WHAT DOES KEY QUANTIZATION ACTUALLY CHANGE IN A DELTA RULE? THE STEP SIZE.

Anonymous authors
Paper under double-blind review

ABSTRACT

Delta-rule linear recurrent neural networks such as DeltaNet, Gated DeltaNet and DeltaProduct update a matrix state with one least-mean-squares step per key. This step shrinks the residual it corrects only if the effective step size βs lies in $(0, 2)$, where β is the step size and s is the squared norm of the key. The models keep $s = 1$ with a key normalizer, but a quantizer placed after the normalizer breaks this: with 4-bit integer (INT4) keys, s deviates from one by up to 9.4%. In models with step sizes below 1, every update then has a relative step-size error of $|s - 1|$. In models with step sizes up to 2, 3.4% of the steps have $\beta s > 2$, where the update flips and enlarges the residual. If β and s are independent, the share of such steps depends only on the tail of β near 2 and the distribution of s, and measuring the two separately predicts the observed shares within 7% for INT4 keys. NormFix divides the step size by the squared norm of the key that the kernel receives, $\eta = \beta/s$. This restores the trained residual factor $1 - \beta$ for every s and makes the propagation of state errors nonexpansive for any quantizer. With 4-bit weights and INT4 keys, NormFix recovers 57% of the long-context state-tracking loss of a Gated DeltaNet with negative eigenvalues and 51% of the loss of a standard one, at less than 4% extra kernel time. On the negative-eigenvalue model it is better than clipping, skipping, key renormalization and 8-bit keys on every state task and on retrieval. Its gain holds across six models and three scale axes, and where the key is normalized after the quantizer, so that $s = 1$, it changes no score by more than 0.01 points.

1 INTRODUCTION

Linear recurrent neural networks (RNNs) built on the delta rule are now a common choice for long-context language modeling (Schlag et al., 2021; Yang et al., 2024; 2025; Siems et al., 2025), and recent large models place them next to attention layers (Qwen Team, 2025; Kimi Team, 2025). Each layer keeps a matrix state $\boldsymbol{S}$. At every token it reads the state along the current key $\boldsymbol{k}$, compares the result with the value $\boldsymbol{v}$, and moves the state toward $\boldsymbol{v}$ by a step of size β. This is one step of the least-mean-squares rule (Widrow & Hoff, 1960), and it multiplies the residual $\boldsymbol{S}\boldsymbol{k} - \boldsymbol{v}$ by $1 - \beta\|\boldsymbol{k}\|_2^2$. The update therefore shrinks the residual only when $0 < \beta\|\boldsymbol{k}\|_2^2 < 2$, the classical step-size condition of the normalized least-mean-squares (NLMS) algorithm (Nagumo & Noda, 1967; Slock, 1993). Delta-rule models meet this condition by design: they normalize the key to unit length and bound β, to $(0, 1)$ in the original models, which we call class I, and to $(0, 2)$ in the class II models, whose transitions can have negative eigenvalues (Grazzi et al., 2025). The condition then holds at every token if the kernel, the fused GPU program that computes the update, receives a key of unit length.

Quantized deployment can break this assumption. Models are served with 4-bit or 8-bit tensors to save memory (Frantar et al., 2023; Lin et al., 2024; Ashkboos et al., 2024), and the key is often among the quantized tensors. When the quantizer runs after the normalizer, the kernel receives a key $\hat{\boldsymbol{k}}$ whose squared norm $s = \|\hat{\boldsymbol{k}}\|_2^2$ is no longer one, and the step it takes is βs instead of β. The product βs is the effective step size of the update. With 4-bit integer (INT4) keys, the 1st percentile of s is 0.96 or 0.97 and the maximum 1.08 or 1.09, depending on where the quantizer sits (Figure 1b). How much an error of a few percent in the step size matters depends on how close β is to the bound 2.

The factor $1 - \beta s$ separates two cases, and β decides which of them a step can reach. If $\beta s < 2$, the update still shrinks the residual, but its step differs from the trained step β by a relative error of $|s - 1|$. If $\beta s > 2$, the factor falls below -1: the update overshoots, flips the sign of the residual and

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.

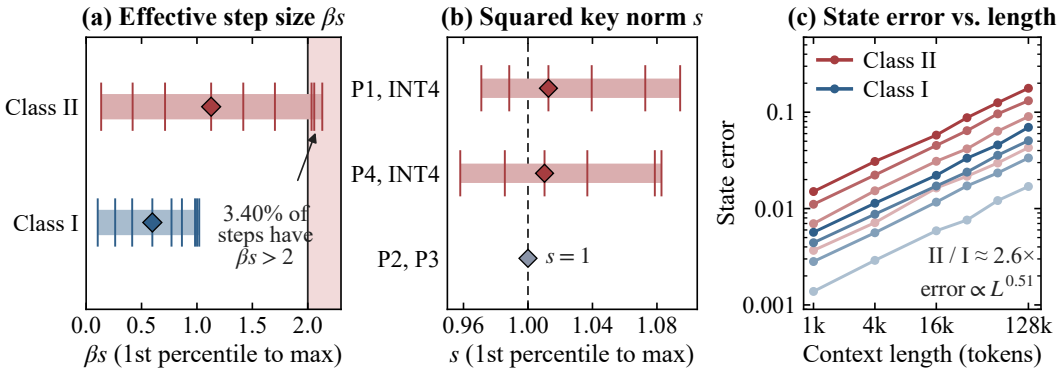

Figure 1: The step size of a quantized delta-rule update. (a) Effective step size βs of class I (quantiles averaged over DeltaNet and Gated DeltaNet) and class II (Gated DeltaNet with negative eigenvalues) with INT4 keys. Bars span the 1st percentile to the maximum, ticks mark the percentiles 1, 10, 25, 50, 75, 90, 99 and 99.9 and the maximum, and diamonds mark the median. In class II, 3.40% of the steps fall in the shaded region $\beta s > 2$, where an update enlarges the residual. (b) Squared norm s of the INT4 key that the kernel receives on each placement of the quantizer (P1 to P4, Section 3.1), drawn as in (a) with ticks at the percentiles 1, 10, 50, 90 and 99.9 and the maximum; s differs from 1 (dashed) only on P1 and P4, where the quantizer runs after the normalizer. (c) State error, the norm of the difference between the perturbed and the unperturbed state, after perturbations of four sizes along the direction of every key (lighter lines are smaller), from 1k to 128k tokens. It grows about like $L^{0.51}$ in the context length L and is 2.6 times larger in class II on average.

makes it larger. A class I step would need $s > 2$ to overshoot, far outside the range of s. A class II step with β close to 2 overshoots as soon as $s - 1$ exceeds about half the margin $2 - \beta$, and such step sizes are not rare: 6.5% of the class II step sizes lie in $[1.8, 2)$, and with INT4 keys 3.4% of the steps overshoot (Figure 1a). The losses are also larger in class II. With 4-bit weights (W4) and INT4 keys, state tracking at 64k tokens drops by 31.0 points in class II and by 6.8 points in class I (Table 1), and the state error left by perturbing every key keeps growing with the context length (Figure 1c). Both regimes come from the norm s, which the kernel never checks, so a repair should act on s itself.

Acting on s is cheap, because the kernel already holds $\hat{\boldsymbol{k}}$. We divide the step size by the squared norm, $\eta = \beta/s$, so that the effective step size equals β again for every s. This is the NLMS normalization, applied to the key that the kernel receives instead of the key that the model computed. We call it NormFix. It changes only the step size and needs no training; the extra work is a reduction over the key and a division of β by s per head and step. Under NormFix the matrix that carries the state error from one step to the next is nonexpansive for any quantizer, and the error recursion of the unrepaired kernel differs from that of NormFix by one rank-one term per step whose size is proportional to $s - 1$.

We test these statements with controlled perturbations of the key and by measuring the state error with and without the repair, and we evaluate NormFix on six delta-rule models, five quantization recipes and three scale axes. If β and s are independent, the share of overshooting steps is fixed by the tail of β near 2 and the distribution of s; measuring the two separately and combining them predicts 3.17% against the measured 3.40%. In class II, NormFix recovers 17.7 of the 31.0 lost points at 3.6% extra kernel time, and in class I it recovers 3.5 of 6.8 points, more than the usual alternatives. Clipping β below 2 removes most overshoots but also shortens the steps near 2 that the model was trained to take, and skipping the overshooting steps removes their updates. More bits shrink $|s - 1|$ but do not bound βs: with 8-bit integer (INT8) keys, 0.59% of the steps still overshoot. Renormalizing the key restores the same residual factor but gains 3.9 points less. The gain of NormFix is larger in models whose β can approach 2, and it stays between 15.6 and 18.2 points on every setting of the three scale axes. Section 2 reviews related work, Section 3 derives and measures the two regimes, Section 4 presents NormFix and its analysis, and Section 5 reports the experiments, including the ablation studies.

2 Related Work

Delta-rule linear RNNs. Linear attention stores key-value outer products in a matrix state (Katharopoulos et al., 2020), and the delta rule replaces this additive write with a correction toward the value (Schlag et al., 2021). DeltaNet parallelizes the rule over the sequence (Yang et al.,

2024), Gated DeltaNet adds a decay (Yang et al., 2025), and the same update appears as online regression at test time (Liu et al., 2025a; Sun et al., 2025; Behrouz et al., 2025; Wang et al., 2025). Grazzi et al. (2025) extend β to $(0, 2)$ so that the transition can have negative eigenvalues, which widens the range of state-tracking problems linear RNNs can solve (Merrill et al., 2024), and DeltaProduct applies several such factors per token (Siems et al., 2025). RWKV-7 uses a vector-valued step size with separate removal and replacement keys (Peng et al., 2025), and Mamba-2 has no key in its transition (Dao & Gu, 2024). NormFix applies without retraining to every model with a scalar step size on one key, from DeltaNet to DeltaProduct, whose several factors it then repairs one at a time.

Post-training quantization. Weight-only GPTQ and AWQ (Frantar et al., 2023; Lin et al., 2024), activation smoothing and outlier handling (Xiao et al., 2023; Dettmers et al., 2022), rotations in QuaRot and SpinQuant (Ashkboos et al., 2024; Liu et al., 2025b), cache quantization (Liu et al., 2024b; Hooper et al., 2024) and low-bit floating-point formats (Micikevicius et al., 2022) reduce the error of their target tensors (Nagel et al., 2021). Quamba, Quamba2 and MambaQuant adapt such recipes to selective state-space models (Chiang et al., 2025b;a; Xu et al., 2025). These methods change $|s - 1|$ when applied to the key and barely when applied to the weights. None of them bounds βs, so in our experiments they are baselines that NormFix can be combined with, and quantization-aware training (QAT) (Liu et al., 2024a) is a reference that retrains the model with quantized weights.

Step size and precision in adaptive filters. With a unit key, the delta rule is the NLMS algorithm (Nagumo & Noda, 1967), whose normalized step size must lie in $(0, 2)$ (Slock, 1993; Haykin, 2002; Sayed, 2003). The same interval is the relaxation range of the Kaczmarz method (Kaczmarz, 1937; Strohmer & Vershynin, 2009), and finite-precision effects in least-mean-squares filters were analyzed for fixed-point arithmetic (Caraiscos & Liu, 1984; Cioffi, 1987). A factor with $\beta = 2$ is a Householder reflection (Householder, 1958), whose computation needs a unit reflecting vector or a division by its squared norm (Golub & Van Loan, 2013). NormFix applies this division to the key a quantized kernel receives. We characterize when a quantizer after the normalizer breaks the condition and how often, and we measure the accuracy this costs a delta-rule language model and the time the repair adds.

