
125M	2	INT3	$+0.57476$	$[+0.52562, +0.62477]$
350M	0	INT4	$+0.01370$	$[+0.01090, +0.01661]$
350M	0	INT3	$+0.42508$	$[+0.39171, +0.46054]$
600M	0	INT4	-0.00805	$[-0.01073, -0.00536]$
600M	0	INT3	-0.21111	$[-0.22261, -0.19958]$
600M	1	INT4	-0.02178	$[-0.02543, -0.01825]$
600M	1	INT3	-0.13188	$[-0.14158, -0.12248]$

for AdamW and 0.01 for Muon hidden matrices, reused across scales. The intervention compares specified recipes, not update rules alone or models matched on loss. The earlier saved causal evaluation ran on Apple MPS; execution-time weight hashes were not recorded. A current on-disk SHA256 manifest now identifies all twelve final checkpoints for future reproduction; it is not a retrospective attestation of the earlier evaluations. The added control runs record model-file hashes at loading, alongside source, input-token, and calibration identities; these attest the new evaluations separately.

Production revisions are listed in Table 12. Run manifests store full source, model-file, and tokenization provenance. The local `mla-cache-reproducibility.zip` combines per-article records, final-quality contrasts, cache-schedule calculations, corrected code, CPU analysis/table scripts, and a table-to-evidence index. Weights and full BF16 tensors are external; package manifests and `MODEL_ACCESS.md` record their identities and access status; their hosting status is given in the reproducibility statement. The recovered original results, implementation copies, arithmetic checks, and experiment-by-experiment evidence ledger are also preserved locally.

Table 12: Production checkpoint revisions and run-manifest SHA256 prefixes. Full manifest hashes are included in the saved analysis report.

Model	Checkpoint revision	Manifest SHA256 prefix
V2-Lite	`604d5664dddd88a0433dbae533b7fe9472482de0`	`e7dc39762c42b724`
Coder	`e434a23f91ba5b4923cf6c9d9a238eb4a08e3a11`	`b6628caa9f73d706`
Moonlight	`4e735b07a89f73647dfab71ab91b840f362ede5b`	`892c91992b28d1d6`

The common article-manifest SHA256 prefix is `4028a8a4ed10b85c`; the frozen causal-reference prefix is `d7730b62ea4c3b46`. The saved-analysis checks cover run completion, case identities, expected coverage, article and token pairing, source/data hashes, and loss-sum reductions. They

recompute the direct contrasts and reproduce the existing twin excess-damage intervals. They do not reload weights, retokenize text, or recertify all historical training and evaluation code.

F GEOMETRY, ACHIEVED FIT, AND TRAINING-SEED DISPERSION

Table 13 joins the original activation descriptors to corrected causal loss measurements for every matched pair. The descriptors are measured on four original 1K chunks, whereas the loss differences use the 47 confirmation articles. Their association is descriptive and cannot identify a common-residual sensitivity coefficient.

Table 13: Geometry and fit for all six pairs. Slash-separated entries are AdamW/Muon. BF16 gap is Muon minus AdamW using the full-forward baseline. Tokenwise and streaming damage use their respective execution baselines.

Size	Seed	Outlier ratio	Token INT3 ΔNLL	Stream INT3 ΔNLL	BF16 gap
125M	0	3.19/1.79	0.04030/0.24407	0.00755/0.02968	-0.09601
125M	1	2.85/1.68	0.04218/0.21873	0.00763/0.03478	-0.10315
125M	2	3.50/1.83	0.04334/0.63642	0.01139/0.02972	-0.09525
350M	0	5.86/1.84	0.07162/0.51151	0.00409/0.01890	-0.09472
600M	0	13.57/1.80	0.28011/0.07261	0.00681/0.01041	-0.13664
600M	1	24.16/1.77	0.21109/0.08285	0.00776/0.01140	-0.08946

Within the six AdamW checkpoints, the descriptive Spearman association between outlier ratio and tokenwise-INT3 damage is 0.886, versus -0.086 for Stream3. The corresponding Muon associations are 0.543 and -0.257. These six-point associations mix architectures, training progress, and seeds; we do not attach a mechanism claim or a discovery p value. The within-size recipe counterexamples in the main text show why the AdamW association cannot be generalized to a universal outlier-based ordering.

One temporal-amplitude proxy can be recovered from the original production channel profiles: the channel-median ratio of maximum to mean absolute activation, averaged across layers, is 5.259/5.192/4.780 for V2-Lite/Coder/Moonlight. Moonlight has the lowest value despite its larger streaming damage. This statistic pools four original 2K chunks; it is neither a 128-token group descriptor nor a realized residual on the confirmation articles. It therefore does not answer the matched-residual question, and we do not treat it as evidence against every temporal-range explanation.

Achieved BF16 loss differs systematically within each pair. Because fit is an outcome of the recipe intervention, it is a possible mediator of its fixed-budget effect, rather than an independently randomized control. The experiments do not identify a direct recipe effect at equal loss. The completed development-only checkpoint selection in Appendix H tests feasibility of a narrower equal-loss comparison, but neither seed meets the 0.01-nat tolerance. Matching checkpoints directly on the reported confirmation losses would introduce additional selection on the evaluation set.

Table 14 summarizes how the excess-damage means vary across the available training pairs. With three, one, and two pairs at the three sizes, these dispersions are descriptive and do not support a training-seed interval.

The nominal 125M models contain about 120.31M stored tensor elements; the other sizes contain 351.15M and 600.01M. Using these counts instead of nominal labels gives approximately 8.31/2.85/1.67 training tokens per stored element. Neither convention matches relative training progress across sizes. The completed fixed-checkpoint group-size control (Table 23) retains the 600M ordering when groups tile width 448 evenly, while narrowing the gap at group size 64. Changing width or token budget together with parameter count still cannot isolate an effect of parameter scale.

Table 14: Training-pair dispersion of Muon-minus-AdamW excess damage. SD is the sample standard deviation across paired training seeds; the bracketed values are the observed minimum and maximum, not a confidence interval. A single pair has no estimable seed SD. Article intervals elsewhere in the paper describe a different source of uncertainty.

Size	Policy	Pairs	Mean	Seed SD	Observed range
125M	Stream4	3	+0.00092	0.00085	$[+0.00042, +0.00189]$
125M	Stream3	3	+0.02253	0.00442	$[+0.01832, +0.02714]$
125M	Token4	3	+0.02432	0.00635	$[+0.01720, +0.02941]$
125M	Token3	3	+0.32447	0.23303	$[+0.17655, +0.59308]$
350M	Stream4	1	+0.00262	–	$[+0.00262, +0.00262]$
350M	Stream3	1	+0.01481	–	$[+0.01481, +0.01481]$
350M	Token4	1	+0.01632	–	$[+0.01632, +0.01632]$
350M	Token3	1	+0.43989	–	$[+0.43989, +0.43989]$
600M	Stream4	2	-0.00019	0.00109	$[-0.00096, +0.00058]$
600M	Stream3	2	+0.00362	0.00003	$[+0.00360, +0.00364]$
600M	Token4	2	-0.01511	0.01080	$[-0.02274, -0.00747]$
600M	Token3	2	-0.16787	0.05605	$[-0.20750, -0.12824]$

G ORIGINAL MEASUREMENTS: COMPLETE RETAINED RECORDS

The recovered original measurements use WikiText-103 *test* text concatenated with blank-line separators and split into fixed token chunks, unlike the whole-article validation partition of the causal revision. Production tokenizers differ. Static FP8 calibration uses chunk 20 and evaluation uses chunks 21–68; these are disjoint token intervals, not a claim of separate underlying articles. Recovered scripts and results are hashed at recovery. Their hashes establish the copied files' identity, not a retrospective guarantee that the current scripts and model revisions are identical to the originals at execution time. No new model run was performed to recover these measurements. Table 15 lists the complete twin statistics and noise measurements.

Table 15: All original twin statistics and noise measurements. Kurtosis, outlier ratio, and offline INT4 temporal-group relative MSE use four 1K chunks. PR is the layer-median participation ratio of the uncentered second moment on two chunks. Noise damage uses eight chunks; uncertainty is saved paired SEM across chunks, not across training seeds or noise draws.

Size	Recipe	Seed	Kurtosis	Outliers	MSE	PR	ΔNLL ($\alpha = .114$)	ΔNLL ($\alpha = .26$)
125M	AdamW	0	10.58	3.19	0.01358	34.08	$+0.00382 \pm 0.00141$	$+0.02355 \pm 0.00504$
125M	AdamW	1	8.50	2.85	0.01328	29.16	$+0.00504 \pm 0.00242$	$+0.02357 \pm 0.00367$
125M	AdamW	2	12.82	3.50	0.01328	37.37	$+0.00633 \pm 0.00207$	$+0.02139 \pm 0.00345$
125M	Muon	0	3.01	1.79	0.00882	5.15	$+0.01937 \pm 0.00319$	$+0.15191 \pm 0.01457$
125M	Muon	1	2.92	1.68	0.00871	3.98	$+0.02192 \pm 0.00563$	$+0.15613 \pm 0.01483$
125M	Muon	2	3.02	1.83	0.00836	3.70	$+0.02699 \pm 0.00454$	$+0.17300 \pm 0.01332$
350M	AdamW	0	15.47	5.86	0.01280	16.19	$+0.00205 \pm 0.00143$	$+0.01762 \pm 0.00441$
350M	Muon	0	3.03	1.84	0.00833	4.14	$+0.01064 \pm 0.00249$	$+0.09083 \pm 0.00970$
600M	AdamW	0	47.53	13.57	0.01054	6.63	$+0.00201 \pm 0.00138$	$+0.01423 \pm 0.00159$
600M	AdamW	1	72.49	24.16	0.01068	4.96	$+0.00343 \pm 0.00111$	$+0.01196 \pm 0.00216$
600M	Muon	0	3.00	1.80	0.00862	3.58	$+0.01137 \pm 0.00192$	$+0.07059 \pm 0.00906$
600M	Muon	1	3.02	1.77	0.00882	3.80	$+0.00921 \pm 0.00254$	$+0.06315 \pm 0.00784$

The production noise probe stores only aggregate perplexities, so the reported log ratios are approximate NLL differences without paired intervals. The statistic captures finite-amplitude response along the perturbed forward trajectory. It does not hold every layer's realized error norm equal across models, and the tested amplitudes should not be interpreted as exact substitutes for INT4 or INT3 residuals. The original residual randomization battery is a different experiment and does use the future-dependent temporal quantizer; it cannot supply a causal explanation of the corrected cache-policy results.

MMLU protection comparison. The same fixed question selection and format were used to protect the first 25% of prompt positions in BF16 and quantize the rest at INT3. Moonlight scores 262/399 (65.66%), versus 269/399 for BF16 and 248/399 for full-prompt INT3. The discordances are 16:9 against BF16 ($p = 0.2295$) and 18:32 against INT3 ($p = 0.0649$). The latter is a favorable point

estimate, not confirmed improvement or equivalence to BF16. These two comparisons complete the eight-test family used above. Saved outcomes preserve question order and agree exactly for the repeated baseline; the original files do not store prompt hashes or predicted letter labels.

Table 16: Original engine context sweep. Paired format settings use the same source text and requested chunk count at each length. ΔNLL is computed as $\log(\mathrm{PPL}_{q4_0}/\mathrm{PPL}_{f16})$ from rounded engine output; no paired chunk interval can be recovered from the saved aggregate log.

Context	Chunks	f16 PPL	q4_0 PPL	ΔNLL
4,096	32	5.7048	5.7941	+0.01553
16,384	8	5.5154	5.6067	+0.01642
32,768	4	5.9023	6.0055	+0.01733

The sweep in Table 16 requests 131,072 input tokens at each length (32/8/4 chunks); this is not a statement that all input tokens enter llama.cpp's perplexity reduction. An earlier separate 24-chunk 4K run reports PPL 5.8880/5.8893/5.9762 for f16/q8_0/q4_0. Its baseline differs from the sweep because the chunk counts differ. The recovered engine checkout is `000547513f1530346ecd163db8b3e13962949961`; the original summary logs do not attest the binary hash. The format measurements do not establish latency, throughput, or the memory frontier of the research policies.

Table 17 reports all 30 original FP8 configurations. Their paired mean changes span -0.00091 to $+0.01152$ nats.