3 The Effective Step Size under Key Quantization

3.1 Preliminaries

A delta-rule layer with head dimension d keeps one matrix state per head. At step t the head receives a key $\boldsymbol{k}_t \in \mathbb{R}^d$ with $\|\boldsymbol{k}_t\|_2 = 1$, a value $\boldsymbol{v}_t \in \mathbb{R}^d$, a query $\boldsymbol{q}_t \in \mathbb{R}^d$, a step size β_t and, in gated models, a decay $\alpha_t \in (0, 1]$. The state $\boldsymbol{S}_t \in \mathbb{R}^{d\times d}$ is updated and read as

$$\boldsymbol{S}_t = \alpha_t \boldsymbol{S}_{t-1}\big(\boldsymbol{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top\big) + \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^\top, \qquad \boldsymbol{o}_t = \boldsymbol{S}_t \boldsymbol{q}_t, \tag{1}$$

where $\boldsymbol{I}$ is the identity and $\boldsymbol{o}_t$ is the output (Yang et al., 2024; 2025). The first term of (1) erases the old value stored along $\boldsymbol{k}_t$, and the second writes the new value along it. The matrix $\boldsymbol{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top$ has eigenvalue $1 - \beta_t$ along $\boldsymbol{k}_t$ and 1 in every orthogonal direction. With σ the logistic sigmoid, a model is in class I if $\beta_t = \sigma(\cdot) \in (0, 1)$, as in DeltaNet and Gated DeltaNet, and in class II if $\beta_t = 2\sigma(\cdot) \in (0, 2)$, as in Gated DeltaNet with negative eigenvalues (Grazzi et al., 2025) and DeltaProduct, which applies n_h such factors per token (Siems et al., 2025). In class II a step size near 2 turns the factor into a near reflection, which nearly flips the component of the state along the key.

Deployment quantizes the key somewhere between its projection and the kernel. We write $\hat{\boldsymbol{k}}_t$ for the key the kernel receives, $s_t = \|\hat{\boldsymbol{k}}_t\|_2^2$ for its squared norm and $\beta_t s_t$ for the effective step size. Whether s_t differs from one depends on the order of the normalizer and the quantizer, and we consider four placements (Appendix B.2). On P1 and P4 the quantizer runs last: on P1 the normalized key is quantized as an activation and used as it is, and on P4 it is quantized on entry to the fused kernel in the kernel's own format. On P2 and P3 the normalizer runs last: on P2 the key is quantized before the normalizer, and on P3 the kernel renormalizes the quantized key. Hence $s_t = 1$ on P2 and P3, while on P1 and P4 s_t carries the rounding error of the quantizer, which changes from one step to the next.

3.2 Two Regimes of the Effective Step Size

Let $\boldsymbol{r}_t = \alpha_t \boldsymbol{S}_{t-1} \hat{\boldsymbol{k}}_t - \boldsymbol{v}_t$ be the residual before the update and $\boldsymbol{r}_t^+ = \boldsymbol{S}_t \hat{\boldsymbol{k}}_t - \boldsymbol{v}_t$ the residual after it, both read with the key that the kernel uses in the update, since it is the only key the kernel has.

Lemma 1 (a posteriori residual). *For every step of (1) computed with $\hat{\boldsymbol{k}}_t$, $\boldsymbol{r}_t^+ = (1 - \beta_t s_t)\,\boldsymbol{r}_t$, whatever the quantizer and the decay are, including decays $\alpha_t < 1$ and received keys of any norm.*

Writing the update as $\boldsymbol{S}_t = \alpha_t \boldsymbol{S}_{t-1} - \beta_t \boldsymbol{r}_t \hat{\boldsymbol{k}}_t^\top$ and multiplying it by $\hat{\boldsymbol{k}}_t$ gives the lemma (Appendix A.2). It is the relation between the a priori and the a posteriori error of least-mean-squares filters (Sayed, 2003), applied here to a key whose norm the kernel does not control: the factor on the residual is $1 - \beta_t s_t$, and the bound on β_t controls it only when $s_t = 1$. On the recorded steps, the measured ratio $\|\boldsymbol{r}_t^+\|_2 / \|\boldsymbol{r}_t\|_2$ agrees with $|1 - \beta_t s_t|$ up to floating-point error on all four placements.

Proposition 2 (two regimes). *Write $s_t = 1 + \varepsilon_t$ with $\varepsilon_t > -1$. (i) If $\beta_t s_t < 2$, the update contracts the residual, and its effective step size differs from β_t by the relative error $|\varepsilon_t|$. (ii) If $\beta_t s_t > 2$, then $1 - \beta_t s_t < -1$: the update flips the sign of the residual and enlarges its norm by the factor $\beta_t s_t - 1$. (iii) If $\beta_t \le 1$, case (ii) needs $\varepsilon_t > 1$. If $\beta_t = 2 - m_t$ with margin $m_t \in (0, 2)$, case (ii) holds if and only if $\varepsilon_t > m_t/(2 - m_t)$, that is, $\varepsilon_t > m_t/2 + O(m_t^2)$ for a small margin m_t below the bound.*

With the largest recorded $s_t = 1.094$, part (iii) rules out case (ii) in class I and allows it in class II only for step sizes within 0.17 of 2. To see how often each case occurs, we record β_t, s_t and both residuals on 6144 steps from the four heads of six layers, under every placement and recipe (Appendix C). In class I no step overshoots; on P1 with INT4 keys the effective step size stays below 1.03, but the 99th percentile of the step-size error is 5.6%. In class II, 3.40% of the steps overshoot on P1 with INT4 keys, a similar share on P4, five to six times fewer with 8-bit floating-point (FP8) keys, and 0.07% in the bfloat16 (BF16) reference. Every one of them has $\beta_t \in [1.8, 2)$, a bin that holds 6.5% of the step sizes. Because the shift $\beta_t(s_t - 1)$ grows with β_t, the upper tail of $\beta_t s_t$ reaches past 2 (Figure 1a).

3.3 Rate of Overshooting Steps

We treat β_t and s_t as random over the steps; how often a step overshoots depends on their joint law.

Theorem 3 (overshoot rate). *Let $\bar{G}(x) = \Pr[\beta_t > x]$ and $h(\delta) = \bar{G}\big(2/(1+\delta)\big)$. If β_t and s_t are independent,*

$$\Pr[\beta_t s_t > 2] = \mathbb{E}\big[h(s_t - 1)\big], \tag{2}$$

where $h(\delta) = 0$ for $\delta \le 0$, and $h(\delta) = 0$ for $\delta \le 1$ in class I. If moreover $s_t \le s_{\max}$ and β_t has a density f on $(2/s_{\max}, 2)$ with left limit $f(2^-)$ and Lipschitz constant ℓ, then, with $(x)_+ = \max(x, 0)$,

$$\Pr[\beta_t s_t > 2] = 2 f(2^-)\,\mathbb{E}\Big[\frac{(s_t - 1)_+}{s_t}\Big] + R, \qquad |R| \le 2\ell\,\mathbb{E}\big[(s_t - 1)_+^2\big]. \tag{3}$$

Without the independence of β_t and s_t, $\Pr[\beta_t s_t > 2] \le h(\delta) + \Pr[s_t > 1 + \delta]$ for every $\delta \ge 0$.

For a small spread of s, the rate is about $2 f(2^-)\,\mathbb{E}[(s-1)_+]$: the density of β just below 2, which the model sets, times the mean positive part of $s - 1$, which the quantizer sets (Appendix A.4). Overshoots therefore occur even when s is one on average, as long as some steps have $s > 1$. More bits reduce the rate only through $\mathbb{E}[(s-1)_+]$, which under a uniform rounding model falls linearly with the grid step and stays positive for every finite grid (Corollary 8). At the spread of INT4 keys the density of β changes quickly below 2, so the predictions below use the exact form (2) of the rate.

The two factors can be measured separately. A shift of s by the same amount δ at every step makes the rate $h(\delta)$, up to the change the shift causes in the step sizes, so shifts applied to the unquantized keys of the class II model sample h directly. A radial shift of amplitude a changes s by about $2a + a^2$ and a tangential one of the same norm by about a^2, and at $a = 0.045$ the radial shift makes 6.5% of the steps overshoot and the tangential one 0.8% (Figure 2a). The rate therefore follows the resulting shift Δs of s rather than the size of the key error, and it increases with Δs for both kinds of shift.

Combining the measured h with the recorded distribution of s through (2) predicts the measured share within 7% with INT4 keys, 3.17% against 3.40% on P1 and 3.09% against 3.24% on P4 (Figure 2b). With FP8 keys, where overshoots are five to six times rarer and rest on the few steps with β within about 0.03 of 2, the prediction is within 5% on P4 and 23% on P1 (Appendix D.4). With INT4 keys the mean of s changes by at most 0.003 between the bins of β, which is consistent with independence (Appendix C.3). Without this assumption, the bound of Theorem 3 still holds; on the recorded steps it gives about 6.5% at its best δ, about twice the measured rate, since it must hold for any joint law of β_t and s_t. As a check of the implementation, the rule of Proposition 2(iii) predicts all 30 cells of a grid with the step size fixed at a margin m below 2 and a known radial shift of the key (Appendix D.2).

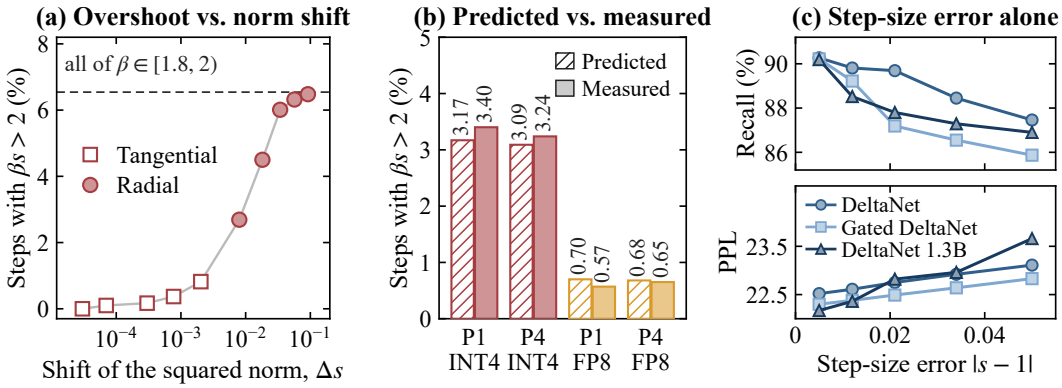

Figure 2: Tests of the overshoot rate and of the step-size error. (a) Share of class II steps with $\beta s > 2$ when the unquantized keys are shifted along the key (radial, circles) or orthogonal to it (tangential, squares), against the resulting shift Δs of the squared norm; this share is $h(\Delta s)$ in Theorem 3. The gray line interpolates the points, and the dashed line is the share of step sizes in $[1.8, 2)$. (b) Share of steps with $\beta s > 2$ predicted by Theorem 3 from h and the recorded distribution of s (hatched), and measured (solid). (c) Associative recall and perplexity of three class I models when the step-size error $|s - 1|$ is set to fixed values; no step overshoots in this test.

A step-size error of a few percent costs accuracy even when no step overshoots. In class I, we set $|s-1|$ to fixed values by scaling the key. From $|s-1| = 0.005$ to 0.05, associative recall (Arora et al., 2024) falls monotonically, by 2.8 to 4.4 points across three models, one of them a 1.3B DeltaNet, and perplexity rises by 0.5 to 1.5 (Figure 2c). In class II, unrepaired sequences with the most overshooting steps lose up to 7.8 more points of state-tracking accuracy than those with the fewest (Appendix D.5).

4 NormFix

4.1 Rescaling the Step Size by the Key Norm

The quantity that must stay in $(0, 2)$ is $\beta_t s_t$, and the kernel already holds $\hat{\boldsymbol{k}}_t$. NormFix computes s_t and replaces the step size by

$$\eta_t = \beta_t / s_t, \qquad \boldsymbol{S}_t = \alpha_t \boldsymbol{S}_{t-1}\big(\boldsymbol{I} - \eta_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top\big) + \eta_t \boldsymbol{v}_t \hat{\boldsymbol{k}}_t^\top. \tag{4}$$

By Lemma 1 the residual is then multiplied by $1 - \eta_t s_t = 1 - \beta_t$, the factor the model was trained with, for every $s_t > 0$, and no step overshoots. The key, the value, the decay and the structure of the kernel stay unchanged, and nothing is trained. In a fused kernel, s_t is accumulated in 32-bit floating point (FP32) over the d coordinates of $\hat{\boldsymbol{k}}_t$, which takes 128 multiplications and 127 additions at $d = 128$, and then divides β_t; our kernel needs two extra registers and no extra memory traffic (Appendix H). On P2 and P3, where $s_t = 1$, NormFix leaves every step unchanged up to rounding. For DeltaProduct, NormFix repairs each of the n_h factors of a token with its own s (Proposition 9).

4.2 Propagation of State Errors

Lemma 1 describes one step. To see what the repair implies over a sequence of L tokens, we isolate the effect of the key: let $\boldsymbol{S}_t^\star$ be the state of a kernel that receives the same α_t, β_t and $\boldsymbol{v}_t$ but the unquantized key $\boldsymbol{k}_t$, and let $\boldsymbol{E}_t = \boldsymbol{S}_t - \boldsymbol{S}_t^\star$ be the error of the deployed state, with $\boldsymbol{E}_0 = \boldsymbol{0}$ initially.

Theorem 4 (error propagation)**.** *For a kernel that uses $\hat{\boldsymbol{k}}_t$ and a step size γ_t, the error obeys $\boldsymbol{E}_t = \boldsymbol{E}_{t-1}\boldsymbol{A}_t + \boldsymbol{F}_t$ with the error transition matrix $\boldsymbol{A}_t = \alpha_t(\boldsymbol{I} - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top)$ and the local error $\boldsymbol{F}_t = \alpha_t \boldsymbol{S}_{t-1}^\star(\beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top) + \boldsymbol{v}_t(\gamma_t \hat{\boldsymbol{k}}_t - \beta_t \boldsymbol{k}_t)^\top$. (i) Under NormFix, $\gamma_t = \eta_t$, the matrix $\boldsymbol{A}_t$ is symmetric with $\|\boldsymbol{A}_t\|_2 = \alpha_t$ for every $s_t > 0$ and $\beta_t \in (0, 2]$. Hence, for any quantizer,*

$$\|\boldsymbol{E}_L\|_F \le \sum_{t=1}^{L} \Big(\prod_{j=t+1}^{L} \alpha_j \Big) \|\boldsymbol{F}_t\|_F. \tag{5}$$

(ii) Without repair, $\gamma_t = \beta_t$ and $\|\boldsymbol{A}_t\|_2 = \alpha_t \max(1, \beta_t s_t - 1)$, which exceeds one if and only if $\alpha_t(\beta_t s_t - 1) > 1$. (iii) The unrepaired error follows the NormFix recursion plus one rank-one term per step,

$$\boldsymbol{E}_t = \boldsymbol{E}_{t-1}\boldsymbol{A}_t^{\mathrm{NF}} + \boldsymbol{F}_t^{\mathrm{NF}} - \beta_t \frac{s_t - 1}{s_t} \boldsymbol{r}_t \hat{\boldsymbol{k}}_t^\top, \tag{6}$$

where $\boldsymbol{A}_t^{\mathrm{NF}}$ and $\boldsymbol{F}_t^{\mathrm{NF}}$ are the terms of part (i) and $\boldsymbol{r}_t$ is the residual of the unrepaired state at step t.

Part (i) holds for any quantizer and any size of error: the recurrence can no longer amplify an error that it carries (Appendix A.5). It does not remove $\boldsymbol{F}_t$, which still holds the direction error of $\hat{\boldsymbol{k}}_t$ and a relative error of $1/\sqrt{s_t} - 1$ in the size of the write term. Part (iii) isolates the difference between the unrepaired and the NormFix recursions in one rank-one term per step, proportional to $s_t - 1$ and to the correction that the update makes. In class I this term changes each correction; in class II the same term and part (ii) let overshooting steps enlarge the error. The theorem also shows where the repair helps less: on P2 and P3, where $s_t = 1$ and the two recursions coincide, and in class I, where $\boldsymbol{A}_t$ is nonexpansive even without the repair and only the rank-one term remains. If the local errors are uncorrelated with the carried one, part (i) bounds the root-mean-square error by a multiple of $\sqrt{L}$ (Corollary 6). This bound also covers the unrepaired kernel in class I, where $\beta_t s_t < 2$ at all steps, and there the error left by perturbing all keys grows like $L^{0.51}$ from 1k to 128k tokens (Figure 1c).

4.3 Relation to Key Renormalization

A second way to correct the norm is to divide the key by $\sqrt{s_t}$. Both repairs belong to one family.

Proposition 5 (scalar repairs)**.** *Let a repair use the key $c_t\hat{\boldsymbol{k}}_t$ and the step size $g_t\beta_t$ with scalars $c_t, g_t > 0$. The residual read with $c_t\hat{\boldsymbol{k}}_t$ is multiplied by $1 - \beta_t$ for every state if and only if $g_t c_t^2 s_t = 1$. All such repairs share the erase term $\alpha_t \boldsymbol{S}_{t-1}(\boldsymbol{I} - \beta_t \bar{\boldsymbol{k}}_t \bar{\boldsymbol{k}}_t^\top)$ with $\bar{\boldsymbol{k}}_t = \hat{\boldsymbol{k}}_t/\sqrt{s_t}$, and hence the matrix $\boldsymbol{A}_t^{\mathrm{NF}}$ of Theorem 4, and differ only in the write term $\beta_t\,(c_t\sqrt{s_t})^{-1}\boldsymbol{v}_t\bar{\boldsymbol{k}}_t^\top$. NormFix is the member with $c_t = 1$ and renormalization the member with $c_t = 1/\sqrt{s_t}$, so their write terms differ by $\sqrt{s_t}$.*

All members satisfy the bound (5), and they differ only in how much they write: renormalization writes the trained amount β_t along the received direction, and NormFix writes $\beta_t/\sqrt{s_t}$, the amount that keeps the step size of the received key. The two write terms differ by a factor of at most 1.046 at the largest recorded s_t. Both repairs can be computed with scalars on the received key, but the usual renormalization kernel rescales the key vector, which needs a second pass over the key and a write-back (Appendix H). Section 5 compares the two repairs empirically, in accuracy and in cost.

5 Experiments

5.1 Experimental Setup

Models. Class I: DeltaNet and Gated DeltaNet. Class II: Gated DeltaNet with $\beta \in (0, 2)$ and DeltaProduct with $n_h \in \{1, 2, 3\}$. All are trained from scratch with 340M parameters on 20B tokens of FineWeb-Edu (Penedo et al., 2024), with 32 layers, four heads per layer and $d = 128$ unless a scale axis changes them; Section 3.3 also uses a 1.3B DeltaNet. **Quantization.** The default recipe quantizes the weights to W4 and the normalized key to INT4 on P1, both with group size 128. The other recipes are W4 per-channel, W4A8 (8-bit activations), QuaRot+W4 and SpinQuant+W4; W4 with an INT8 or FP8 key, 6-bit weights (W6), GPTQ or AWQ W4 and QAT-W4 are further baselines. **Tasks.** Perplexity on PG-19 (Rae et al., 2020) and Books3 (Gao et al., 2020) at 8k, 32k and 128k tokens; RULER (Hsieh et al., 2024), averaged over 4k to 128k tokens, and LongBench (Bai et al., 2024); state tracking on parity (Delétang et al., 2023), S_5 (composing permutations of five elements) (Merrill et al., 2024) and Dyck, trained at 4k and tested up to 128k tokens. **Default setting.** Class II Gated DeltaNet on P1 with the default recipe, state tracking at 64k tokens. **Protocol.** Kernel time is the median of five runs on one H100 GPU, throughput is the decode rate at batch size one, and intervals are 95% confidence intervals of paired differences over 4096 sequences (Appendix B).

Table 1: Main results on P1 with W4 weights and INT4 keys unless the row says otherwise, for class II (top) and class I (bottom). PPL: perplexity at 8k, 32k and 128k tokens, geometric mean of PG-19 and Books3. RULER and LongBench (LB). Parity, S_5 and Dyck at 64k tokens and their average (Avg). $\beta s{>}2$: share of steps whose effective step size exceeds 2 (%). Err: 99th percentile of the step-size error (%), not reported for skipping. In class I, clipping and skipping leave the model unchanged and share the No fix row. Bold and underline mark the best and second best post-training row in each quality column; gray rows are BF16 or retrained references and blue rows NormFix. Appendix H reports kernel time and throughput.