H Completed controls for schedule, geometry, and calibration

The added study re-evaluates all twelve existing final checkpoints on the same frozen confirmation articles. It adds no training seeds and does not change the earlier production, perturbation, MMLU, FP8, or engine panels. Twelve policy jobs each contain 47 articles and ten conditions (5,640 records); four 600M geometry jobs each contain 47 articles and four conditions (752 records); sixteen saved-stage selection jobs each contain eight development articles and one BF16 condition (128 records). All 32 jobs are complete. The 6,520 records include repeated baselines and reference conditions, rather than 6,520 independent samples. The separately completed V2-Lite production extension adds 470 records (Appendix H.5); it does not add training replicates.

H.1 Policies and execution

The ten-condition panel contains full-forward BF16 and tokenwise INT3/INT4 with channel groups of 128, sequential BF16 and streaming INT3/INT4, Tail3/INT4, causal rotation plus a calibrated codebook (RtC3), and asymmetric temporal INT3 (Asym3). Each article has 1,024 input tokens and 1,023 scored predictions. Tokenwise full-forward damage uses its full-forward baseline; all closure-based policies use sequential BF16. The latter retain the same 128-token closure-before-attention schedule, BF16 open tail, and unquantized positional key as the reference.

Matched-tail tokenwise quantization. At each temporal-block closure, Tail3/4 quantize each of its tokens independently in 128-channel groups, using the same narrow-range symmetric integer codes and FP16 scales as the tokenwise reference. The incomplete temporal block remains BF16. Comparing Tail3 with Stream3 therefore holds write timing and open-tail treatment fixed while changing which latent coordinates share a scale. It does not hold the resulting perturbation or its propagation fixed.

Asymmetric and calibrated alternatives. Asym3 stores a per-channel FP16 minimum and FP16 step for each closed 128-token block, then rounds and clips to codes 0–7. RtC3 first applies a fixed orthonormal block-Hadamard transform of width 64, which tiles all three twin latent widths. It stores an FP16 temporal mean and an FP16 centered-maximum scale for each rotated channel, clips normalized values to $[-1, 1]$, assigns the nearest of eight calibrated scalar levels, and inverts the transform. Both reconstructions are returned to BF16. Scales are lower-bounded by the smallest positive normal FP16 value. The RtC codebook is shared across layers within each checkpoint and fitted only on the four calibration articles: 4,096 sampled normalized values per layer per article,

Table 17: All 30 original FP8 configurations. Values are paired mean ΔNLL $\pm$ SEM over 48 chunks, recomputed from saved per-chunk losses. Static block scales are frozen per layer and 128-channel block; static tensor scales use one scalar per layer. In the latter, cal is the calibration scale, lo0.5/hi2.0 multiply it by 0.5/2, and raw uses scale 1. Dynamic block scales are recomputed within each token. These are E4M3 simulations, not throughput measurements of an FP8 serving kernel.

Model	Scale family	Setting	ΔNLL $\pm$ SEM
V2-Lite	static block	m0.9	-0.00004 ± 0.00019
V2-Lite	static block	m0.95	$+0.00031 \pm 0.00019$
V2-Lite	static block	m1.0	-0.00042 ± 0.00019
V2-Lite	static block	m1.05	$+0.00040 \pm 0.00021$
V2-Lite	static block	m1.1	$+0.00036 \pm 0.00018$
V2-Lite	static tensor	cal	$+0.00013 \pm 0.00020$
V2-Lite	static tensor	lo0.5	$+0.00372 \pm 0.00048$
V2-Lite	static tensor	hi2.0	$+0.00005 \pm 0.00021$
V2-Lite	static tensor	raw	-0.00017 ± 0.00018
V2-Lite	dynamic token block	fp8	$+0.00014 \pm 0.00019$
Coder	static block	m0.9	$+0.00336 \pm 0.00049$
Coder	static block	m0.95	-0.00078 ± 0.00057
Coder	static block	m1.0	$+0.00221 \pm 0.00053$
Coder	static block	m1.05	$+0.00088 \pm 0.00056$
Coder	static block	m1.1	$+0.00088 \pm 0.00055$
Coder	static tensor	cal	$+0.00141 \pm 0.00047$
Coder	static tensor	lo0.5	$+0.01152 \pm 0.00103$
Coder	static tensor	hi2.0	$+0.00049 \pm 0.00046$
Coder	static tensor	raw	-0.00026 ± 0.00050
Coder	dynamic token block	fp8	$+0.00036 \pm 0.00045$
Moonlight	static block	m0.9	$+0.00120 \pm 0.00036$
Moonlight	static block	m0.95	$+0.00049 \pm 0.00035$
Moonlight	static block	m1.0	$+0.00060 \pm 0.00031$
Moonlight	static block	m1.05	-0.00091 ± 0.00040
Moonlight	static block	m1.1	$+0.00108 \pm 0.00035$
Moonlight	static tensor	cal	$+0.00003 \pm 0.00040$
Moonlight	static tensor	lo0.5	$+0.00250 \pm 0.00056$
Moonlight	static tensor	hi2.0	-0.00013 ± 0.00040
Moonlight	static tensor	raw	$+0.00130 \pm 0.00039$
Moonlight	dynamic token block	fp8	$+0.00037 \pm 0.00037$

sample seed 20260911, and 40 Lloyd iterations on CPU in FP32. No confirmation article fits the codebook. This is the specified rotation-and-codebook control, not a claim of a bit-identical reproduction of another quantization system.

The symmetric temporal reference has seven reconstruction levels and one 16-bit scale per 128 elements: its ideal closed-block payload is 3.125 bits per element. Asym3 and RtC3 have eight levels and two 16-bit metadata values, giving 3.25 bits per element before the shared RtC codebook. These calculations exclude the BF16 open tail and positional branch. The implementation reconstructs expanded K/V for quality evaluation; the comparison does not measure a packed implementation's memory or speed.

Recorded checks and provenance. Runs use BF16 weights and eager attention on Apple MPS with PyTorch 2.13.0 and Transformers 5.14.1. For every final checkpoint, all seven sequential conditions passed a 257-token development suffix test with zero measured prefix-logit change; Stream3/4 also matched the frozen single-token reference with zero measured logit drift. The accelerated BF16 chunk path failed the 10^{-5} gate, so sequential execution was used. Small-model CPU tests additionally cover 16 policy cases, eight expanded-K/V and positional-key reconstructions, and four rotation widths. These are bounded checks, not exhaustive proofs. Execution identities record model, tokenizer, source, article, and token hashes; calibration records save the codebooks and article identities. The saved-data audit found no coverage, duplicate-record, pairing, nonfinite-loss, loss-reduction, source-identity, calibration-overlap, or recorded-gate inconsistency in the 6,520 completed records.

H.2 All checkpoint outcomes and recipe contrasts

Table 18: Completed policy controls: mean ΔNLL on 47 confirmation articles per checkpoint. All five columns use the same sequential BF16 baseline. Tail3/4 quantize tokenwise only at temporal-block closure; RtC3 and Asym3 use eight levels and additional metadata. Stream3 repeats the original seven-level symmetric reference. These are the same twelve trained checkpoints, not new seeds.

Size	Seed	Recipe	Stream3	Tail3	Tail4	RtC3	Asym3
125M	0	AdamW	0.00755	0.01004	0.00243	0.01823	0.00771
125M	0	Muon	0.02968	0.06256	0.00996	0.02179	0.01031
125M	1	AdamW	0.00763	0.00893	0.00262	0.01795	0.00778
125M	1	Muon	0.03478	0.05687	0.00776	0.02005	0.01244
125M	2	AdamW	0.01139	0.01188	0.00470	0.01449	0.00863
125M	2	Muon	0.02972	0.23897	0.01714	0.02846	0.01140
350M	0	AdamW	0.00409	0.00840	0.00100	0.01155	0.00322
350M	0	Muon	0.01890	0.22805	0.00765	0.02206	0.00923
600M	0	AdamW	0.00681	0.02070	0.00320	0.01378	0.00238
600M	0	Muon	0.01041	0.01278	0.00093	0.01068	0.00341
600M	1	AdamW	0.00776	0.02249	0.00907	0.01066	0.00282
600M	1	Muon	0.01140	0.01181	0.00234	0.01143	0.00266

The repeated symmetric Stream3 condition retains greater Muon damage and better Muon final NLL in every pair; both corresponding paired-article intervals exclude zero in their respective directions. Matched-tail INT3 retains greater Muon damage for all four 125M/350M pairs and lower Muon damage for both 600M pairs. Final-quality ordering differs: Muon is worse for 125M seed 2 and 350M seed 0, and better for the other four pairs. The open-tail schedule attenuates the immediate-tokenwise losses but does not remove their checkpoint dependence.

Asym3 reduces mean damage against Stream3 on each Muon checkpoint by 51.2–76.7%, with the level-count and metadata differences specified above. The recipe ordering is not universal across quantizers: for example, Asym3 has slightly lower mean damage for Muon than AdamW at 600M seed 1. RtC3 improves on Stream3 for all three 125M Muon seeds, but not consistently for the remaining checkpoints. Table 18 reports every checkpoint instead of selecting the most favorable baseline.

Tables 19, 20, 21, and 22 give all six recipe contrasts for each new policy. These post-run descriptive intervals use 10,000 paired-article bootstrap resamples with seed 20260913; they do not increase the number of training replicates or provide simultaneous coverage across the contrast family. All per-

checkpoint means and intervals, including repeated reference conditions, are retained in the analysis data.

Table 19: All six recipe contrasts for `tail3`: mean and pointwise 95% paired-article interval. Excess damage is Muon minus AdamW damage; final NLL is Muon minus AdamW quantized NLL. Negative final NLL favors Muon. Intervals condition on the checkpoints and are not adjusted for the family of descriptive comparisons.

Size	Seed	Excess damage [95% CI]	Final NLL difference [95% CI]
125M	0	$+0.05252$ $[+0.04755, +0.05772]$	-0.04371 $[-0.05446, -0.03292]$
125M	1	$+0.04794$ $[+0.04320, +0.05263]$	-0.05525 $[-0.06759, -0.04216]$
125M	2	$+0.22710$ $[+0.20010, +0.25618]$	$+0.13236$ $[+0.10670, +0.15986]$
350M	0	$+0.21965$ $[+0.20054, +0.24006]$	$+0.12514$ $[+0.10350, +0.14804]$
600M	0	-0.00793 $[-0.01104, -0.00487]$	-0.14454 $[-0.15660, -0.13261]$
600M	1	-0.01068 $[-0.01485, -0.00668]$	-0.09990 $[-0.11187, -0.08766]$

Table 20: All six recipe contrasts for `tail4`: mean and pointwise 95% paired-article interval. Excess damage is Muon minus AdamW damage; final NLL is Muon minus AdamW quantized NLL. Negative final NLL favors Muon. Intervals condition on the checkpoints and are not adjusted for the family of descriptive comparisons.

Size	Seed	Excess damage [95% CI]	Final NLL difference [95% CI]
125M	0	$+0.00753$ $[+0.00594, +0.00912]$	-0.08870 $[-0.09879, -0.07833]$
125M	1	$+0.00514$ $[+0.00319, +0.00706]$	-0.09805 $[-0.10939, -0.08580]$
125M	2	$+0.01244$ $[+0.01009, +0.01482]$	-0.08230 $[-0.09296, -0.07183]$
350M	0	$+0.00665$ $[+0.00500, +0.00835]$	-0.08786 $[-0.09970, -0.07629]$
600M	0	-0.00228 $[-0.00352, -0.00109]$	-0.13889 $[-0.15105, -0.12688]$
600M	1	-0.00674 $[-0.00838, -0.00511]$	-0.09595 $[-0.10740, -0.08452]$

Table 21: All six recipe contrasts for `rtc3`: mean and pointwise 95% paired-article interval. Excess damage is Muon minus AdamW damage; final NLL is Muon minus AdamW quantized NLL. Negative final NLL favors Muon. Intervals condition on the checkpoints and are not adjusted for the family of descriptive comparisons.

Size	Seed	Excess damage [95% CI]	Final NLL difference [95% CI]
125M	0	$+0.00356$ $[+0.00068, +0.00637]$	-0.09267 $[-0.10331, -0.08178]$
125M	1	$+0.00210$ $[-0.00043, +0.00462]$	-0.10109 $[-0.11267, -0.08845]$
125M	2	$+0.01397$ $[+0.01060, +0.01724]$	-0.08077 $[-0.09077, -0.07127]$
350M	0	$+0.01051$ $[+0.00722, +0.01402]$	-0.08401 $[-0.09554, -0.07269]$
600M	0	-0.00310 $[-0.00466, -0.00147]$	-0.13972 $[-0.15170, -0.12787]$
600M	1	$+0.00077$ $[-0.00070, +0.00226]$	-0.08844 $[-0.09934, -0.07755]$

H.3 Evenly tiled channel groups at 600M

Both new group sizes eliminate the 64-channel remainder of grouping width 448 by 128. At group size 112, the number of scales remains four per token, and Muon's lower tokenwise damage persists in both seeds. At group size 64, seven scales reduce damage particularly strongly for AdamW, substantially narrowing the gap while retaining its direction in the means. Thus a remainder group is not necessary for the observed ordering. This control does not independently vary latent width, training progress, or architecture.