	PPL ↓			Retrieval ↑		State tracking ↑				Step size	
Method	8k	32k	128k	RULER	LB	Parity	S_5	Dyck	Avg	$\beta s{>}2$	Err
Class II: Gated DeltaNet with negative eigenvalues, $\beta \in (0, 2)$											
BF16 reference	15.21	18.61	21.39	41.71	16.31	96.14	91.02	93.20	93.45	0.07	0.05
No fix	16.88	20.73	26.14	31.10	10.64	65.40	58.99	62.92	62.44	3.40	5.95
Clip β at 1.9	16.68	21.12	26.04	34.19	11.86	71.96	63.85	72.12	69.31	0.11	6.01
Clip β at 1.0	16.88	21.20	26.75	31.15	10.33	63.72	56.44	60.40	60.19	0.00	51.54
Skip if $\beta s > 2$	16.14	20.86	25.41	34.47	12.08	76.12	68.27	72.00	72.13	0.00	n/a
Renormalize key	16.45	20.17	24.91	35.39	13.07	79.16	73.72	75.91	76.26	0.00	0.00
W6 weights	16.46	20.38	24.96	37.46	13.82	82.45	75.54	75.75	77.91	3.34	5.89
INT8 key	16.17	**19.43**	**23.65**	36.96	13.73	81.86	75.08	75.61	77.52	0.59	1.01
FP8 key	**15.99**	21.09	25.03	36.70	13.19	78.96	71.04	80.40	76.80	0.57	1.01
QuaRot + W4	16.13	20.63	25.62	34.94	12.96	78.95	72.62	73.45	75.01	3.48	6.11
SpinQuant + W4	16.86	21.23	25.70	33.75	11.77	74.14	65.08	71.27	70.16	3.34	5.78
GPTQ / AWQ W4	16.46	20.60	25.27	33.83	11.37	71.81	62.53	66.94	67.09	3.37	5.81
QAT-W4 (retrained)	15.88	19.06	23.46	40.14	15.91	87.64	84.84	87.66	86.71	3.34	6.09
NormFix	16.18	19.95	24.12	**37.70**	**14.12**	**83.29**	**76.16**	**81.07**	**80.17**	0.00	0.00
Class I: Gated DeltaNet, $\beta \in (0, 1)$											
BF16 reference	15.05	19.18	22.50	43.40	18.72	99.97	91.20	95.80	95.66	0.00	0.07
No fix (= clip, skip)	16.49	19.49	24.15	40.09	17.32	94.43	83.69	88.36	88.83	0.00	5.56
Renormalize key	16.48	19.86	23.68	42.68	18.15	95.16	84.81	90.48	90.15	0.00	0.00
W6 weights	**15.26**	19.78	24.40	42.98	**18.70**	95.41	84.15	**94.73**	91.43	0.00	5.92
INT8 key	15.83	19.84	23.55	43.09	17.60	94.99	86.89	94.35	92.08	0.00	1.02
FP8 key	15.70	20.33	24.19	42.85	18.33	96.01	84.78	93.01	91.27	0.00	1.00
QuaRot + W4	15.81	19.70	**23.18**	42.70	17.04	91.92	83.44	90.93	88.76	0.00	6.01
SpinQuant + W4	15.64	19.59	23.57	42.44	17.78	94.31	**87.65**	90.45	90.80	0.00	6.04
GPTQ / AWQ W4	15.81	19.59	23.33	42.29	17.87	90.44	84.78	90.01	88.41	0.00	5.83
QAT-W4 (retrained)	15.83	19.07	22.91	43.38	19.51	100.00	88.68	94.28	94.32	0.00	6.07
NormFix	15.58	**19.42**	23.42	**43.10**	18.35	**97.20**	86.90	92.80	**92.30**	0.00	0.00

5.2 Main Results

Table 1 compares NormFix with repairs of the step, more bits, other post-training recipes, and references that use BF16 or retrain the model. In class II, NormFix raises the state-tracking average at 64k tokens from 62.4 to 80.2, a paired gain of 17.7 points with the 95% interval $[16.3, 19.2]$, and it is the best post-training row on all three state tasks, on RULER and on LongBench. It also lowers the perplexity at 128k tokens from 26.14 to 24.12; at 32k and 128k tokens the INT8 key reaches a lower perplexity, 19.43 and 23.65 against 19.95 and 24.12, and at 8k tokens several rows are within 0.2.

Skipping removes every overshoot and clipping at 1.9 removes 97% of them, but they reach only 72.1 and 69.3, since both change the steps near 2 that carry the reflections. Clipping at 1.0 falls below the unrepaired model, as it turns the reflections into near-complete erasures. More bits leave the product βs unconstrained: W6 weights keep the INT4 key and its overshoots, INT8 and FP8 keys still overshoot on 0.59% and 0.57% of the steps, and the three reach 76.8 to 77.9. Renormalizing the key restores the same residual factor and reaches 76.3, 3.9 points below NormFix with the interval $[3.3, 4.6]$. QAT reaches 86.7 by retraining the model, although 3.34% of its steps still overshoot.

In class I no step overshoots, so clipping and skipping leave the model unchanged. NormFix still removes the step-size error and gains 3.5 points on the state tasks with the interval $[2.8, 4.2]$, against 1.3 for renormalization and 2.4 for the FP8 key. The INT8 key comes closest: it is 0.22 points below NormFix on the mean of the state tasks and level with it on RULER. NormFix also lowers the perplexity at 128k tokens from 24.15 to 23.42. In class II, NormFix adds 3.6% kernel time and lowers decode throughput by 2.8%, and renormalization adds about twice as much (Appendix H).

Table 2: Ablations in the default setting: parity and S_5 accuracy at 64k tokens (↑) and the share of the NormFix gain kept on their mean (%), where 0 is the unrepaired model and 100 is NormFix; in rows 1 to 13, darker cells keep more. Row 0 is NormFix, and the gray row is the unrepaired model. Table 27 adds perplexity.

#	Variant	Parity	S_5	Kept
0	**NormFix**	83.3	76.2	100.0
1	Skip	76.1	68.3	57.0
2	Clip at 1.9	72.0	63.9	32.6
3	Renormalize	79.2	73.7	81.3
4	Only $\beta s > 2$	81.3	74.3	89.1
5	Only $\beta s > 1.9$	82.0	74.4	91.3
6	Only $\beta s > 1$	82.7	75.7	97.0
7	First 8 layers	76.9	68.1	58.8
8	First 16 layers	79.8	71.7	77.3
9	First 24 layers	81.8	74.0	89.6
10	s before quantization	65.4	59.0	0.0
11	Per-layer mean s	82.1	75.0	93.4
12	16-bit accumulation	82.2	75.0	93.8
13	Head exemption	82.0	75.4	94.4
	No fix	65.4	59.0	0.0

5.3 Ablation Studies

Table 2 changes one design decision at a time, and Appendix F gives every sweep with its perplexity.

Rescaling the step size outperforms skipping, clipping and key renormalization. Skipping the overshooting steps keeps 57% of the gain, clipping at 1.9 keeps 33% and renormalizing the key 81% (rows 1 to 3), and a cap at 1.5 removes every overshoot yet keeps only 19% (Appendix F.1). Skipping drops an update and clipping shrinks every step above the cap, while rescaling keeps the trained factor $1 - \beta_t$ at every step. Renormalization keeps this factor too and differs from NormFix only in the write term (Proposition 5), which therefore accounts for the remaining 19% of the full gain.

Repairing only some steps loses part of the gain. Restricting NormFix to the steps with βs above 2, above 1.9 or above 1 reaches 89%, 91% and 97% of the gain (rows 4 to 6). The overshooting steps, only 3.4% of all steps, carry most of the gain, and the step-size error of the other steps carries the rest.

Every block of layers contributes. Repairing only the first 8, 16 or 24 of the 32 layers yields 59%, 77% and 90% of the gain (rows 7 to 9); only the full repair removes all overshoots (Appendix F.3).

The norm of the received key at each step works best. Computing s from the key before the quantizer gives $s = 1$ and no gain (row 10), and a per-layer mean of s loses 7% of the gain (row 11).

Lower precision for s costs part of the gain. Accumulating s in 16-bit floating point instead loses 6% of the gain (row 12), so the kernel accumulates s in FP32 and then divides β_t by it exactly.

Exempting heads does not help. On a 64-step test sequence that starts from an outdated state, some heads reach a lower error without the repair. Leaving these heads unrepaired costs 6% of the gain (row 13), so they also benefit over long sequences, and the repair runs on every head (Appendix F.4).

5.4 Results on Other Models and Settings

The gain holds across lengths and models. At every length from 4k to 128k tokens, NormFix is above the three other methods on all three state tasks (Figure 3a). Across the six delta-rule models, the gain follows the range of the step size (Figure 3b): the two class I models gain 2.5 to 4.4 points per state task and 1.9 to 3.0 points on RULER, and the four class II models 17.1 to 18.3 and 5.8 to 6.6 points. The 95% intervals of the two class means do not overlap on any task (Appendix G).

The gain is stable across scales. Along the context length, head dimension and number of layers, the gain stays between 15.6 and 18.2 points, with its smallest value at 4k tokens (Figure 3c). Before the repair, the share of overshooting steps stays between 3.3% and 3.5% on all three axes (Appendix G).

NormFix also improves other recipes. It adds 5.9 points to QuaRot and 8.8 points to SpinQuant, less than on the default recipe, so the gains of rotation and of NormFix partly overlap (Appendix G).

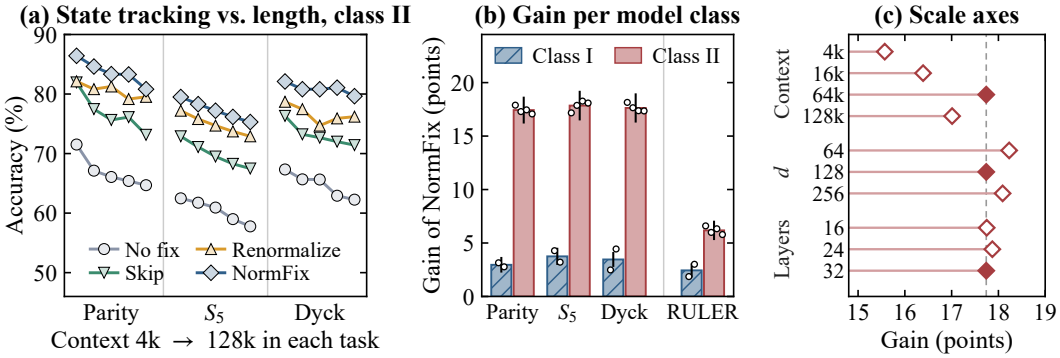

Figure 3: Results across lengths, models and scales. (a) State-tracking accuracy of the default class II model from 4k to 128k tokens (five points per task) without repair, with skipping of the steps with $\beta s > 2$, with key renormalization and with NormFix. (b) Gain of NormFix per task; bars are class means with 95% intervals, and dots are the individual models, two in class I and four in class II. (c) Gain on the state tasks at each setting of three scale axes: the context length, the head dimension d and the number of layers. Filled diamonds mark the default setting, and the dashed line is its gain.

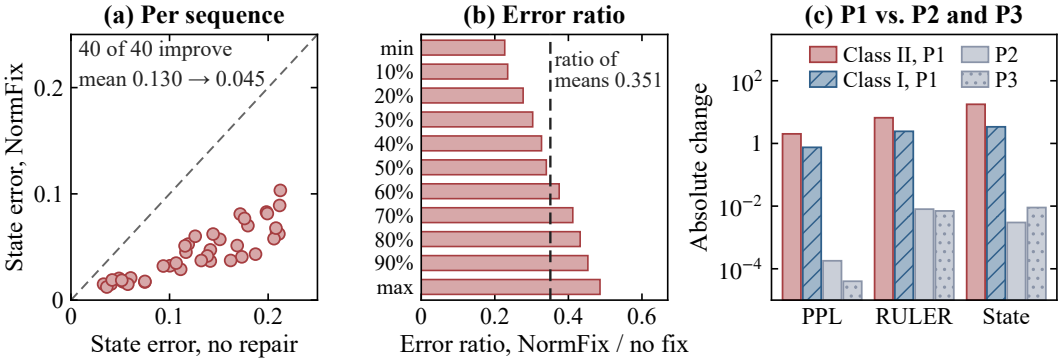

Figure 4: State error after the repair. (a) Norm of the state error $\boldsymbol{E}_L$ of 40 class II sequences without and with NormFix; the dashed line is $y = x$. (b) Ratio of the state error with NormFix to that without: minimum, deciles and maximum over the 40 sequences, and the ratio of the means (dashed). (c) Absolute change caused by NormFix in perplexity at 128k tokens and in RULER and state-tracking accuracy (points), on P1, where $s \neq 1$, and on P2 and P3, where $s = 1$, on a logarithmic axis; class I averages its two models.

5.5 State Error Comparison

We return to the state error $\boldsymbol{E}_L$, the quantity that Theorem 4 bounds, and measure its norm on 40 class II sequences with and without NormFix. The repair lowers the state error on every sequence, to between 0.23 and 0.49 of its unrepaired value, and the mean falls from 0.130 to 0.045 (Figure 4a,b). On P2 and P3, where $s = 1$, the repair changes perplexity by at most 0.0002 and accuracy by at most 0.01 points, more than two orders of magnitude below its effect on P1 even in class I (Figure 4c).

6 Conclusion

We studied the step size of delta-rule updates when the key is quantized after its normalizer. The kernel then takes the effective step size βs, which changes every correction by a relative step-size error of $|s - 1|$ and, where β is close to 2, overshoots on a few percent of the steps. If β and s are independent, the overshoot rate follows from the tail of β near 2 and the distribution of s. NormFix divides the step size by the squared norm of the received key, which restores the residual factor $1 - \beta$ of the trained model and makes the propagation of state errors nonexpansive for any quantizer. It recovers roughly half of the quantization loss in state tracking at less than 4% extra kernel time.

Limitations. We evaluate NormFix on models with 340M parameters, and our analysis covers scalar step sizes; vector-valued step sizes such as those of RWKV-7 need their own treatment. NormFix corrects the norm of the quantized key but leaves its direction error in place, as Theorem 4 shows.

Reproducibility Statement

Section 5.1 and Appendix B describe the models, the training data, the quantization recipes, the placements of the quantizer, the tasks and the measurement protocol. Appendix A contains proofs of all results, Appendix H gives the kernel change and its cost, and Appendices C to G list the numbers behind all tables and most figures. We will release the recorded steps, the kernel patch and the scripts.

AI Use Statement

Generative AI tools were used only to edit the wording of text written by the authors, to search for related work and check bibliographic details, and to write code for plots, tables and typesetting. They were not used for any task that the ICLR 2027 AI policy requires authors to disclose: developing theory, formulating or proving mathematical claims, proposing or refining hypotheses, designing, implementing or running the method or experiments, generating, cleaning or analyzing data, translating text, or interpreting results. The models in our experiments are objects of study, not writing aids. The authors reviewed all AI-assisted work and take full responsibility for the content of this paper.

References

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 100213–100240, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 113506–113543, 2025.

C. Caraiscos and Bede Liu. A roundoff error analysis of the LMS adaptive algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(1):34–41, 1984. doi: 10.1109/TASSP.1984.1164286.

Hung-Yueh Chiang, Chi-Chih Chang, Natalia Frumkin, Kai-Chiang Wu, Mohamed S. Abdelfattah, and Diana Marculescu. Quamba2: A robust and scalable post-training quantization framework for selective state space models. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10411–10427. PMLR, 2025a.

Hung-Yueh Chiang, Chi-Chih Chang, Natalia Frumkin, Kai-Chiang Wu, and Diana Marculescu. Quamba: A post-training quantization recipe for selective state space models. In *International Conference on Learning Representations (ICLR)*, 2025b.

John M. Cioffi. Limited-precision effects in adaptive filtering. *IEEE Transactions on Circuits and Systems*, 34(7):821–833, 1987. doi: 10.1109/TCS.1987.1086209.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 2024.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky hierarchy. In *International Conference on Learning Representations (ICLR)*, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 30318–30332, 2022. Also known as LLM.int8() (arXiv:2208.07339).

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023. Also known as GPTQ (arXiv:2210.17323).

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4th edition, 2013. ISBN 978-1-4214-0794-4.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *International Conference on Learning Representations (ICLR)*, 2025.

Simon Haykin. *Adaptive Filter Theory*. Prentice Hall, Upper Saddle River, NJ, 4th edition, 2002. ISBN 978-0-13-090126-2.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 1270–1303, 2024.

Alston S. Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM*, 5 (4):339–342, 1958. doi: 10.1145/320941.320947.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024.

Stefan Kaczmarz. Angenäherte Auflösung von Systemen linearer Gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles, Série A, Sciences Mathématiques*, pp. 355–357, 1937.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020.

Kimi Team. Kimi Linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 6, pp. 87–100, 2024.

Bo Liu, Rui Wang, Lemeng Wu, Yihao Feng, Peter Stone, and Qiang Liu. Longhorn: State space models are amortized online learners. In *International Conference on Learning Representations (ICLR)*, 2025a.

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

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

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 *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344. PMLR, 2024b.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35492–35506. PMLR, 2024.

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

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Jin-Ichi Nagumo and Atsuhiko Noda. A learning method for system identification. *IEEE Transactions on Automatic Control*, 12(3):282–287, 1967. doi: 10.1109/TAC.1967.1098599.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 30811–30849, 2024.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S. Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. RWKV-7 "goose" with expressive dynamic state evolution. In *Conference on Language Modeling (COLM)*, 2025.

Qwen Team. Qwen3-Next: Towards ultimate training & inference efficiency. Blog post, https://qwen.ai/blog?id=4074cca80393150c248e508aa62983f9cb7d27cd, September 2025.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations (ICLR)*, 2020.

Ali H. Sayed. *Fundamentals of Adaptive Filtering*. Wiley-IEEE Press, 2003. ISBN 978-0-471-46126-5.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 2021.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via Householder products. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 153738–153782, 2025.

Dirk T. M. Slock. On the convergence behavior of the LMS and the normalized LMS algorithms. *IEEE Transactions on Signal Processing*, 41(9):2811–2825, 1993. doi: 10.1109/78.236504.

Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009. doi: 10.1007/s00041-008-9030-4.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57503–57522. PMLR, 2025.

Ke Alexander Wang, Jiaxin Shi, and Emily B. Fox. Test-time regression: a unifying framework for designing sequence models with associative memory. *arXiv preprint arXiv:2501.12352*, 2025.

Bernard Widrow and Marcian E. Hoff. Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pp. 96–104, New York, 1960. IRE.

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 *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023.

Zukang Xu, Yuxuan Yue, Xing Hu, Zhihang Yuan, Zixu Jiang, Zhixuan Chen, Jiangyong Yu, Chen Xu, Sifan Zhou, and Dawei Yang. MambaQuant: Quantizing the Mamba family with variance aligned rotation methods. In *International Conference on Learning Representations (ICLR)*, 2025.

Songlin Yang and Yu Zhang. FLA: A Triton-based library for hardware-efficient implementations of linear attention mechanism. GitHub repository, https://github.com/fla-org/flash-linear-attention, January 2024.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 115491–115522, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations (ICLR)*, 2025.

CONTENTS OF THE APPENDIX

A Proofs and Derivations **14**
A.1 The Delta Rule as a Normalized Least-Mean-Squares Step . . . 15
A.2 Proof of Lemma 1 . . . 15
A.3 Proof of Proposition 2 . . . 15
A.4 Proof of Theorem 3 . . . 16
A.5 Proof of Theorem 4 and Growth of the State Error . . . 17
A.6 Proof of Proposition 5 and Comparison of the Write Terms . . . 18
A.7 Size of the Norm Error under Rounding . . . 18
A.8 Product-Form Updates . . . 19
A.9 A Worked Example . . . 20

B Experimental Details **20**
B.1 Models and Training . . . 20
B.2 Quantization Recipes and Placements . . . 20
B.3 Tasks and Metrics . . . 21
B.4 Recording, Timing and Statistics . . . 21

C Recorded Statistics **21**
C.1 Step Sizes across Layers . . . 21
C.2 Squared Key Norm . . . 22
C.3 Step-Size Bins . . . 22
C.4 Statistics per Model and Placement . . . 23
C.5 Effective Step Size . . . 24

D Controlled Experiments **25**
D.1 Radial and Tangential Shifts . . . 25
D.2 Test of the Margin Condition . . . 25
D.3 Step-Size Error in Class I . . . 26
D.4 Predicted Overshoot Rate . . . 26
D.5 Overshoot Share and Accuracy Loss across Sequences . . . 26

E Full Results **27**
E.1 Perplexity per Corpus . . . 27
E.2 Paired Differences . . . 28
E.3 Context Length . . . 28
E.4 Step Statistics of All Configurations . . . 29
E.5 State Error . . . 29

F Ablation Studies **30**
F.1 Clipping the Step Size . . . 30
F.2 Applying NormFix above a Threshold . . . 31
F.3 Repaired Layers . . . 32
F.4 Exempting Heads from the Repair . . . 32

G Results on Other Models and Settings **33**

H Kernel Implementation and Cost **36**

I Notation **37**

A PROOFS AND DERIVATIONS

Throughout, $\boldsymbol{S}$ is the state before a step, α the decay, β the step size, $\boldsymbol{v}$ the value, $\boldsymbol{k}$ the normalized key with $\|\boldsymbol{k}\|_2 = 1$, $\hat{\boldsymbol{k}} = \boldsymbol{k} + \boldsymbol{\epsilon}$ the key the kernel receives, $s = \|\hat{\boldsymbol{k}}\|_2^2$ its squared norm, $\bar{\boldsymbol{k}} = \hat{\boldsymbol{k}}/\sqrt{s}$ its

direction, and $\boldsymbol{r} = \alpha \boldsymbol{S}\hat{\boldsymbol{k}} - \boldsymbol{v}$ and $\boldsymbol{r}^+$ the residuals before and after the update. We drop the step index t whenever a statement involves one step only, and an unlabeled norm of a vector is the Euclidean norm. Table 33 in Appendix I lists the main symbols of the paper and where each one is defined.

A.1 The Delta Rule as a Normalized Least-Mean-Squares Step

The update (1) is one gradient step on the squared error of the state on the current key-value pair, with a step size divided by the squared norm of the key, as in the normalized least-mean-squares algorithm.