H.4 Development-only loss-match feasibility

The selection set is the eight development articles, disjoint by article hash from calibration and confirmation. The panel evaluates the three saved intermediate stages and final checkpoint for each

Table 22: All six recipe contrasts for `asym3`: mean and pointwise 95% paired-article interval. Excess damage is Muon minus AdamW damage; final NLL is Muon minus AdamW quantized NLL. Negative final NLL favors Muon. Intervals condition on the checkpoints and are not adjusted for the family of descriptive comparisons.

Size	Seed	Excess damage [95% CI]	Final NLL difference [95% CI]
125M	0	$+0.00261$ $[-0.00018, +0.00520]$	-0.09362 $[-0.10445, -0.08273]$
125M	1	$+0.00466$ $[+0.00219, +0.00703]$	-0.09854 $[-0.10972, -0.08655]$
125M	2	$+0.00277$ $[+0.00057, +0.00496]$	-0.09197 $[-0.10198, -0.08192]$
350M	0	$+0.00601$ $[+0.00366, +0.00839]$	-0.08850 $[-0.10013, -0.07712]$
600M	0	$+0.00103$ $[-0.00040, +0.00254]$	-0.13558 $[-0.14746, -0.12377]$
600M	1	-0.00016 $[-0.00185, +0.00150]$	-0.08937 $[-0.10054, -0.07825]$

Table 23: Fixed-checkpoint tokenwise INT3 geometry control at 600M: mean damage and pointwise 95% article-bootstrap intervals. All conditions use the same full-forward BF16 baseline and 47 articles. Groups of 64 and 112 tile the latent width 448 evenly; 112 retains four scales per token, as at group size 128, while 64 uses seven.

Seed	Group	AdamW damage [95% CI]	Muon damage [95% CI]
0	128	$+0.28011$ $[+0.26705, +0.29377]$	$+0.07261$ $[+0.06770, +0.07764]$
0	64	$+0.08854$ $[+0.08263, +0.09446]$	$+0.06310$ $[+0.05861, +0.06760]$
0	112	$+0.23130$ $[+0.22076, +0.24201]$	$+0.07280$ $[+0.06809, +0.07754]$
1	128	$+0.21109$ $[+0.20190, +0.22008]$	$+0.08285$ $[+0.07849, +0.08725]$
1	64	$+0.06913$ $[+0.06420, +0.07420]$	$+0.06380$ $[+0.06000, +0.06768]$
1	112	$+0.17432$ $[+0.16620, +0.18237]$	$+0.07349$ $[+0.06976, +0.07733]$

Table 24: All saved-stage selection outcomes at 600M: mean BF16 NLL on eight development articles. Final AdamW is the target within each seed. The nearest available Muon stage is step 11,443 for both seeds, with absolute gaps 0.01783/0.04214 nats, exceeding the specified 0.01-nat tolerance. These are selection losses, not confirmation outcomes.

Recipe	Training step	Seed 0	Seed 1
AdamW	3814	4.61620	4.49193
AdamW	7629	4.13923	4.08068
AdamW	11443	3.87264	3.83800
AdamW	final	3.73116	3.71751
Muon	3814	4.30278	4.29953
Muon	7629	3.96091	3.99800
Muon	11443	3.74899	3.75965
Muon	final	3.61139	3.62986

of four 600M runs. It selects the available Muon stage closest in mean BF16 loss to the same-seed final AdamW target. Step 11,443 is closest for both seeds, but neither gap meets the specified 0.01-nat tolerance. The threshold was not relaxed and no interpolated checkpoint was constructed.

The failure is one of resolution rather than range. In both seeds the AdamW target lies between Muon step 11,443 and the Muon final checkpoint, so the target loss is reached somewhere inside the saved trajectory. Those two stages differ by 0.13760 and 0.12979 nats, about thirteen times the tolerance, so a stage grid of this spacing can land within 0.01 nats only by coincidence. We therefore report the attempt as a negative result: post-hoc stage selection cannot supply an equal-fit comparison here, and a supporting experiment would need a denser checkpoint schedule or a run trained to a specified target loss. No confirmation comparison of these intermediate candidates is reported. This feasibility result leaves achieved fit unresolved as a mediator; it cannot establish a direct recipe effect at equal loss.

H.5 Completed production extension: DeepSeek-V2-Lite

The V2-Lite extension is complete for all 47 confirmation articles and ten conditions (470 records). It uses the same checkpoint revision as Table 41, with identical model-file hashes and scored token sequences, but a separate Apple MPS execution. All within-panel contrasts use the corresponding MPS BF16 baseline. The four added policies follow Appendix H.1; the RtC rotation has width 512 here, rather than the twins' width 64. Its eight-level codebook is fitted on 442,368 normalized values from the four disjoint calibration articles (27 layers, 4,096 samples per layer per article).

Table 25: Completed V2-Lite control panel on Apple MPS: all 47 confirmation articles and ten conditions (470 records). Full-forward/sequential BF16 mean NLLs are 1.910471/1.910788; tokenwise/closure-based damage uses the respective baseline. Intervals are pointwise paired-article 95% bootstrap intervals; wins count negative damage. This panel is separate from the H100 results in Table 41.

Policy	Final NLL	ΔNLL	95% CI	Wins
Token INT3	1.98783	$+0.07736$	$[+0.07203, +0.08299]$	0/47
Token INT4	1.92066	$+0.01018$	$[+0.00816, +0.01228]$	2/47
Stream INT3	1.92014	$+0.00935$	$[+0.00777, +0.01101]$	2/47
Stream INT4	1.90951	-0.00128	$[-0.00232, -0.00024]$	29/47
Tail INT3	1.92726	$+0.01647$	$[+0.01443, +0.01858]$	0/47
Tail INT4	1.91305	$+0.00227$	$[+0.00099, +0.00349]$	13/47
RtC INT3	1.93035	$+0.01956$	$[+0.01811, +0.02100]$	0/47
Asym INT3	1.91872	$+0.00793$	$[+0.00585, +0.00999]$	7/47

Schedule and geometry both matter. In Table 25, matching the BF16 open tail reduces Token3 damage from 0.07736 to 0.01647. Temporal grouping with the same closure schedule further reduces it to 0.00935. The paired Tail3-minus-Stream3 difference is $+0.00712$, with interval $[+0.00474, +0.00947]$. Tail4 also has more damage than Stream4, by $+0.00355$ $[+0.00211, +0.00502]$. These comparisons separate the schedule change from the grouping-axis change at this checkpoint; they do not identify a common-residual sensitivity mechanism.

Additional levels and calibration are not a uniform improvement. RtC3 has greater damage than Stream3 by $+0.01021$ $[+0.00844, +0.01189]$. Asym3's mean damage is lower than Stream3, but its paired difference is -0.00142 $[-0.00413, +0.00118]$, so the interval does not establish an advantage. The eight-level alternatives retain their additional metadata cost. Table 26 reports all four added-policy comparisons with the same-bit temporal reference. Intervals use 10,000 paired-article resamples (seed 20260913) and are pointwise, post-run descriptive intervals without multiplicity adjustment.

Execution dependence near zero. The original H100 panel remains unchanged. Model-file identities agree, and the saved MPS token lists reproduce all 282 H100 token hashes after applying the original compact-JSON serialization. Both executions have greatest mean damage for Token3 and least for Stream4; the closely spaced Token4 and Stream3 means exchange order. Their damage estimates differ between executions by at most 0.00107 nats in magnitude. For Stream4, the MPS estimate is -0.00128 with interval $[-0.00232, -0.00024]$, whereas the H100 interval includes zero

Table 26: All four added V2-Lite policies versus the same-bit symmetric temporal reference on MPS. Negative differences favor the added policy. The baseline cancels within each paired article. Intervals are pointwise and descriptive.

Contrast	Damage difference	95% CI
Tail INT3 − Stream INT3	$+0.00712$	$[+0.00474, +0.00947]$
Tail INT4 − Stream INT4	$+0.00355$	$[+0.00211, +0.00502]$
RtC INT3 − Stream INT3	$+0.01021$	$[+0.00844, +0.01189]$
Asym INT3 − Stream INT3	-0.00142	$[-0.00413, +0.00118]$

(Table 41). Whether the interval excludes zero therefore differs between these execution stacks; we do not infer a portable INT4 quality improvement or pool the two evaluations. Identical checkpoints and inputs do not make these implementations bitwise identical, and this comparison does not isolate hardware from the rest of the execution stack.

Validation and scope. The complete panel passed the saved-data coverage, token-pairing, finite-loss, loss-reduction, calibration-separation, and recorded-source checks. Its seven sequential policies have zero measured prefix-logit change on the 257-token development probe, and Stream3/4 match the frozen reference with zero measured logit drift. The chunked path failed the tolerance gate and was excluded. The completed file SHA256 prefix is `a162409d7738a68b`; full identities, diagnostics, and calculations are retained with the analysis. Coder's completed ten-condition panel follows in Appendix H.6; Moonlight's completed panel is reported separately in Appendix L.

H.6 Completed production extension: DeepSeek-Coder-V2-Lite-Instruct

The Coder extension is complete for all 47 confirmation articles and ten conditions (470 records). It uses the same checkpoint and scored token sequences as the earlier H100 Coder panel, but a separate Apple MPS execution with its own full-forward and sequential BF16 baselines. The four added policies follow Appendix H.1. RtC uses width-512 rotation and eight codepoints fitted on 442,368 values from four disjoint calibration articles, as in the V2-Lite extension.

Table 27: Completed Coder-V2-Lite-Instruct control panel on Apple MPS: all 47 confirmation articles and ten conditions (470 records). Full-forward/sequential BF16 mean NLLs are 2.261971/2.261966; tokenwise/closure-based damage uses the respective baseline. Intervals are pointwise paired-article 95% bootstrap intervals; wins count negative damage. This panel is separate from the H100 results in Table 41.

Policy	Final NLL	ΔNLL	95% CI	Wins
Token INT3	2.28723	$+0.02526$	$[+0.01632, +0.03439]$	12/47
Token INT4	2.25254	-0.00943	$[-0.01335, -0.00574]$	36/47
Stream INT3	2.25415	-0.00781	$[-0.01106, -0.00443]$	35/47
Stream INT4	2.24706	-0.01491	$[-0.01749, -0.01250]$	45/47
Tail INT3	2.26896	$+0.00700$	$[+0.00360, +0.01047]$	13/47
Tail INT4	2.25658	-0.00538	$[-0.00733, -0.00345]$	40/47
RtC INT3	2.28749	$+0.02552$	$[+0.02170, +0.02926]$	2/47
Asym INT3	2.27330	$+0.01133$	$[+0.00416, +0.01886]$	14/47

Policy-dependent damage. In Table 27, Token3 gives mean damage $+0.02526$, Tail3 gives $+0.00700$, and Stream3 gives -0.00781. At the same closure schedule, the Tail3-minus-Stream3 difference is $+0.01481$, with interval $[+0.01100, +0.01835]$. Tail4 also has greater damage than Stream4, by $+0.00952$ $[+0.00669, +0.01243]$. Thus matching the BF16 open tail does not remove the grouping-dependent difference at this checkpoint.

RtC3 and Asym3 have positive mean damage, $+0.02552$ and $+0.01133$, respectively, and greater damage than Stream3 in both paired comparisons (Table 28). The eight-level alternatives therefore do not improve this checkpoint's WikiText NLL over the seven-level temporal reference, despite their additional metadata. The negative Stream3/4 and Token4 damage estimates do not establish improved code generation or identify a regularization mechanism.

Table 28: All four added Coder policies versus the same-bit symmetric temporal reference on MPS. Negative differences favor the added policy. The sequential BF16 baseline cancels within each paired article. Intervals are pointwise and descriptive; Asym3/RtC3 use extra metadata.