Take the loss $J(\boldsymbol{S}) = \frac{1}{2}\|\boldsymbol{S}\boldsymbol{k} - \boldsymbol{v}\|_2^2$ and a gradient step of size μ from the decayed state $\alpha\boldsymbol{S}$:

$$\begin{aligned} \nabla J(\alpha \boldsymbol{S}) &= (\alpha \boldsymbol{S}\boldsymbol{k} - \boldsymbol{v})\, \boldsymbol{k}^\top, \\ \boldsymbol{S}' &= \alpha \boldsymbol{S} - \mu\, (\alpha \boldsymbol{S}\boldsymbol{k} - \boldsymbol{v})\, \boldsymbol{k}^\top = \alpha \boldsymbol{S}\big(\boldsymbol{I} - \mu\, \boldsymbol{k}\boldsymbol{k}^\top\big) + \mu\, \boldsymbol{v}\boldsymbol{k}^\top, \\ \boldsymbol{S}'\boldsymbol{k} - \boldsymbol{v} &= (\alpha \boldsymbol{S}\boldsymbol{k} - \boldsymbol{v}) - \mu\, (\alpha \boldsymbol{S}\boldsymbol{k} - \boldsymbol{v})\, \boldsymbol{k}^\top \boldsymbol{k} = \big(1 - \mu\|\boldsymbol{k}\|_2^2\big)(\alpha \boldsymbol{S}\boldsymbol{k} - \boldsymbol{v}), \\ \mu = \beta/\|\boldsymbol{k}\|_2^2 &\implies \boldsymbol{S}'\boldsymbol{k} - \boldsymbol{v} = (1 - \beta)\, (\alpha \boldsymbol{S}\boldsymbol{k} - \boldsymbol{v}), \\ \|\boldsymbol{k}\|_2 = 1 &\implies \boldsymbol{S}' = \alpha \boldsymbol{S}\big(\boldsymbol{I} - \beta\, \boldsymbol{k}\boldsymbol{k}^\top\big) + \beta\, \boldsymbol{v}\boldsymbol{k}^\top. \end{aligned} \tag{7}$$

The first line is the gradient, a rank-one matrix. The second is the gradient step, regrouped so that the first term erases the state along $\boldsymbol{k}$ and the second writes the value along $\boldsymbol{k}$. The third line reads the new state with the same key: the residual is multiplied by $1 - \mu\|\boldsymbol{k}\|_2^2$, so what survives depends on the product of the step size and the squared length of the key. The fourth line is the choice of Nagumo & Noda (1967), which divides the step size by the squared norm so that the surviving fraction is $1 - \beta$ for a key of any length. The last line is (1): a delta-rule model performs the division once, through its normalizer, and from then on the kernel uses $\mu = \beta$ on the assumption that $\|\boldsymbol{k}\|_2 = 1$.

With a quantizer after the normalizer, the kernel receives $\hat{\boldsymbol{k}}$ and still uses $\mu = \beta$, so the third line gives the surviving fraction $1 - \beta s$ in place of $1 - \beta$. NormFix restores the fourth line for the key that the kernel holds, $\mu = \beta/s$. The derivation does not depend on how $\hat{\boldsymbol{k}}$ came to differ from $\boldsymbol{k}$.

A.2 Proof of Lemma 1

The update multiplies the residual by $1 - \beta s$ for any received key. Because the proof uses no property of the quantizer, the lemma can be checked on recorded steps without a model of the quantizer.

Since $\alpha \boldsymbol{S}_{t-1}(\boldsymbol{I} - \beta \hat{\boldsymbol{k}}\hat{\boldsymbol{k}}^\top) + \beta \boldsymbol{v}\hat{\boldsymbol{k}}^\top = \alpha \boldsymbol{S}_{t-1} - \beta(\alpha \boldsymbol{S}_{t-1}\hat{\boldsymbol{k}} - \boldsymbol{v})\hat{\boldsymbol{k}}^\top$,

$$\begin{aligned} \boldsymbol{S}_t &= \alpha \boldsymbol{S}_{t-1} - \beta\, \boldsymbol{r}\, \hat{\boldsymbol{k}}^\top, \\ \boldsymbol{S}_t\hat{\boldsymbol{k}} &= \alpha \boldsymbol{S}_{t-1}\hat{\boldsymbol{k}} - \beta\, \boldsymbol{r}\, (\hat{\boldsymbol{k}}^\top \hat{\boldsymbol{k}}) = \alpha \boldsymbol{S}_{t-1}\hat{\boldsymbol{k}} - \beta s\, \boldsymbol{r}, \\ \boldsymbol{r}^+ &= \boldsymbol{S}_t\hat{\boldsymbol{k}} - \boldsymbol{v} = (\alpha \boldsymbol{S}_{t-1}\hat{\boldsymbol{k}} - \boldsymbol{v}) - \beta s\, \boldsymbol{r} = (1 - \beta s)\, \boldsymbol{r}. \end{aligned} \tag{8}$$

Only $\hat{\boldsymbol{k}}^\top \hat{\boldsymbol{k}} = s$ enters, so the identity holds for any received key, quantized or not, and for any decay, which enters only through $\boldsymbol{r}$. □

A.3 Proof of Proposition 2

Both regimes follow from the sign and size of $1 - \beta s$, and the margin condition from simple algebra.

Parts (i) and (ii). Since $\beta > 0$ and $s > 0$,

$$\begin{aligned} |1 - \beta s| < 1 &\iff -1 < 1 - \beta s < 1 \iff 0 < \beta s < 2, \\ \frac{|(1 - \beta s) - (1 - \beta)|}{\beta} &= |s - 1| = |\varepsilon|, \\ \beta s > 2 &\iff 1 - \beta s < -1 \iff \|\boldsymbol{r}^+\| = (\beta s - 1)\, \|\boldsymbol{r}\| > \|\boldsymbol{r}\|. \end{aligned} \tag{9}$$

In the contracting regime the factor differs from the trained one by $\beta|\varepsilon|$, which is a relative step-size error of $|\varepsilon|$, and once βs passes 2 the residual changes sign and grows by the factor $\beta s - 1 > 1$.

Part (iii). If $\beta \le 1$, then $\beta s > 2$ forces $s > 2/\beta \ge 2$, that is $\varepsilon > 1$. With $\beta = 2 - m$ and $0 < m < 2$,

$$\begin{aligned}(2-m)(1+\varepsilon) > 2 &\iff 2 + 2\varepsilon - m - m\varepsilon > 2 \iff \varepsilon(2-m) > m \\ &\iff \varepsilon > \frac{m}{2-m} = \frac{m}{2} + \frac{m^2}{4} + \frac{m^3}{8} + O(m^4),\end{aligned} \tag{10}$$

so for a small margin the threshold is half the margin. □

For the state error, the relevant matrix is $\alpha(\boldsymbol{I} - \beta s \bar{\boldsymbol{k}}\bar{\boldsymbol{k}}^\top)$ with $\bar{\boldsymbol{k}} = \hat{\boldsymbol{k}}/\sqrt{s}$. It is symmetric with eigenvalue $\alpha(1-\beta s)$ along $\bar{\boldsymbol{k}}$ and α in every orthogonal direction, so its spectral norm is $\alpha \max(1, |1-\beta s|)$. This exceeds one only if $\alpha(\beta s - 1) > 1$, a stricter condition than $\beta s > 2$ whenever $\alpha < 1$: with $\alpha = 0.5$ and $\beta s = 2.1$ the factor is 0.55, which still contracts the state error although $\beta s > 2$.

A.4 Proof of Theorem 3

If β and s are independent, the share of steps with $\beta s > 2$ is the mean over s of the tail of β at $2/s$, and to first order it is linear in $(s-1)_+$ with twice the density of β just below 2 as the coefficient.

Exact form. Since $s > 0$, the event $\{\beta s > 2\}$ equals $\{\beta > 2/s\}$. Conditioning on s and using independence,

$$\begin{aligned}&\Pr[\beta s > 2 \mid s] = \Pr[\beta > 2/s \mid s] = \bar{G}(2/s), \\ &\Pr[\beta s > 2] = \mathbb{E}\big[\Pr[\beta s > 2 \mid s]\big] = \mathbb{E}\big[\bar{G}(2/s)\big] = \mathbb{E}\big[h(s-1)\big], \qquad h(\delta) = \bar{G}\Big(\frac{2}{1+\delta}\Big).\end{aligned} \tag{11}$$

The first line fixes s and uses independence to replace the conditional probability by the tail of β. The second averages over s and substitutes $\delta = s - 1$. For $\delta \le 0$ the threshold $2/(1+\delta)$ is at least 2, above every step size in $(0, 2)$, so $h(\delta) = 0$ and values of s below one never produce an overshoot. If $\beta \le 1$, the threshold is at least 1 for $\delta \le 1$, so $h(\delta) = 0$ there too, as in part (iii) of Proposition 2.

First-order form. Suppose $s \le s_{\max}$ and β has a density f on $(2/s_{\max}, 2)$ with a left limit $f(2^-)$ at 2 and $|f(b) - f(b')| \le \ell |b - b'|$ there. For $0 < \delta \le s_{\max} - 1$ let $w(\delta) = 2 - 2/(1+\delta) = 2\delta/(1+\delta)$ be the length of the interval of step sizes that overshoot. Then

$$\begin{aligned}h(\delta) &= \int_{2-w(\delta)}^{2} f(b)\, db = f(2^-)\, w(\delta) + \int_{2-w(\delta)}^{2} \big(f(b) - f(2^-)\big)\, db, \\ \Big|\int_{2-w}^{2} \big(f(b) - f(2^-)\big)\, db\Big| &\le \int_0^w \ell\, u\, du = \frac{\ell w^2}{2} = \frac{2\ell\,\delta^2}{(1+\delta)^2} \le 2\ell\,\delta^2, \\ h(\delta) &= \frac{2 f(2^-)\,\delta}{1+\delta} + R(\delta), \qquad |R(\delta)| \le 2\ell\,\delta^2, \\ \Pr[\beta s > 2] &= 2 f(2^-)\, \mathbb{E}\Big[\frac{(s-1)_+}{s}\Big] + \mathbb{E}\big[R\big((s-1)_+\big)\big], \\ \big|\mathbb{E}\big[R\big((s-1)_+\big)\big]\big| &\le 2\ell\, \mathbb{E}\big[(s-1)_+^2\big].\end{aligned} \tag{12}$$

The first line writes the tail as an integral of the density over the overshooting interval and splits it into its value at 2 times the length plus a remainder. The second bounds the remainder with the Lipschitz condition after substituting $u = 2 - b$ and collects the two terms. The last two lines take the expectation over s, using $h(\delta) = 0$ for $\delta \le 0$, so only $(s-1)_+$ enters, and $\delta/(1+\delta) = (s-1)/s$.

Without independence. For every $\delta \ge 0$, a step with $s \le 1 + \delta$ and $\beta s > 2$ has $\beta > 2/s \ge 2/(1+\delta)$. Hence

$$\{\beta s > 2\} \subseteq \Big\{\beta > \frac{2}{1+\delta}\Big\} \cup \{s > 1 + \delta\} \quad \Longrightarrow \quad \Pr[\beta s > 2] \le h(\delta) + \Pr[s > 1 + \delta], \tag{13}$$

for any joint distribution of β and s. On the recorded steps of the default model this bound is about 6.5% at its best δ, about twice the measured rate, while the form (11) that assumes independence comes within 7% of the measured rate. □

Consequences. To first order, the rate is twice the density of β just below 2, which the model sets, times the mean positive part of $s - 1$, which the quantizer sets. Overshoots occur even when s is one on average, as long as some steps have $s > 1$. More bits reduce the rate only through $\mathbb{E}[(s-1)_+]$,

never to zero while β has mass near 2 (Corollary 8). A step size in $(0, 1)$ contributes nothing for any quantizer with $s < 2$, since its tail vanishes on the interval $[1, 2)$, so class I models cannot overshoot.

Measuring the two factors. A perturbation that shifts s by the same δ at every step makes the rate $h(\delta)$, by the first line of (11) with $s \equiv 1 + \delta$, if the step sizes stay unchanged. The radial and tangential shifts of Appendix D.1 are such perturbations and sample h at ten points between $\delta = 3 \times 10^{-5}$ and $\delta = 0.092$. For the prediction of Figure 2b we interpolate h linearly between its measured points, with $h(0) = 0$, and average it over the recorded values of s in each cell (Table 15). The measured h also gives the density near 2: $h(0.002) = 0.8\%$ implies an average density of about 2 over $(1.996, 2)$, six times the average 0.33 over $[1.8, 2)$, so the class II step sizes concentrate near 2, where a small shift of s already produces overshoots. The average density over the overshooting interval, $h(\delta)/w(\delta)$, falls from about 2 at $\delta = 0.002$ to 0.39 at $\delta = 0.092$, so the first-order form (12) holds only for fine grids, as in Corollary 8, and the predictions of Table 15 use the exact form (11).

A.5 Proof of Theorem 4 and Growth of the State Error

With NormFix the matrix that carries the state error from step to step is nonexpansive, so the deployed error is at most the sum of the local errors. The same algebra shows how the unrepaired kernel differs.

Error recursion. The deployed and the unquantized updates are $\boldsymbol{S}_t = \alpha_t \boldsymbol{S}_{t-1}(\boldsymbol{I} - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top) + \gamma_t \boldsymbol{v}_t \hat{\boldsymbol{k}}_t^\top$ and $\boldsymbol{S}_t^\star = \alpha_t \boldsymbol{S}_{t-1}^\star(\boldsymbol{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top) + \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^\top$. Subtracting them and adding and subtracting $\alpha_t \boldsymbol{S}_{t-1}^\star(\boldsymbol{I} - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top)$,

$$\begin{aligned}
\boldsymbol{E}_t &= \alpha_t \boldsymbol{S}_{t-1}\big(\boldsymbol{I} - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top\big) - \alpha_t \boldsymbol{S}_{t-1}^\star\big(\boldsymbol{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top\big) + \boldsymbol{v}_t\big(\gamma_t \hat{\boldsymbol{k}}_t - \beta_t \boldsymbol{k}_t\big)^\top \\
&= \alpha_t\big(\boldsymbol{S}_{t-1} - \boldsymbol{S}_{t-1}^\star\big)\big(\boldsymbol{I} - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top\big) + \boldsymbol{v}_t\big(\gamma_t \hat{\boldsymbol{k}}_t - \beta_t \boldsymbol{k}_t\big)^\top \\
&\quad + \alpha_t \boldsymbol{S}_{t-1}^\star\big[\big(\boldsymbol{I} - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top\big) - \big(\boldsymbol{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top\big)\big] \\
&= \underbrace{\boldsymbol{E}_{t-1}\, \alpha_t\big(\boldsymbol{I} - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top\big)}_{\boldsymbol{E}_{t-1}\boldsymbol{A}_t} + \underbrace{\alpha_t \boldsymbol{S}_{t-1}^\star\big(\beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top - \gamma_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top\big) + \boldsymbol{v}_t\big(\gamma_t \hat{\boldsymbol{k}}_t - \beta_t \boldsymbol{k}_t\big)^\top}_{\boldsymbol{F}_t}.
\end{aligned} \tag{14}$$

The first line subtracts the two updates, the next two add and subtract the unquantized state times the deployed erase factor to separate the carried error from the new one, and the last names the two parts.

Part (i). Under NormFix, $\gamma_t = \eta_t = \beta_t / s_t$ and $\eta_t \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top = \beta_t \bar{\boldsymbol{k}}_t \bar{\boldsymbol{k}}_t^\top$. The matrix $\boldsymbol{A}_t$ is symmetric and

$$\begin{aligned}
\boldsymbol{A}_t \boldsymbol{z} &= \alpha_t \boldsymbol{z} \quad \text{for every } \boldsymbol{z} \perp \bar{\boldsymbol{k}}_t, \qquad \boldsymbol{A}_t \bar{\boldsymbol{k}}_t = \alpha_t\big(\bar{\boldsymbol{k}}_t - \beta_t \bar{\boldsymbol{k}}_t\big) = \alpha_t(1 - \beta_t)\, \bar{\boldsymbol{k}}_t, \\
\|\boldsymbol{A}_t\|_2 &= \alpha_t \max\big(1, |1 - \beta_t|\big) = \alpha_t \qquad \text{for every } \beta_t \in (0, 2] \text{ and every } s_t > 0.
\end{aligned} \tag{15}$$

For any matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}_i^\top$ and any symmetric $\boldsymbol{A}$, the rows of $\boldsymbol{X}\boldsymbol{A}$ are $(\boldsymbol{A}\boldsymbol{x}_i)^\top$, so $\|\boldsymbol{X}\boldsymbol{A}\|_F^2 = \sum_i \|\boldsymbol{A}\boldsymbol{x}_i\|^2 \le \|\boldsymbol{A}\|_2^2 \|\boldsymbol{X}\|_F^2$. The triangle inequality and induction on L from $\boldsymbol{E}_0 = \boldsymbol{0}$ give

$$\begin{aligned}
\|\boldsymbol{E}_t\|_F &\le \|\boldsymbol{E}_{t-1}\boldsymbol{A}_t\|_F + \|\boldsymbol{F}_t\|_F \le \alpha_t \|\boldsymbol{E}_{t-1}\|_F + \|\boldsymbol{F}_t\|_F, \\
\|\boldsymbol{E}_L\|_F &\le \sum_{t=1}^{L} \Big(\prod_{j=t+1}^{L} \alpha_j \Big) \|\boldsymbol{F}_t\|_F \le \sum_{t=1}^{L} \|\boldsymbol{F}_t\|_F.
\end{aligned} \tag{16}$$

Part (ii). Without the repair, $\gamma_t = \beta_t$ and the same computation gives the eigenvalue $\alpha_t(1 - \beta_t s_t)$ along $\bar{\boldsymbol{k}}_t$, hence $\|\boldsymbol{A}_t\|_2 = \alpha_t \max(1, |1 - \beta_t s_t|)$. Since $\beta_t s_t > 0$, $|1 - \beta_t s_t| > 1$ exactly when $\beta_t s_t > 2$, and the norm is then $\alpha_t(\beta_t s_t - 1)$, which exceeds one exactly when $\alpha_t(\beta_t s_t - 1) > 1$.

Part (iii). Write $\boldsymbol{A}_t^{\mathrm{raw}}$ and $\boldsymbol{F}_t^{\mathrm{raw}}$ for the terms of the unrepaired kernel, $\gamma_t = \beta_t$, and compare them with the NormFix terms, using $\beta_t - \eta_t = \beta_t(s_t - 1)/s_t$:

$$\begin{aligned}
\boldsymbol{A}_t^{\mathrm{raw}} - \boldsymbol{A}_t^{\mathrm{NF}} &= -\alpha_t(\beta_t - \eta_t)\, \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top, \\
\boldsymbol{F}_t^{\mathrm{raw}} - \boldsymbol{F}_t^{\mathrm{NF}} &= -\alpha_t(\beta_t - \eta_t)\, \boldsymbol{S}_{t-1}^\star \hat{\boldsymbol{k}}_t \hat{\boldsymbol{k}}_t^\top + (\beta_t - \eta_t)\, \boldsymbol{v}_t \hat{\boldsymbol{k}}_t^\top, \\
\boldsymbol{E}_{t-1}\boldsymbol{A}_t^{\mathrm{raw}} + \boldsymbol{F}_t^{\mathrm{raw}} &= \boldsymbol{E}_{t-1}\boldsymbol{A}_t^{\mathrm{NF}} + \boldsymbol{F}_t^{\mathrm{NF}} - (\beta_t - \eta_t)\big(\alpha_t \boldsymbol{E}_{t-1} \hat{\boldsymbol{k}}_t + \alpha_t \boldsymbol{S}_{t-1}^\star \hat{\boldsymbol{k}}_t - \boldsymbol{v}_t\big) \hat{\boldsymbol{k}}_t^\top \\
&= \boldsymbol{E}_{t-1}\boldsymbol{A}_t^{\mathrm{NF}} + \boldsymbol{F}_t^{\mathrm{NF}} - (\beta_t - \eta_t)\big(\alpha_t \boldsymbol{S}_{t-1} \hat{\boldsymbol{k}}_t - \boldsymbol{v}_t\big) \hat{\boldsymbol{k}}_t^\top \\
&= \boldsymbol{E}_{t-1}\boldsymbol{A}_t^{\mathrm{NF}} + \boldsymbol{F}_t^{\mathrm{NF}} - \beta_t \frac{s_t - 1}{s_t}\, \boldsymbol{r}_t \hat{\boldsymbol{k}}_t^\top.
\end{aligned} \tag{17}$$

The difference of the two recursions collects every extra term into one rank-one matrix along $\hat{\boldsymbol{k}}_t$, and $\boldsymbol{E}_{t-1} + \boldsymbol{S}^\star_{t-1} = \boldsymbol{S}_{t-1}$ turns this matrix into $s_t - 1$ times the correction $-\eta_t \boldsymbol{r}_t \hat{\boldsymbol{k}}_t^\top$ that NormFix would make from the unrepaired state. □

Corollary 6 (growth of the state error)**.** *Under NormFix, if* $\mathbb{E}\langle \boldsymbol{E}_{t-1}\boldsymbol{A}_t, \boldsymbol{F}_t\rangle_F = 0$ *and* $\mathbb{E}\|\boldsymbol{F}_t\|_F^2 \le \rho^2$ *for every* t*, then* $\mathbb{E}\|\boldsymbol{E}_L\|_F^2 \le L\rho^2$*, so the root-mean-square state error grows at most like* $\rho\sqrt{L}$*.*

Proof. Expand the square of the recursion, take expectations and drop the cross term:

$$\begin{aligned} &\mathbb{E}\|\boldsymbol{E}_t\|_F^2 = \mathbb{E}\|\boldsymbol{E}_{t-1}\boldsymbol{A}_t\|_F^2 + 2\,\mathbb{E}\langle \boldsymbol{E}_{t-1}\boldsymbol{A}_t, \boldsymbol{F}_t\rangle_F + \mathbb{E}\|\boldsymbol{F}_t\|_F^2 \le \mathbb{E}\|\boldsymbol{E}_{t-1}\|_F^2 + \rho^2, \\ &\mathbb{E}\|\boldsymbol{E}_L\|_F^2 \le L\,\rho^2. \end{aligned} \tag{18}$$

The inequality uses $\|\boldsymbol{E}_{t-1}\boldsymbol{A}_t\|_F \le \|\boldsymbol{A}_t\|_2 \|\boldsymbol{E}_{t-1}\|_F \le \|\boldsymbol{E}_{t-1}\|_F$, which holds at every step because $\|\boldsymbol{A}_t\|_2 = \alpha_t \le 1$, and the second line follows by induction from $\boldsymbol{E}_0 = \boldsymbol{0}$. □

The same argument applies to the unrepaired kernel of class I, whose $\boldsymbol{A}_t$ is also nonexpansive because $\beta_t s_t < 2$ on every recorded step. There, the state error left by perturbing every key in Figure 1c grows like $L^{0.51}$ from 1k to 128k tokens, by a factor of 11 to 13 over this range. In class II, which the corollary does not cover without the repair, the error grows at about the same rate and is 2.4 to 2.8 times as large as in class I at every length and for each of the four perturbation sizes (Table 21).