Contrast	Damage difference	95% CI
Tail INT3 − Stream INT3	$+0.01481$	$[+0.01100, +0.01835]$
Tail INT4 − Stream INT4	$+0.00952$	$[+0.00669, +0.01243]$
RtC INT3 − Stream INT3	$+0.03334$	$[+0.02905, +0.03757]$
Asym INT3 − Stream INT3	$+0.01915$	$[+0.01160, +0.02675]$

Execution comparison. Model-file identities and all 282 repeated-condition token sequences agree with the H100 panel. Both executions give positive Token3 damage and negative Token4, Stream3, and Stream4 damage. Their mean damage estimates differ by at most 0.00308 nats. For Stream3, MPS-minus-H100 damage is -0.00187 with pointwise interval $[-0.00364, -0.00016]$; the two panels are retained separately rather than pooled or treated as bitwise identical. This comparison does not isolate hardware from the rest of the execution stack.

Validation and scope. The saved records passed coverage, token-pairing, finite-loss, loss-reduction, diagnostic, calibration, and source-identity checks. The recorded sequential gates have zero prefix-logit change, and Stream3/4 match the frozen reference with zero measured drift on the 257-token probe. The failing chunked path was excluded. Intervals use 10,000 paired-article resamples with seed 20260913, conditional on this checkpoint and without multiplicity adjustment. The raw file SHA256 prefix is `0dda194c2f7134e7`; full identities, all 28 pairwise damage contrasts, and table-generation inputs are archived. Moonlight's completed extension is reported separately in Appendix L.

Artifacts and remaining scope. The additive tables are generated from the completed-run audit by `paper/scripts/build_completed_controls_tables.py`, which verifies the hashes of all 32 input files before writing tables. The audit retains record-level provenance and per-checkpoint and paired-article summaries. Per-layer diagnostics store latent energy, squared/absolute/maximum reconstruction error, and temporal-range aggregates; they do not store a common residual intervention or identify attention-weighted sensitivity. The V2-Lite extension and all four added-policy contrasts are generated by `paper/scripts/analyze_v2lite_controls_20260914.py` from its completed, hash-verified raw file. Its analysis also retains all 28 pairwise damage contrasts and the six repeated-condition comparisons with H100. The completed Coder extension is generated by `paper/scripts/build_coder_controls_tables_20260915.py` from its hash-verified raw file and saved-data audit. Moonlight's completed extension and table builder are reported in Appendix L. The corrected code-domain and 4K extensions have not run. Their planned outcomes are not included in these tables. The existing completed H100 production and independent engine results remain separate evidence.

I Implementation corrections and provenance

Validation record and corrections. The original temporal hook allowed later group members to affect earlier next-token predictions; prefix-invariance tests demonstrated this dependency. The causal reference passed independent CPU checks of completed-group reconstruction, positional keys, open groups, and the first closure boundary on small DeepSeek-V2/V3 models, in addition to the three production probes. The H100 accelerated chunked path failed its development gate and was excluded; the evaluated single-token path matched the frozen reference with zero measured drift. These are bounded numerical checks, not exhaustive validation of all inputs. Earlier duplicate-hook, prompt-budget, and training-validation selection issues affect the specific studies identified below or in Appendix B. Arithmetic and frozen source/data hashes were checked; weights were not reloaded during the saved-data analyses.

Current final-checkpoint hashes and metadata were recorded from the saved files on 11 September. The source manifest also lists twelve 600M intermediate checkpoints (three training stages for each of four runs). Development-only selection of these candidates and the final checkpoints is now complete (Table 24). Neither selected Muon candidate meets the 0.01-nat tolerance, and neither in-

termediate checkpoint has been evaluated on the confirmation set in this extension. The weights and an anonymous public artifact are not yet released. Thus the manifest improves future file identification without closing the public reproducibility gap or proving which exact weight bytes produced old results.

The original full-sequence temporal RTN next-token NLL results, including the code-domain study, research long-context curves, early-cache NLL mitigation, temporal rotation/codebook comparisons, and residual battery, depend on within-group future statistics. They are preserved as historical measurements but cannot establish autoregressive quality or explain the corrected policies. Causal production and twin results replace the primary RTN comparisons. The separate unquantized statistics, token-local noise, fixed/dynamic-token FP8, full-prompt MMLU, and engine studies are retained.

Passkey scores were teacher-forced, including the repaired prompt-budget version; they do not establish generated retrieval success. A duplicated RtC hook was repaired in the original work, but removing that duplication does not remove the temporal-scale dependency. The new causal RtC control on all twelve twins and all three completed production extensions is reported in Appendices H and L. This does not validate the earlier future-dependent comparison. The same-model expanded-K statistics include replicated RoPE dimensions, explicitly labeled in Table 40. Tokenwise equalization pilots remain development measurements in the evidence archive.

J INT3 RECONSTRUCTION ERROR AND LOSS DAMAGE ON SAVED TRAJECTORIES

This post-run analysis compares reconstruction error and downstream loss at the same nominal precision. It extends the original offline INT4 statistics in Section M.1 with diagnostics from the completed causal INT3 runs. It reuses saved evaluations, with no additional inference or training replicates. The distinction between relative error, absolute error, damage, and final quality is retained throughout.

J.1 SCOPE, ESTIMANDS, AND UNCERTAINTY

The original analysis includes fourteen completed Apple MPS checkpoint panels: the twelve twins and the V2-Lite and Coder extensions in Appendix H. Each contains the same 47 selected confirmation articles and ten conditions. We check 6,580 raw records and 107,536 layer diagnostic records, and analyze the 3,290 article-policy records and 67,210 layer records for Stream3, Tail3, Token3, RtC3, and Asym3. Repeated conditions and layers are not independent training replicates. Moonlight's subsequently completed extension is reported separately in Appendix L, preserving this panel. These results are not pooled with the earlier H100 panel. Token hashes match within all recipe and policy contrasts. V2-Lite and Coder share tokens with each other, but differ from the twins on all 47 articles; their NLL contrasts remain in separate token units.

For article i, layer ℓ, and policy p, let $S_{i\ell p}$ be the saved sum of squared reconstruction residuals, $U_{i\ell p}$ the sum of squared pre-quantization latent values over the same elements, and $N_{i\ell p}$ their count. Define

$$R_p = \frac{\sum_{i,\ell} S_{i\ell p}}{\sum_{i,\ell} U_{i\ell p}}, \qquad A_p = \frac{\sum_{i,\ell} S_{i\ell p}}{\sum_{i,\ell} N_{i\ell p}}, \qquad E_p = \frac{\sum_{i,\ell} U_{i\ell p}}{\sum_{i,\ell} N_{i\ell p}}, \quad R_p = A_p/E_p. \tag{2}$$

R_p is dimensionless pooled relative MSE, not the unweighted mean of layer ratios; A_p is raw MSE per element, and E_p is latent energy per element. The pre-quantization values are those reached on the policy's own perturbed trajectory. The saved diagnostics accumulate squared differences and energies in FP32. Damage uses Equation 1: closure-based policies subtract sequential BF16, whereas Token3 subtracts its full-forward BF16 baseline. Final NLL remains a separate outcome.

Matched write coverage. Stream3 and Tail3 each record seven closed 128-token blocks, or 896 latent token writes per layer, while scoring 1,023 next-token predictions from a 1,024-token sequence. They share the closure schedule and BF16 open tail. Each block contributes once to reconstruction error, not once per subsequent attention read. Token3 diagnostics instead include all 1,024 forwarded positions, including the last unscored position. Its same-policy recipe comparison is retained, but its aggregate error is not an exposure-matched comparison with Stream3 or Tail3.

Uncertainty and validation. We use 10,000 paired-article percentile bootstrap resamples, seed 20260915, with article order fixed by SHA256. Each resample recomputes the sums and their ratios using the same article indices for both arms. Intervals are pointwise and conditional on the checkpoints, without multiplicity adjustment. Spearman intervals use the first 2,000 paired resamples and average ranks for ties. This new resampling seed does not overwrite the earlier controls' intervals. Source hashes, coverage, pairing, finite diagnostics, loss arithmetic, and recorded gate status pass the saved-data audit. All 70 recomputed INT3 damage means agree with the prior audits within 10^{-12}. Independent reductions check all 70 policy summaries and six primary intervals across representative twin and production contrasts; no inference gate is rerun by this analysis.

J.2 Streaming recipe contrasts and the role of normalization

Table 29: Streaming INT3 recipe contrasts: Muon minus AdamW. R is pooled relative MSE (Equation 2); d is NLL damage in nats per shared twin token. Brackets are pointwise 95% paired-article intervals, conditional on each checkpoint pair (47 articles; 10,000 resamples).

Pair	$R_{\mathrm{Muon}} - R_{\mathrm{AdamW}}$ [95% CI]	$d_{\mathrm{Muon}} - d_{\mathrm{AdamW}}$ [95% CI]
125M s0	-0.02212 $[-0.02230, -0.02194]$	$+0.02213$ $[+0.01895, +0.02529]$
125M s1	-0.02258 $[-0.02273, -0.02242]$	$+0.02714$ $[+0.02338, +0.03109]$
125M s2	-0.02337 $[-0.02354, -0.02320]$	$+0.01832$ $[+0.01488, +0.02197]$
350M s0	-0.02299 $[-0.02315, -0.02283]$	$+0.01481$ $[+0.01151, +0.01790]$
600M s0	-0.01621 $[-0.01631, -0.01610]$	$+0.00360$ $[+0.00132, +0.00594]$
600M s1	-0.01360 $[-0.01371, -0.01348]$	$+0.00364$ $[+0.00141, +0.00579]$

Muon has 22.0–33.4% lower pooled relative MSE under Stream3 in all six pairs, yet greater NLL damage. Every relative-MSE difference interval is negative and every damage-difference interval is positive (Table 29). Muon nevertheless retains lower final streaming NLL in each pair because its BF16 starting point is better. This extends the error–damage ordering disagreement to the same precision and confirmation articles; it does not equate relative distortion with absolute perturbation size.

Table 30: Streaming INT3 normalization and layer coverage. All ratios are Muon/AdamW: relative error R, raw MSE per element A (with pointwise 95% CI), and latent energy per element E. The final column counts layers with lower Muon pooled relative MSE; no layers are excluded. Ratios compare learned coordinate systems, not matched-norm perturbations.

Pair	R ratio	A ratio [95% CI]	E ratio	Layers
125M s0	0.689	0.855 $[0.854, 0.857]$	1.241	12/12
125M s1	0.682	0.857 $[0.855, 0.859]$	1.256	12/12
125M s2	0.670	0.836 $[0.835, 0.838]$	1.248	12/12
350M s0	0.666	1.310 $[1.307, 1.313]$	1.967	24/24
600M s0	0.744	1.825 $[1.819, 1.830]$	2.454	19/28
600M s1	0.780	2.158 $[2.152, 2.164]$	2.766	21/28

The raw-error comparison changes with size (Table 30). At 125M, Muon's raw MSE is 0.836–0.857 times AdamW's. At 350M and the two 600M checkpoints, it is instead 1.310, 1.825, and 2.158 times AdamW's, while latent energy ratios are 1.967, 2.454, and 2.766. Lower normalized reconstruction error therefore does not mean universally smaller absolute errors. Cross-recipe raw scales also refer to different learned latent coordinate systems; dividing damage by either error metric does not identify a causal sensitivity coefficient.

The relative-MSE ordering survives both equal-article weighting of article-relative MSE and equal-layer weighting of article-pooled layer ratios for every streaming pair. All layers have lower Muon relative MSE at 125M/350M; 19/28 and 21/28 layers do at 600M. The remaining layers are included. Saved profiles characterize reconstruction error by layer but do not allocate the model's total loss damage to particular layers.

Table 31: Matched-tail minus streaming INT3 at each fixed checkpoint. R is relative MSE; damage differences use the same sequential BF16 baseline and scored tokens. Both policies quantize the same seven closed blocks (896 latent writes per layer). The production block has different tokens from the twins; no cross-block NLL ranking is intended. All brackets are pointwise 95% paired-article intervals.

Checkpoint	$R_{\text{Tail}}/R_{\text{Stream}}$	$R_{\text{Tail}} - R_{\text{Stream}}$ [95% CI]	$d_{\text{Tail}} - d_{\text{Stream}}$ [95% CI]
125M s0 AdamW	1.394	$+0.02807$ $[+0.02769, +0.02846]$	$+0.00249$ $[+0.00067, +0.00433]$
125M s0 Muon	1.555	$+0.02722$ $[+0.02707, +0.02738]$	$+0.03288$ $[+0.02765, +0.03840]$
125M s1 AdamW	1.379	$+0.02693$ $[+0.02656, +0.02730]$	$+0.00130$ $[-0.00070, +0.00324]$
125M s1 Muon	1.502	$+0.02431$ $[+0.02417, +0.02446]$	$+0.02209$ $[+0.01738, +0.02707]$
125M s2 AdamW	1.381	$+0.02702$ $[+0.02659, +0.02746]$	$+0.00049$ $[-0.00166, +0.00262]$
125M s2 Muon	1.586	$+0.02781$ $[+0.02764, +0.02799]$	$+0.20926$ $[+0.18244, +0.23788]$
350M s0 AdamW	1.922	$+0.06339$ $[+0.06260, +0.06417]$	$+0.00431$ $[+0.00286, +0.00580]$
350M s0 Muon	1.649	$+0.02970$ $[+0.02953, +0.02987]$	$+0.20916$ $[+0.19023, +0.22955]$
600M s0 AdamW	2.182	$+0.07474$ $[+0.07388, +0.07561]$	$+0.01390$ $[+0.01159, +0.01638]$
600M s0 Muon	1.581	$+0.02730$ $[+0.02710, +0.02750]$	$+0.00237$ $[-0.00023, +0.00493]$
600M s1 AdamW	2.427	$+0.08821$ $[+0.08721, +0.08925]$	$+0.01473$ $[+0.01194, +0.01756]$
600M s1 Muon	1.551	$+0.02655$ $[+0.02635, +0.02677]$	$+0.00041$ $[-0.00227, +0.00308]$
Production token unit (separate from twins)			
V2-Lite	2.688	$+0.11361$ $[+0.11237, +0.11487]$	$+0.00712$ $[+0.00474, +0.00947]$
Coder	2.458	$+0.10095$ $[+0.09990, +0.10200]$	$+0.01481$ $[+0.01113, +0.01837]$

J.3 Matched-tail versus streaming at fixed checkpoints

With closure timing and write coverage held fixed, Tail3 has 1.38–2.69 times Stream3 relative MSE across all fourteen checkpoints. All relative-MSE difference intervals are positive. Mean Tail3 damage is also higher in every checkpoint, with ten intervals above zero and four including zero: AdamW 125M seeds 1/2 and Muon 600M seeds 0/1 (Table 31). This comparison orders error and damage in the same direction on average; it is not itself a lower-error, higher-damage reversal. It shows that grouping differences remain relevant under the same BF16 open-tail schedule.

Table 32: Matched-tail INT3 recipe contrasts: Muon minus AdamW. Relative-MSE differences are negative in every pair, whereas damage differences are positive at 125M/350M and negative at 600M. All pointwise paired-article intervals exclude zero in their respective directions.

Pair	$R_{\text{Muon}} - R_{\text{AdamW}}$ [95% CI]	$d_{\text{Muon}} - d_{\text{AdamW}}$ [95% CI]
125M s0	-0.02296 $[-0.02341, -0.02253]$	$+0.05252$ $[+0.04756, +0.05777]$
125M s1	-0.02519 $[-0.02561, -0.02478]$	$+0.04794$ $[+0.04318, +0.05252]$
125M s2	-0.02258 $[-0.02308, -0.02206]$	$+0.22710$ $[+0.20004, +0.25594]$
350M s0	-0.05668 $[-0.05750, -0.05584]$	$+0.21965$ $[+0.20051, +0.24037]$
600M s0	-0.06365 $[-0.06450, -0.06281]$	-0.00793 $[-0.01112, -0.00489]$
600M s1	-0.07525 $[-0.07622, -0.07431]$	-0.01068 $[-0.01480, -0.00666]$

Under Tail3, Muon again has lower relative MSE in every recipe pair, but greater damage at 125M/350M and lower damage at 600M (Table 32). All six damage intervals exclude zero in their respective directions. The different recipe orderings under Stream3 and Tail3 reinforce the policy dependence already reported in Appendix H; they do not identify parameter scale, group geometry, or achieved fit as an isolated cause.

J.4 Within-checkpoint article associations

For this diagnostic only, article relative MSE pools the layer sums within each article. We correlate it with that article's damage, then correlate Tail3-minus-Stream3 changes in both quantities (Table 33). These latter Spearman coefficients range from -0.378 to $+0.117$. No positive interval excludes zero; one negative interval does (AdamW 600M seed 0), without multiplicity adjustment. The small, mixed associations describe variation within these 47-article panels. They do not demon-

Table 33: Within-checkpoint Spearman associations across 47 articles. The first two columns correlate article relative MSE and damage within one policy; the last correlates their Tail3-minus-Stream3 differences. Only the last column displays intervals (2,000 paired-article resamples; pointwise, unadjusted). No model or layer pooling is used.

Checkpoint	Stream3 ρ	Tail3 ρ	Difference ρ [95% CI]
125M s0 AdamW	-0.231	-0.160	-0.171 $[-0.491, +0.172]$
125M s0 Muon	-0.037	-0.022	-0.183 $[-0.474, +0.132]$
125M s1 AdamW	0.157	-0.218	-0.119 $[-0.389, +0.182]$
125M s1 Muon	-0.143	0.159	-0.163 $[-0.490, +0.160]$
125M s2 AdamW	0.119	-0.302	-0.152 $[-0.430, +0.154]$
125M s2 Muon	0.180	0.006	$+0.012$ $[-0.294, +0.295]$
350M s0 AdamW	0.112	0.049	-0.031 $[-0.320, +0.279]$
350M s0 Muon	0.160	0.198	-0.134 $[-0.421, +0.175]$
600M s0 AdamW	0.144	-0.355	-0.378 $[-0.620, -0.082]$
600M s0 Muon	0.159	-0.085	$+0.117$ $[-0.169, +0.407]$
600M s1 AdamW	0.104	-0.332	-0.299 $[-0.559, +0.010]$
600M s1 Muon	-0.010	-0.013	-0.015 $[-0.285, +0.270]$
Production token unit (separate from twins)			
V2-Lite	0.038	-0.106	-0.076 $[-0.350, +0.228]$
Coder	-0.304	0.017	-0.012 $[-0.298, +0.272]$

strate absence of an error–quality relationship under other perturbations or establish out-of-sample predictive failure. Neither models nor layers are pooled as independent observations.

What the saved diagnostics establish. The streaming recipe result is not solely an INT4-error/INT3-damage comparison. Relative reconstruction MSE alone does not preserve the observed recipe ordering of INT3 damage, and the matched-tail results separately retain a grouping-policy effect at matched write coverage. These conclusions remain conditional on the specified checkpoints, quantizers, and evaluation sample.

What remains unresolved. Each quantized policy changes later activations, so its recorded residuals are evaluated on its own realized trajectory. The archives inspected for this saved-data analysis contain aggregate diagnostics and one-dimensional channel profiles, not the token-by-channel BF16 tensors needed for an identical-input reconstruction comparison. First-layer Tail3/Stream3 energy aggregates agree on all 47 articles for every checkpoint; equality of summaries is not proof of identical tensors throughout the model. This analysis does not isolate equal-norm sensitivity, fit mediation, residual direction, or Hessian/attention mechanisms. It adds neither a new task nor a longer-context experiment. A subsequent capture control supplies identical BF16 reference inputs for both quantizers within each of the twelve twins (Appendix K); its reconstruction results are reported separately from these realized-trajectory diagnostics.

J.5 Secondary quantizers and complete recipe contrasts

In Table 34, RtC3 lowers relative MSE in all fourteen checkpoints but has greater mean damage than Stream3 in eleven; eight of those intervals are positive. Of its three negative mean differences, two intervals are negative. Asym3 also lowers relative MSE in all fourteen, with eleven lower mean damages (nine negative intervals) and three higher means (one positive interval, Coder). These outcomes retain improvements, reversals, and inconclusive comparisons. RtC3 and Asym3 change codebooks and metadata in addition to reconstruction error; they do not isolate rotation, offset, or equal-storage effects.

Table 35 retains the remaining eighteen same-policy recipe contrasts, completing all six pairs for each of the five INT3 policies. The error coverage and BF16-path distinction for Token3 are specified in Appendix J.1. These results supplement, rather than replace, the full final-NLL and damage tables in Appendix H.

Reproduction. The saved analysis includes source hashes, article-policy and article-layer records, pooled layer profiles, all 70 policy summaries, 30 recipe contrasts, 14 primary policy contrasts, and 28 secondary policy contrasts. A source-verified table builder reads `analysis.json` with

Table 34: All secondary INT3 policy comparisons at fixed checkpoints. Error ratios divide each added policy by Stream3; damage differences subtract Stream3. Brackets are pointwise 95% paired-article intervals. RtC3 and Asym3 have eight levels and additional metadata versus seven-level Stream3, so these are not matched-storage interventions. Production rows remain in their own token unit.

Checkpoint	Relative MSE ratio	Raw MSE ratio	Damage difference [95% CI]
RtC3 minus Stream3			
125M s0 AdamW	0.413	0.413	$+0.01068$ $[+0.00788, +0.01344]$
125M s0 Muon	0.349	0.349	-0.00789 $[-0.01130, -0.00445]$
125M s1 AdamW	0.418	0.417	$+0.01032$ $[+0.00774, +0.01301]$
125M s1 Muon	0.334	0.334	-0.01473 $[-0.01859, -0.01080]$
125M s2 AdamW	0.416	0.416	$+0.00310$ $[+0.00043, +0.00575]$
125M s2 Muon	0.325	0.325	-0.00126 $[-0.00518, +0.00277]$
350M s0 AdamW	0.402	0.401	$+0.00747$ $[+0.00514, +0.00978]$
350M s0 Muon	0.302	0.302	$+0.00316$ $[-0.00072, +0.00711]$
600M s0 AdamW	0.388	0.388	$+0.00697$ $[+0.00465, +0.00943]$
600M s0 Muon	0.302	0.302	$+0.00027$ $[-0.00272, +0.00342]$
600M s1 AdamW	0.388	0.387	$+0.00290$ $[+0.00016, +0.00556]$
600M s1 Muon	0.309	0.309	$+0.00003$ $[-0.00230, +0.00239]$
Production token unit (separate from twins)			
V2-Lite	0.383	0.387	$+0.01021$ $[+0.00842, +0.01198]$
Coder	0.381	0.387	$+0.03334$ $[+0.02904, +0.03756]$
Asym3 minus Stream3			
125M s0 AdamW	0.558	0.558	$+0.00016$ $[-0.00212, +0.00248]$
125M s0 Muon	0.451	0.451	-0.01937 $[-0.02266, -0.01600]$
125M s1 AdamW	0.561	0.561	$+0.00015$ $[-0.00188, +0.00215]$
125M s1 Muon	0.438	0.438	-0.02234 $[-0.02588, -0.01888]$
125M s2 AdamW	0.562	0.562	-0.00276 $[-0.00521, -0.00032]$
125M s2 Muon	0.426	0.426	-0.01831 $[-0.02221, -0.01452]$
350M s0 AdamW	0.544	0.543	-0.00086 $[-0.00226, +0.00048]$
350M s0 Muon	0.397	0.397	-0.00966 $[-0.01246, -0.00681]$
600M s0 AdamW	0.525	0.525	-0.00443 $[-0.00592, -0.00297]$
600M s0 Muon	0.402	0.402	-0.00699 $[-0.00923, -0.00480]$
600M s1 AdamW	0.523	0.522	-0.00495 $[-0.00720, -0.00269]$
600M s1 Muon	0.412	0.412	-0.00875 $[-0.01098, -0.00648]$
Production token unit (separate from twins)			
V2-Lite	0.519	0.526	-0.00142 $[-0.00422, +0.00123]$
Coder	0.519	0.530	$+0.01915$ $[+0.01179, +0.02671]$

Table 35: Remaining same-policy INT3 recipe contrasts, retaining all six pairs for every policy. Damage uses each policy's matching BF16 path. Token3 (full-position) reconstruction diagnostics cover 1,024 forwarded positions; its damage scores 1,023 predictions. Its reconstruction coverage is not matched to the 896 writes of closure-based policies. Brackets are pointwise 95% paired-article intervals.

Pair	$R_{\mathrm{Muon}} - R_{\mathrm{AdamW}}$ [95% CI]	$d_{\mathrm{Muon}} - d_{\mathrm{AdamW}}$ [95% CI]
Token3 (full-position diagnostics): Muon minus AdamW		
125M s0	-0.02266 $[-0.02306, -0.02227]$	$+0.20377$ $[+0.19256, +0.21590]$
125M s1	-0.02483 $[-0.02522, -0.02444]$	$+0.17655$ $[+0.16386, +0.18932]$
125M s2	-0.02243 $[-0.02292, -0.02192]$	$+0.59308$ $[+0.54392, +0.64467]$
350M s0	-0.05616 $[-0.05694, -0.05536]$	$+0.43989$ $[+0.40546, +0.47591]$
600M s0	-0.06098 $[-0.06165, -0.06033]$	-0.20750 $[-0.21956, -0.19550]$
600M s1	-0.07333 $[-0.07409, -0.07260]$	-0.12824 $[-0.13764, -0.11892]$
RtC3: Muon minus AdamW		
125M s0	-0.01232 $[-0.01242, -0.01222]$	$+0.00356$ $[+0.00077, +0.00635]$
125M s1	-0.01346 $[-0.01355, -0.01337]$	$+0.00210$ $[-0.00035, +0.00457]$
125M s2	-0.01406 $[-0.01417, -0.01394]$	$+0.01397$ $[+0.01053, +0.01726]$
350M s0	-0.01380 $[-0.01390, -0.01369]$	$+0.01051$ $[+0.00723, +0.01409]$
600M s0	-0.01037 $[-0.01044, -0.01030]$	-0.00310 $[-0.00468, -0.00151]$
600M s1	-0.00909 $[-0.00917, -0.00902]$	$+0.00077$ $[-0.00068, +0.00227]$
Asym3: Muon minus AdamW		
125M s0	-0.01764 $[-0.01778, -0.01750]$	$+0.00261$ $[-0.00011, +0.00522]$
125M s1	-0.01865 $[-0.01877, -0.01853]$	$+0.00466$ $[+0.00221, +0.00707]$
125M s2	-0.01958 $[-0.01975, -0.01941]$	$+0.00277$ $[+0.00057, +0.00502]$
350M s0	-0.01923 $[-0.01939, -0.01907]$	$+0.00601$ $[+0.00368, +0.00839]$
600M s0	-0.01426 $[-0.01436, -0.01415]$	$+0.00103$ $[-0.00043, +0.00259]$
600M s1	-0.01250 $[-0.01262, -0.01238]$	-0.00016 $[-0.00189, +0.00151]$

SHA256 prefix `e80e2ff7016c87c5`; all tables here preserve its estimates and intervals. The artifact inventory and scripts accompany the local research archive; artifact hosting status is given in the reproducibility statement.

K INT3 RECONSTRUCTION ON IDENTICAL BF16 REFERENCE INPUTS

K.1 REFERENCE CAPTURE AND PAIRED COMPARISON

Appendix J measures INT3 residuals on each quantized policy's realized trajectory. We add a separate control that applies both quantizers to the same unquantized reference tensors within each checkpoint and article. All twelve trained twins are complete: six AdamW/Muon pairs at 125M (seeds 0–2), 350M (seed 0), and 600M (seeds 0–1), with the same 47 confirmation articles and token sequences as the completed causal controls. This yields 564 BF16 captures and 1,128 article–policy reconstruction records. The corresponding V2-Lite and Coder reference captures are not yet complete and contribute no rows here; their previously completed policy evaluations remain in Appendices H.5 and H.6.

For each article, an unquantized BF16 full-forward pass captures the post-normalization MLA latent entering the cache at every layer. Capture hooks leave model outputs unchanged. Full 1,024-position tensors are archived, and reconstruction is measured only on the first seven complete 128-token blocks (896 positions), matching the closed-write coverage in Appendix J.1. Stream3 shares one scale per channel across each 128-token block; Tail3 shares one scale per 128-channel group within each token of those same blocks. Both use the reference seven-level symmetric INT3 quantizer, FP16 scales, and BF16 reconstruction. At width 448, Tail3 retains the original 64-real-channel final group and zero-padding convention; this control does not change the tiling. Neither reconstructed tensor is fed back into the model.

We denote reference-tensor relative MSE by R^{B} and raw MSE per element by A^{B}. They use the pooled sums in Equation 2, now evaluated on the captured BF16 tensors: sum squared reconstruction error over all measured positions, channels, layers, and articles, then divide by total reference

energy for R^{B} or the number of real elements for A^{B}. Padding is excluded from both denominators. Quantization runs on Apple MPS; squared error and energy are reduced on CPU in float64. Stream3 and Tail3 have identical input hashes, reference energy, and element counts within each checkpoint/article/layer. Different recipes still have different learned activations and energies.

The damage column d remains the previously saved causal Stream3/Tail3 evaluation, with its original sequential BF16 baseline and 1,023 scored next-token predictions per article. We do not subtract the new full-forward reference NLL from old quantized NLL. Thus the control tests whether reconstruction orderings persist without quantization feedback in the error measurement; it does not repeat the loss intervention on these fixed tensors. For all contrasts, 10,000 paired-article bootstrap resamples (seed 20260915) give percentile 95% intervals, recomputing pooled error ratios within each resample. Intervals are pointwise, unadjusted for multiplicity, and conditional on the fixed checkpoints; they do not quantify training-seed uncertainty.

Execution validation. All 564 new full-forward BF16 article NLLs exactly match the saved full-forward BF16 baselines. Every checkpoint passes a separate development-prefix gate: capture versus uncaptured logits have zero measured maximum absolute difference, and changing tokens after the prefix leaves its logits and captured latents exactly unchanged. Full-forward BF16 is not claimed bitwise equivalent to sequential BF16. Matching full-forward baselines validates the new reference execution without changing the old damage baseline.

All 564 archived activation-file hashes were verified. An independent explicit-slice quantizer and `math.fsum` reduction reproduce 464 layer–policy metric sets, covering all layers and both policies for one article selected by smallest article hash at each checkpoint. This is a sampled independent arithmetic check, not a second computation over every article. Raw results, paired article summaries, source hashes, analysis code, and table-generation code accompany the control; full BF16 tensors remain archived on the execution host; artifact hosting status is given in the reproducibility statement.

K.2 Streaming recipe ordering persists on BF16 inputs

Across all six pairs, Muon's streaming relative reconstruction MSE is 21.8–33.6% lower than AdamW's on the BF16 reference captures (Table 4). Every ratio interval lies below one, while all six saved Muon-minus-AdamW damage intervals lie above zero. This retains the same-precision recipe ordering after removing the policy-dependent activation trajectory from the reconstruction measurement. The new percentages describe this reference control; the 22.0–33.4% range in Appendix J describes the distinct realized-trajectory diagnostics.

Normalization remains consequential: Muon's raw streaming MSE is lower only at 125M and is 1.31 times AdamW's at 350M and 1.83/2.17 times at 600M seeds 0/1. The relative-error result therefore does not establish smaller absolute perturbations at all scales or greater loss response at equal perturbation norm. It establishes that relative reconstruction error alone does not preserve the observed recipe ordering of INT3 damage in this fixed panel, even when error is measured without quantization feedback.

Interpretation. The added evidence addresses the identical-input reconstruction gap for the twelve twins. It does not match activations across recipes, equalize perturbation norms, mediate achieved BF16 fit, or identify residual direction, attention, gradient, or Hessian mechanisms. No new task or longer-context result is implied. The geometry controls, final-NLL comparisons, noise probes, and task/engine results elsewhere in the paper retain their separate roles and evidence limits.

K.3 Grouping comparison and all matched-tail recipe pairs

At every one of the twelve checkpoints, Tail3 has greater reference reconstruction error than Stream3: the relative-MSE ratio is 1.38–2.46, with every ratio interval above one and every difference interval above zero (Table 36). Because the two quantizers share the same BF16 inputs, energy denominator, and measured positions within each checkpoint, this ordering also holds for raw MSE. The previously measured Tail3-minus-Stream3 damage is positive in mean at all twelve checkpoints; eight intervals are positive and four include zero (125M AdamW seeds 1/2 and 600M Muon seeds 0/1). This is a twelve-twin comparison, distinct from the fourteen-checkpoint panel in Appendix J that also includes the production models' realized-trajectory diagnostics.

Table 36: Matched-tail versus streaming INT3 on byte-identical BF16 inputs within each checkpoint/article. R^{B} pools error over the same 896 positions in every layer. T/S denote Tail3/Stream3; A/M denote AdamW/Muon. All reconstruction-error intervals favor Stream3; eight saved damage-difference intervals are positive and four include zero. Damage retains the original sequential BF16 baseline. Brackets are pointwise 95% paired-article intervals.

Checkpoint	$R^{\mathrm{B}}_{\mathrm{T}}/R^{\mathrm{B}}_{\mathrm{S}}$ [95% CI]	$R^{\mathrm{B}}_{\mathrm{T}} - R^{\mathrm{B}}_{\mathrm{S}}$ [95% CI]	$d_{\mathrm{T}} - d_{\mathrm{S}}$ [95% CI]
125M s0 A	$1.395\ [1.389, 1.400]$	$+0.02809\ [+0.02771, +0.02848]$	$+0.00249\ [+0.00067, +0.00433]$
125M s0 M	$1.550\ [1.546, 1.554]$	$+0.02711\ [+0.02697, +0.02725]$	$+0.03288\ [+0.02765, +0.03840]$
125M s1 A	$1.380\ [1.374, 1.385]$	$+0.02695\ [+0.02658, +0.02732]$	$+0.00130\ [-0.00070, +0.00324]$
125M s1 M	$1.492\ [1.488, 1.496]$	$+0.02404\ [+0.02390, +0.02418]$	$+0.02209\ [+0.01738, +0.02707]$
125M s2 A	$1.383\ [1.377, 1.389]$	$+0.02713\ [+0.02669, +0.02756]$	$+0.00049\ [-0.00166, +0.00262]$
125M s2 M	$1.609\ [1.604, 1.614]$	$+0.02871\ [+0.02855, +0.02889]$	$+0.20926\ [+0.18244, +0.23788]$
350M s0 A	$1.933\ [1.921, 1.945]$	$+0.06417\ [+0.06334, +0.06499]$	$+0.00431\ [+0.00286, +0.00580]$
350M s0 M	$1.659\ [1.653, 1.666]$	$+0.03012\ [+0.02993, +0.03032]$	$+0.20916\ [+0.19023, +0.22955]$
600M s0 A	$2.201\ [2.186, 2.217]$	$+0.07596\ [+0.07504, +0.07689]$	$+0.01390\ [+0.01159, +0.01638]$
600M s0 M	$1.582\ [1.575, 1.588]$	$+0.02734\ [+0.02715, +0.02754]$	$+0.00237\ [-0.00023, +0.00493]$
600M s1 A	$2.462\ [2.441, 2.483]$	$+0.09034\ [+0.08918, +0.09155]$	$+0.01473\ [+0.01194, +0.01756]$
600M s1 M	$1.547\ [1.541, 1.554]$	$+0.02645\ [+0.02625, +0.02665]$	$+0.00041\ [-0.00227, +0.00308]$

For Tail3, Muon has lower reference relative MSE in all six recipe pairs, but saved damage remains greater at 125M/350M and lower at 600M (Table 37). All six relative-error ratio intervals are below one; damage intervals exclude zero in their respective directions. Raw Tail3 MSE, like raw streaming MSE, is lower for Muon at 125M and higher at 350M/600M. We retain the 600M reversal: fixed BF16 inputs do not turn reconstruction quality into a universal recipe-level predictor of downstream damage.

Table 37: Matched-tail INT3 recipe contrasts on captured BF16 reference tensors. Error ratios compare Muon with AdamW; damage differences retain the original causal evaluations and sequential baselines. All six relative-error ratio intervals are below one, while damage intervals are positive at 125M/350M and negative at 600M. Brackets are pointwise 95% paired-article intervals.

Pair	$R^{\mathrm{B}}_{\mathrm{M}}/R^{\mathrm{B}}_{\mathrm{A}}$ [95% CI]	$A^{\mathrm{B}}_{\mathrm{M}}/A^{\mathrm{B}}_{\mathrm{A}}$ [95% CI]	$d_{\mathrm{M}} - d_{\mathrm{A}}$ [95% CI]
125M s0	$0.770\ [0.766, 0.773]$	$0.955\ [0.953, 0.958]$	$+0.05252\ [+0.04756, +0.05777]$
125M s1	$0.744\ [0.741, 0.747]$	$0.935\ [0.932, 0.937]$	$+0.04794\ [+0.04318, +0.05252]$
125M s2	$0.774\ [0.770, 0.778]$	$0.967\ [0.964, 0.970]$	$+0.22710\ [+0.20004, +0.25594]$
350M s0	$0.570\ [0.567, 0.574]$	$1.124\ [1.120, 1.128]$	$+0.21965\ [+0.20051, +0.24037]$
600M s0	$0.534\ [0.531, 0.538]$	$1.311\ [1.307, 1.316]$	$-0.00793\ [-0.01112, -0.00489]$
600M s1	$0.491\ [0.488, 0.495]$	$1.362\ [1.356, 1.368]$	$-0.01068\ [-0.01480, -0.00666]$

L Completed Moonlight production controls

The Moonlight-16B-A3B-Instruct extension is complete for all 47 confirmation articles and ten conditions (470 records). Together with V2-Lite and Coder (Appendices H.5–H.6), the three completed production extensions contain 1,410 records. This Apple MPS panel uses the same recorded checkpoint and scored tokens as the earlier H100 Moonlight panel, with its own full-forward and sequential BF16 baselines. The additional policies follow Appendix H.1; RtC uses width-512 rotation and an eight-level codebook fitted on 442,368 values from four disjoint calibration articles. No additional inference was needed for this saved-data analysis.

L.1 Matched-tail and calibrated-policy results

The grouping ordering reverses. In Table 38, Token3 has mean damage $+0.06006$, Tail3 $+0.03481$, and Stream3 $+0.10926$. Tail3 reduces mean Stream3 damage by 68.1%. The paired Tail3-minus-Stream3 difference is $-0.07445\ [-0.08317, -0.06618]$, and all 47 articles favor Tail3. Tail4 also has lower mean damage than Stream4, by $-0.00741\ [-0.01267, -0.00242]$. Both comparisons hold the closure schedule and BF16 open tail fixed. Thus the lower-damage grouping

Table 38: Completed Moonlight-16B-A3B-Instruct control panel on Apple MPS: 47 confirmation articles and ten conditions (470 records). Full-forward/sequential BF16 mean NLLs are 2.057321/2.056638; tokenwise/closure-based damage uses the respective baseline. Intervals are pointwise paired-article 95% bootstrap intervals; wins count negative damage. The earlier H100 panel in Table 41 is retained separately.

Policy	Final NLL	ΔNLL	95% CI	Wins
Token INT3	2.11738	$+0.06006$	$[+0.04873, +0.07318]$	0/47
Token INT4	2.07284	$+0.01552$	$[+0.01286, +0.01816]$	5/47
Stream INT3	2.16590	$+0.10926$	$[+0.09730, +0.12376]$	0/47
Stream INT4	2.07256	$+0.01592$	$[+0.01107, +0.02145]$	5/47
Tail INT3	2.09145	$+0.03481$	$[+0.02743, +0.04350]$	0/47
Tail INT4	2.06515	$+0.00851$	$[+0.00631, +0.01063]$	5/47
RtC INT3	2.10665	$+0.05001$	$[+0.04266, +0.05815]$	0/47
Asym INT3	2.11693	$+0.06030$	$[+0.05173, +0.06975]$	0/47

direction differs from the V2-Lite and Coder extensions, whose Tail3-minus-Stream3 differences are $+0.00712$ and $+0.01481$, respectively, with positive paired intervals.

RtC3 and Asym3 give mean damage $+0.05001$ and $+0.06030$, reducing Moonlight's mean Stream3 damage by 54.2% and 44.8%, respectively. Their paired differences from Stream3 are negative with intervals excluding zero (Table 39). These outcomes are checkpoint dependent: RtC3 worsens V2-Lite and Coder; Asym3 worsens Coder and is inconclusive for V2-Lite. Neither alternative is a uniform improvement across the three production checkpoints. Both use eight levels and additional metadata compared with the seven-level reference, so these results do not establish an equal-storage or offset-only advantage.

Table 39: All four added Moonlight policies versus the same-bit symmetric temporal reference on MPS. Negative differences favor the added policy; the sequential BF16 baseline cancels within each paired article. Wins count articles favoring the added policy. Intervals are pointwise; Asym3/RtC3 use eight levels and additional metadata.

Contrast	Damage difference	95% CI	Wins
Tail INT3 − Stream INT3	-0.07445	$[-0.08317, -0.06618]$	47/47
Tail INT4 − Stream INT4	-0.00741	$[-0.01267, -0.00242]$	34/47
RtC INT3 − Stream INT3	-0.05925	$[-0.06954, -0.05003]$	47/47
Asym INT3 − Stream INT3	-0.04896	$[-0.05743, -0.04088]$	46/47

L.2 Same-precision error and damage

Using the estimands in Equation 2, Tail3 has $1.24264\times$ Stream3 relative MSE, with interval $[1.23992, 1.24562]$, and $1.23777\times$ raw MSE per element, $[1.23490, 1.24076]$. Despite these larger reconstruction errors, its loss damage is lower. All 47 articles have both higher Tail3 relative error and lower damage, and all 27 layers have higher pooled Tail3 relative MSE. Equal-article and equal-layer weighting preserve the error direction. The reversal therefore appears in raw and normalized error and is not solely a consequence of the normalization denominator.

This result extends, rather than replaces, the fourteen-checkpoint panel in Appendix J: all of those earlier checkpoints have higher mean Tail3 error and damage, whereas Moonlight has higher error and lower damage. Both policies here measure seven closed blocks, or 896 positions per layer, while scoring 1,023 predictions. Their diagnostics come from their own realized trajectories, not identical BF16 reference activations. The twin-only control in Appendix K does not supply that missing production control. The results establish an error–damage ordering reversal under these policies, without identifying an equal-norm sensitivity or optimizer mechanism.

L.3 Execution comparison, validation, and scope

The model/tokenizer identities and all 282 repeated-condition token sequences agree with the earlier H100 Moonlight panel. All four repeated quantized conditions retain positive mean damage. The largest absolute MPS-minus-H100 damage difference is 0.00457 nats for Token3; its signed difference is -0.00457 $[-0.00826, -0.00094]$. For Stream3, the difference is $+0.00095$ $[-0.00191, +0.00378]$. The panels retain their own BF16 baselines and are not pooled; these differences do not isolate hardware from the execution stack.

All 1,410 production records and 30,456 diagnostic layer records pass the saved-data coverage, source-identity, token-pairing, finite-value, loss/diagnostic arithmetic, and calibration-separation checks. The recorded sequential prefix checks and Stream3/4 reference checks show zero measured drift on the 257-token probes; the failed accelerated chunk candidate is excluded. Intervals here use 10,000 paired-article resamples with seed 20260913 and confirmation-manifest article order; they are pointwise and conditional on these fixed checkpoints. Layers are not independent replicates. Model identities also agree with the separately hashed local source inventories; weights were not reloaded during this analysis.

The completed Moonlight raw SHA256 prefix is `83af9915d16ed43b`. All eight policy estimates, all 28 pairwise damage contrasts, reconstruction summaries, and full source identities are retained in `paper/analysis/moonlight_controls_20260916/`. An independent reduction verifies the policy means and paired intervals, error/energy aggregates, and error-ratio intervals; the tables are generated by `paper/scripts/build_moonlight_controls_tables_20260916.py`. These are teacher-forced WikiText results, not generation or code-quality measurements. Fit mediation, production BF16 reference captures, and equal-storage serving measurements remain unresolved.

M Detailed activation, production, and complementary evidence

This appendix retains the full activation and H100 production panels, all per-pair damage intervals, and the complementary perturbation and task protocols. The primary final-quality and BF16-reference tables appear in the main text; their underlying measurements are unchanged.

M.1 Quantizer geometry and activation statistics

Unquantized activation measurements characterize the tensors seen by the quantizers. Four 2,048-token WikiText test chunks per production model were used to capture post-normalization latents and cached K/V. Table 40 reports layer means. The channel-outlier ratio is $\max_c a_c / \operatorname{median}_c a_c$, where $a_c = \max_t |z_{tc}|$; kurtosis is the standardized fourth central moment over all captured elements in a layer. Relative MSE is measured offline on these fixed tensors, without feeding quantization back into the model. Temporal grouping is valid for this reconstruction measurement; it is not a causal loss experiment. The capture passes themselves do not quantize activations.

Table 40: Original activation controls, recomputed from saved layer statistics. INT4 relative MSE uses groups of 128 along the named axis. *Expanded K includes the replicated positional branch; it is not a pure non-RoPE control. GQA rows describe reconstruction, not downstream GQA accuracy.

Representation	Outlier ratio	Kurtosis	Temporal MSE	Channel MSE
V2-Lite latent	6.05	16.78	0.01296	0.03806
V2-Lite expanded K*	9.14	44.44	0.00550	0.07671
V2-Lite expanded V	3.01	4.65	0.01440	0.01668
Coder latent	5.56	9.93	0.01316	0.03251
Moonlight latent	1.58	3.32	0.01277	0.01532
Qwen3-8B K	7.71	25.71	0.00746	0.05421
Qwen3-8B V	5.09	18.15	0.01699	0.02024

Which coordinates share a scale? Ignoring metadata rounding, the INT3 scale for temporal group g, channel c is $s_{gc} = \max_{t \in g} |z_{tc}|/3$; the tokenwise scale for token t, channel group h is $s_{th} = \max_{c \in h} |z_{tc}|/3$. A high-amplitude channel can enlarge other channels' steps only in the

Table 41: Corrected causal production evaluation: 47 matched articles of 1,024 tokens, with 1,023 predictions per article. Negative paired ΔNLL indicates improvement. Intervals are paired article bootstrap 95% CIs; win counts are descriptive. Each method uses its own matched baseline.

Model	Method	Final NLL	ΔNLL	95% CI	Wins
V2-Lite	Stream INT4	1.90977	-0.00070	$[-0.00172, +0.00031]$	21/47
V2-Lite	Stream INT3	1.92089	$+0.01042$	$[+0.00880, +0.01208]$	2/47
V2-Lite	Token INT4	1.92021	$+0.00977$	$[+0.00751, +0.01209]$	4/47
V2-Lite	Token INT3	1.98846	$+0.07803$	$[+0.07301, +0.08342]$	0/47
Coder	Stream INT4	2.24684	-0.01493	$[-0.01761, -0.01232]$	44/47
Coder	Stream INT3	2.25583	-0.00594	$[-0.00889, -0.00305]$	29/47
Coder	Token INT4	2.25143	-0.01101	$[-0.01492, -0.00698]$	40/47
Coder	Token INT3	2.28462	$+0.02218$	$[+0.01308, +0.03171]$	14/47
Moonlight	Stream INT4	2.07304	$+0.01675$	$[+0.01178, +0.02247]$	7/47
Moonlight	Stream INT3	2.16459	$+0.10830$	$[+0.09638, +0.12241]$	0/47
Moonlight	Token INT4	2.07345	$+0.01710$	$[+0.01453, +0.01969]$	1/47
Moonlight	Token INT3	2.12098	$+0.06462$	$[+0.05458, +0.07684]$	0/47

latter. Between-channel amplitude differences are therefore not, by themselves, a reason to expect large temporal-group error. Pooled kurtosis can mix between-channel amplitude variation with within-channel temporal tails; the reported outlier ratio does not distinguish those geometries.

The observed AdamW tokenwise-INT3 damage rises from 0.040–0.043 at 125M, to 0.072 at 350M, to 0.211–0.280 at 600M, alongside outlier ratios near 3, 6, and 14–24. This is consistent with the shared-scale account, although size and training progress also change. It does not establish a predictor across recipes: Muon's outlier ratios are only 1.68–1.84, yet its tokenwise damage is 0.219–0.636 at 125M and 0.512 at 350M, exceeding AdamW's in all four pairs. At 600M its tokenwise damage is instead 0.073–0.083. Figure 1 shows both policies so the scope of the descriptor comparison is visible.

Across all six pairs, Muon has lower offline INT4 MSE and higher streaming-INT3 loss damage. Different subsets and precision policies prevent interpreting this as amplification at a matched residual. The original capture did not save confirmation-article temporal ranges or full residual tensors; its per-channel profiles and article losses cannot reconstruct them. Qwen3 and expanded-K/V statistics provide reconstruction context rather than an MLA–GQA downstream-quality control.

Same-precision reconstruction and damage. Saved INT3 diagnostics (Appendix J) show 22.0–33.4% lower Muon relative MSE but greater streaming damage in all six pairs; both paired intervals exclude zero. Absolute MSE is instead higher for Muon at 350M/600M, alongside greater latent energy. In the original fourteen-checkpoint panel, matched-tail INT3 has 1.38–2.69 times streaming relative MSE and higher mean damage; four damage intervals include zero. The completed Moonlight panel instead has 24.3% higher matched-tail relative MSE but 68.1% lower mean damage (Appendix L). Lower relative error therefore does not guarantee lower damage. These diagnostics follow each policy's perturbed trajectory, not identical BF16 reference tensors, and do not identify equal-norm sensitivity or fit mediation. On identical BF16 inputs within each checkpoint, all six pairs retain 21.8–33.6% lower Muon streaming relative MSE (Appendix K).

M.2 Production results within each checkpoint

The full-forward and incremental BF16 mean NLLs, respectively, are 1.910436 and 1.910467 for V2-Lite, 2.262441 and 2.261770 for Coder, and 2.056355 and 2.056290 for Moonlight. Tokenwise results use the former; streaming results use the latter. Coder improves under Stream4 and INT3, with paired intervals below zero. V2-Lite's streaming-INT4 interval includes zero, while Moonlight's streaming-INT4 and INT3 intervals are above zero (Table 41).

Coder's sign change is checkpoint-specific. Its own tokenwise-INT3 NLL increases by $+0.02218$, despite improving under Stream3. Improvement therefore is not invariant to the tested cache policy. Coder is instruction-tuned and WikiText is not its primary code domain; regularization, domain mismatch, and residual alignment are possible explanations, not identified mechanisms. The original

code-domain measurements use the future-dependent temporal hook and cannot establish whether the corrected improvement transfers to code.

Comparison units. V2-Lite is a base model; Coder and Moonlight are instruction-tuned. Their predictive distributions and training histories differ. V2-Lite and Coder have matching token-ID hashes on all 47 articles, whereas Moonlight differs on every article and may score different character spans. We therefore lead with within-checkpoint changes rather than a ratio or ordering across production models. The previously computed mixed-token contrasts remain in Appendix C for traceability, without a tokenizer-invariant interpretation. The controlled twins below share a tokenizer and identical scored token sequences.

M.3 Additional controlled-twin contrasts and seed variation

We reuse the completed causal evaluation of 12 controlled checkpoints: three matched seed pairs at 125M, one at 350M, and two at 600M. Initialization, data order, architecture, and schedule were matched within each pair. The intervention combines optimizer family with its independently selected learning rate; other optimizer-specific settings also belong to the recipe. The saved-input audit verified all 3,948 records, frozen evaluation-source and data hashes, matched token hashes, loss reductions, and saved intervals. Of these records, 564 are a noncausal legacy diagnostic excluded from causal conclusions; the six causal conditions contain 3,384 records. The added control study re-evaluates these saved checkpoints without retraining or adding training seeds (Appendix H).

Stream3. Muon's mean damage exceeds AdamW's in all six pairs. Each paired article interval for the excess is positive (Table 2). Ratios of seed-mean damage are $3.54\times$, $4.62\times$, and $1.50\times$ at 125M, 350M, and 600M. However, Muon retains lower absolute final NLL after Stream3 and Stream4 in every pair (Table 1). Relative perturbation sensitivity and final quantized quality are different outcomes.

Tokenwise absolute quality. Table 3 shows the complementary outcome omitted by a streaming-only final-loss comparison. Muon has lower tokenwise-INT4 final NLL in all six pairs. At Token3 it has higher final NLL in all four 125M/350M pairs, with Muon-minus-AdamW gaps $+0.108$, $+0.073$, $+0.498$, and $+0.345$ nats/token. The two 600M gaps are negative (-0.344 and -0.218). All corresponding post-run paired-article intervals exclude zero; the full absolute-quality intervals are included in the analysis addendum.

The largest case is 125M seed 2: Muon's NLL rises from 3.84919 to 4.48561 (damage $+0.63642$), while AdamW rises from 3.94444 to 3.98777 ($+0.04334$). Its $+0.59308$ excess damage overcomes Muon's BF16 advantage, leaving a $+0.49783$ final-loss gap (95% CI $[+0.45058, +0.54704]$). The original activation measurements for this Muon checkpoint show kurtosis 3.02 and outlier ratio 1.83, so a simple extreme-outlier explanation is not supported by those aggregate statistics. They do not measure block-level residuals on the confirmation articles. The case remains an observed failure of this checkpoint-policy combination without an identified mechanism.

Policy and precision counterexamples. At 600M, Token3 reverses the recipe ordering in both seeds: the excess-damage intervals are $[-0.21933, -0.19542]$ and $[-0.13780, -0.11898]$. The corresponding changes from streaming excess damage are -0.21111 and -0.13188, with intervals $[-0.22261, -0.19958]$ and $[-0.14158, -0.12248]$. The seed-mean tokenwise-INT3 Muon/AdamW damage ratio is $0.316\times$.

Stream4 also lacks a uniform recipe effect: its seed-mean damage ratios are $1.28\times$, $3.10\times$, and $0.90\times$. Four of its six pairwise excess-damage intervals include zero. Token4 has positive excess in the four smaller-scale pairs and negative excess in both 600M pairs (Appendix D). These are descriptive patterns across a small set of fixed checkpoints. With only 3/1/2 pairs, they do not establish a population-level recipe-by-scale interaction or a scaling law.

Training-seed variation. At 125M, tokenwise-INT3 excess damage is 0.20377, 0.17655, and 0.59308 across seeds. Its seed mean is 0.32447 and sample SD is 0.23303 ($n = 3$); the corresponding streaming-INT3 SD is 0.00442. Narrow article intervals cannot describe this training variation. We retain each pair and give means, SDs, and ranges for every policy in Appendix F, without constructing a precise training-population interval from 3/1/2 pairs.

What changes at 600M. Every model sees 1B training tokens, equivalent to approximately 8.0/2.86/1.67 tokens per nominal parameter at 125M/350M/600M. Thus size is not varied at matched relative training progress. Latent widths are 256/384/448; only 448 leaves a 64-channel remainder when grouped by 128. That group receives its own scale. The observed reversal may depend on training progress, width, or their interaction with the recipe. The added group-size control tiles 448 evenly with groups of 64 or 112: Muon's lower tokenwise damage persists in both seeds, with a substantially smaller gap at 64 (Table 23). We describe a reversal in the tested configurations, not an isolated effect of parameter scale.

M.4 Perturbation sensitivity beyond marginal shape

The noise probes measure response to a token-local perturbation independent of temporal RTN scale selection. At each latent hook, they add $\epsilon_t = \alpha\,\mathrm{RMS}(z_t)\,\xi_t$, with independent standard Gaussian coordinates in ξ_t. The amplitude uses only the current token's channels. No later token determines an earlier token's noise scale. Both $\alpha = 0.114$ and 0.26 were tested, on three 2K chunks for production and eight 1K chunks for each controlled checkpoint. These are finite perturbations injected at every layer, not measured quantizer residuals.

At $\alpha = 0.114$, production ΔNLL is approximately $+0.00299/+0.01051/+0.01282$ for V2-Lite/Coder/Moonlight. These NLL differences are derived as log ratios of the saved aggregate perplexities. Across all six controlled pairs, Muon has larger noise-induced NLL damage at both tested amplitudes (Appendix G). Thus the recipe dependence extends to a token-local perturbation that does not depend on RTN scale selection. The saved twin summaries retain paired SEMs but not per-chunk losses or repeated noise draws, so we do not attach a new paired significance test or a seed-population interval to this comparison.

The original participation-ratio probe uses the uncentered second moment $S = \sum_t z_t z_t^\top$ and reports $\mathrm{PR}(S) = (\mathrm{tr}\,S)^2/\mathrm{tr}(S^2)$, rather than centered covariance rank. Median PR is 42.45/53.70/61.41 out of 512 for the production models. In the twins it is *lower* for Muon in every pair, even though noise damage is higher. Elevated PR is therefore not a necessary feature of the observed Muon sensitivity. These measurements motivate separating distributional shape from perturbation sensitivity; they do not establish rank mediation, quadratic dominance, or a universal predictor of quality.

M.5 Task accuracy, FP8, and independent engine measurements

Task scoring, FP8 precision controls, and an independent engine context sweep complement the causal NLL comparison under the protocols below.

Full-prompt MMLU. The original 5-shot evaluation selected seven questions per subject across 57 subjects, with a fixed seed. For each question, one forward pass quantized latent temporal groups over the entire supplied prompt, then scored only the final-position logits for the four answer-letter tokens. The test answer is not supplied to the model. All prompt tokens are known at that decision, so within-prompt scale dependence does not expose an unseen test answer. This evaluates a full-prompt quantized classifier; it is distinct from incrementally constructing a streaming cache and does not measure generated-answer accuracy. For these 399-question accuracies, a simple binomial approximation gives an individual-accuracy standard error of about 2.3–2.5 percentage points. This is distinct from the paired-discordance test below; the small stratified sample limits subject-level conclusions.

Moonlight decreases from 67.42% to 62.16% under prompt INT3; its $p = 0.00460$ remains below 0.05 after Bonferroni adjustment over the eight original comparisons ($p_{\mathrm{adj}} = 0.0368$), including two protection-policy comparisons detailed in the appendix. The other comparisons do not establish equivalence. This provides task evidence of sensitivity under the specified prompt policy, without claiming that the new streaming policy has identical task effects.

FP8 scale controls. Original latent E4M3 experiments retain BF16 RoPE keys and evaluate 48 consecutive 2K test chunks. Dynamic scales use only each token's 128-channel block; static scales are frozen from a separate calibration chunk. Dynamic FP8 mean damage is $+0.00014/+0.00036/+0.00037$ for V2-Lite/Coder/Moonlight. Across 30 tested model/configuration combinations, the largest mean increase is Coder's $+0.01152$ under a half-sized per-tensor calibration scale; Moon-

light's largest is $+0.00250$. The sub-8-bit findings therefore do not establish a general FP8 failure of the Moonlight checkpoint. Full grids and paired SEMs are retained in Appendix G.

Real engine and longer context. Independent llama.cpp (Gerganov et al., 2023) Metal runs use V2-Lite Q8_0 GGUF weights with f16 versus q4_0 cache formats. The native MLA path stores the normalized latent together with the positional branch; q4_0 uses blocks within token rows rather than the research hook's temporal absmax. Saved perplexities imply ΔNLL $+0.01553/+0.01642/+0.01733$ at 4K/16K/32K, respectively. These are genuine packed-format engine measurements, with different weights, quantization granularity, and RoPE precision from our causal reference. They support a V2-Lite engine result through 32K, not numerical equivalence to streaming INT4 or a measured serving benefit for that reference implementation.

N Expanded configuration and reproduction record

Table 42 reports the architecture fields in all twelve saved final-checkpoint configuration files. Within each size, the listed fields agree across recipes and seeds. Every checkpoint has vocabulary size 50,257, tied input/output embeddings, no query low-rank bottleneck, RoPE/non-RoPE key head dimensions 32/64, value head dimension 64, and maximum-position configuration 2,048. The evaluated training sequence length is 1,024. These are dense models: the configured number of initial dense layers equals the total number of layers.

Table 42: Architecture dimensions read from the twelve saved final-checkpoint configuration files. Each row applies to both recipes and all reported seeds.

Nominal size	Hidden width	MLP width	Layers	Heads	Latent width
125M	768	2,048	12	12	256
350M	1,024	2,816	24	16	384
600M	1,280	3,456	28	20	448

Retained implementation specification. The retained CUDA training source specifies micro-batches of 32/16/16 sequences with accumulation 2/4/4 for 125M/350M/600M, respectively: 64 sequences or 65,536 input tokens per optimizer step. For a requested 10^9 tokens, its integer step budget is 15,258 steps and 457 warmup steps. Warmup scales the learning rate linearly from zero; the subsequent cosine multiplier decays toward 0.1 of the peak. All optimizer parameter groups receive that multiplier, and gradients are clipped at norm 1.0.

For the AdamW recipe, hidden two-dimensional matrices have weight decay 0.1; embeddings, the output head, and other parameters have zero decay. AdamW uses $\beta = (0.9, 0.95)$. For the Muon recipe, hidden matrices use momentum 0.95, the implemented Nesterov combination, five Newton–Schulz iterations, and weight decay 0.01. The matrix update additionally scales the learning rate by $\max(1, n_{\text{rows}}/n_{\text{cols}})^{1/2}$. Other parameters use AdamW with peak learning rate 3×10^{-4}, the same betas, and zero weight decay. Recipe peak rates for the hidden matrices are those in Appendix E.

The source sets the initialization and data-shuffle seed together, packs tokenized FineWeb-Edu text with EOS separators, and uses the same sequence and batch configuration for a matched pair. Its forward execution uses BF16 autocast and gradient accumulation. The FineWeb validation-stream selection limitations in Appendix B continue to apply.

What these records establish. Checkpoint configurations specify the saved architectures; retained source documents the training implementation without attesting every historical run. Configuration files and source hashes accompany the evidence record, with separate provenance for original activations, causal evaluations, and completed controls. Anonymous public access to the runnable artifact and checkpoints remains pending; these details establish neither a public release nor a new experiment.