A.6 Proof of Proposition 5 and Comparison of the Write Terms

All scalar repairs that restore the factor $1 - \beta$ share one erase term and differ only in the size of the write term. This section proves it and then compares the two members that matter most in practice.

Let a repair use the key $c\hat{\boldsymbol{k}}$ and the step size $g\beta$. By Lemma 1 applied to the key $c\hat{\boldsymbol{k}}$, whose squared norm is $c^2 s$,

$$\begin{aligned} \boldsymbol{S}_t\,(c\hat{\boldsymbol{k}}) - \boldsymbol{v} &= \big(1 - g\beta\, c^2 s\big)\big(\alpha \boldsymbol{S}_{t-1}(c\hat{\boldsymbol{k}}) - \boldsymbol{v}\big), \\ 1 - g\beta c^2 s = 1 - \beta \text{ for every state } &\iff g c^2 s = 1, \\ g\beta\,(c\hat{\boldsymbol{k}})(c\hat{\boldsymbol{k}})^\top &= g c^2 s\,\beta\,\bar{\boldsymbol{k}}\bar{\boldsymbol{k}}^\top = \beta\,\bar{\boldsymbol{k}}\bar{\boldsymbol{k}}^\top, \\ g\beta\,\boldsymbol{v}\,(c\hat{\boldsymbol{k}})^\top &= g c\sqrt{s}\,\beta\,\boldsymbol{v}\bar{\boldsymbol{k}}^\top = \frac{\beta}{c\sqrt{s}}\,\boldsymbol{v}\bar{\boldsymbol{k}}^\top. \end{aligned} \tag{19}$$

The residual is nonzero for a generic state, so the condition of the second line is also necessary. The last two lines give the erase factor $\alpha(\boldsymbol{I} - \beta\bar{\boldsymbol{k}}\bar{\boldsymbol{k}}^\top)$ that all members share and the write term, $\beta/(c\sqrt{s})$ times $\boldsymbol{v}\bar{\boldsymbol{k}}^\top$. NormFix has $c = 1$ and $g = 1/s$, and renormalization has $c = 1/\sqrt{s}$ and $g = 1$. □

The two write terms. The trained model writes $\beta\boldsymbol{v}\boldsymbol{k}^\top$ along a unit key. Renormalization writes $\beta\boldsymbol{v}\bar{\boldsymbol{k}}^\top$, the trained magnitude along the received direction, and NormFix writes $(\beta/\sqrt{s})\boldsymbol{v}\bar{\boldsymbol{k}}^\top$, which keeps the step size of the received key. Their states after one step from the same $\boldsymbol{S}_{t-1}$ differ by $\beta(1-1/\sqrt{s})\boldsymbol{v}\bar{\boldsymbol{k}}^\top$, whose relative size $|1-1/\sqrt{s}|$ is at most 0.044 over the recorded values of s. Each repair multiplies the residual read with its own key by $1 - \beta$, renormalization the residual read with $\bar{\boldsymbol{k}}$ and NormFix the residual read with $\hat{\boldsymbol{k}}$. The algebra does not rank the two write terms; in Table 1, NormFix is ahead by 3.9 points in class II and 2.2 in class I, with a faster kernel (Appendix H).

A.7 Size of the Norm Error under Rounding

To relate the size of $s - 1$ to the key format, we derive the mean and the variance of s under a simple model of rounding, and then turn Theorem 3 into a statement about the grid step of the format.

Proposition 7 (moments of s)**.** *Let* $\|\boldsymbol{k}\|_2 = 1$ *and* $\hat{\boldsymbol{k}} = \boldsymbol{k} + \boldsymbol{\epsilon}$ *with independent, zero-mean and symmetric coordinates* ϵ_i *of variance* σ_i^2*. Then* $\mathbb{E}[s] = 1 + \sum_i \sigma_i^2$ *and* $\mathrm{Var}[s] = 4\sum_i k_i^2\sigma_i^2 + \sum_i(\mathbb{E}[\epsilon_i^4] - \sigma_i^4)$*. For errors uniform on* $[-\tau/2, \tau/2]$*, the usual model of rounding to a grid of step* τ *in each of the* d *coordinates, it follows that* $\mathbb{E}[s] - 1 = d\tau^2/12$ *and that* $\mathrm{Var}[s] = \tau^2/3 + d\tau^4/180$*.*

Proof. Expand the squared norm and use that odd moments of a symmetric error vanish:

$$
\begin{aligned}
s &= \|\boldsymbol{k}\|_2^2 + 2\langle \boldsymbol{k}, \boldsymbol{\epsilon}\rangle + \|\boldsymbol{\epsilon}\|_2^2 = 1 + 2\sum_i k_i\epsilon_i + \sum_i \epsilon_i^2,\\
\mathbb{E}[s] &= 1 + 2\sum_i k_i\,\mathbb{E}[\epsilon_i] + \sum_i \mathbb{E}[\epsilon_i^2] = 1 + \sum_i \sigma_i^2,\\
\mathrm{Var}[s] &= \mathrm{Var}\Big[2\sum_i k_i\epsilon_i\Big] + \mathrm{Var}\Big[\sum_i \epsilon_i^2\Big] + 2\,\mathrm{Cov}\Big[2\sum_i k_i\epsilon_i,\ \sum_j \epsilon_j^2\Big],\\
\mathrm{Cov}\big[\epsilon_i, \epsilon_j^2\big] &= \mathbb{E}[\epsilon_i\epsilon_j^2] - \mathbb{E}[\epsilon_i]\,\mathbb{E}[\epsilon_j^2] = 0,\\
\mathrm{Var}[s] &= 4\sum_i k_i^2\sigma_i^2 + \sum_i \big(\mathbb{E}[\epsilon_i^4] - \sigma_i^4\big).
\end{aligned}
\tag{20}
$$

Splitting s into one, a term linear in the error and a term quadratic in it, the linear term averages to zero and the quadratic term adds the total error variance, so the mean of s lies above one. The covariance of the two random terms vanishes, as an odd moment if $i = j$ and by independence otherwise. For uniform errors $\sigma_i^2 = \tau^2/12$ and $\mathbb{E}[\epsilon_i^4] = \tau^4/80$, and $\sum_i k_i^2 = 1$ gives $\mathrm{Var}[s] = \tau^2/3 + d(\tau^4/80 - \tau^4/144) = \tau^2/3 + d\tau^4/180$. □

Corollary 8 (finer grids reduce the rate but never remove it)**.** *Under the model of Proposition 7, let $s - 1$ be approximately normal with mean $\bar\varepsilon = d\tau^2/12$ and standard deviation σ_ε, $\sigma_\varepsilon^2 = \tau^2/3 + d\tau^4/180$. Then $\mathbb{E}[(s-1)_+] = \bar\varepsilon\Phi(\bar\varepsilon/\sigma_\varepsilon) + \sigma_\varepsilon\varphi(\bar\varepsilon/\sigma_\varepsilon)$, where Φ and φ are the standard normal distribution and density, and as $\tau \to 0$, with β and s independent,*

$$
\mathbb{E}\big[(s-1)_+\big] = \frac{\tau}{\sqrt{6\pi}} + \frac{d\tau^2}{24} + O(\tau^3), \qquad \Pr[\beta s > 2] = \frac{2f(2^-)}{\sqrt{6\pi}}\,\tau + O(\tau^2). \tag{21}
$$

Proof. For $X \sim \mathcal{N}(\bar\varepsilon, \sigma_\varepsilon^2)$, substitute $x = \bar\varepsilon + \sigma_\varepsilon z$ and use $\varphi'(z) = -z\varphi(z)$:

$$
\begin{aligned}
\mathbb{E}[X_+] &= \int_{-\bar\varepsilon/\sigma_\varepsilon}^{\infty} (\bar\varepsilon + \sigma_\varepsilon z)\,\varphi(z)\,dz = \bar\varepsilon\,\Phi(\bar\varepsilon/\sigma_\varepsilon) + \sigma_\varepsilon\,\varphi(\bar\varepsilon/\sigma_\varepsilon),\\
\Phi(r) &= \tfrac{1}{2} + r\,\varphi(0) + O(r^3), \qquad \varphi(r) = \varphi(0)\big(1 - \tfrac{r^2}{2}\big) + O(r^4),\\
\mathbb{E}[X_+] &= \sigma_\varepsilon\varphi(0) + \frac{\bar\varepsilon}{2} + \frac{\bar\varepsilon^2}{2\sigma_\varepsilon}\varphi(0) + O\Big(\frac{\bar\varepsilon^4}{\sigma_\varepsilon^3}\Big).
\end{aligned}
\tag{22}
$$

With $\sigma_\varepsilon = (\tau/\sqrt{3})(1 + O(d\tau^2))$ the ratio $\bar\varepsilon/\sigma_\varepsilon = O(d\tau)$ vanishes with τ, $\sigma_\varepsilon\varphi(0) = \tau/\sqrt{6\pi}$ and $\bar\varepsilon/2 = d\tau^2/24$. The second equation of (21) is the first-order form of Theorem 3, whose remainder is $O(\mathbb{E}[(s-1)_+^2]) = O(\tau^2)$; the normal model has an unbounded upper tail, but its mass above any fixed $s_{\max}$ vanishes faster than every power of τ. □

For fine grids the rate is linear in the grid step: one more bit halves τ and, once $\bar\varepsilon \ll \sigma_\varepsilon$, roughly halves the rate, but no finite number of bits removes it while $f(2^-) > 0$. This is the quantitative form of the statement that more bits shrink $|s-1|$ without bounding βs. As an order-of-magnitude check for $d = 128$, a symmetric 4-bit grid set by a largest coordinate near 0.29 has $\tau \approx 0.041$ and gives $\mathbb{E}[s] - 1 \approx 0.018$ and a standard deviation near 0.024. On P1 with INT4 keys the recorded median of $s - 1$ is 0.013, and the 10th and 90th percentiles imply a standard deviation of about 0.020. The rough model thus gives the right order of magnitude for both the center and the spread of s on P1.

A.8 Product-Form Updates

DeltaProduct applies n_h factors per token, $\boldsymbol{S}_t = \alpha_t \boldsymbol{S}_{t-1}\prod_{j=1}^{n_h}(\boldsymbol{I} - \beta_t^{(j)}\hat{\boldsymbol{k}}_t^{(j)}\hat{\boldsymbol{k}}_t^{(j)\top})$ plus the write terms, each with its own key and step size (Siems et al., 2025). The repair applies factor by factor.

Proposition 9 (product-form updates)**.** *Applying (4) to factor j with $\eta_t^{(j)} = \beta_t^{(j)}/s_t^{(j)}$ restores its eigenvalue to $1 - \beta_t^{(j)}$. A norm $\bar s_t$ shared by the factors gives factor j the step-size error $|s_t^{(j)} - \bar s_t|/\bar s_t$.*

Proof. Lemma 1 applied to factor j alone gives the eigenvalue $1 - \beta^{(j)}s^{(j)}$ along $\hat{\boldsymbol{k}}^{(j)}$, and the step size $\beta^{(j)}/s^{(j)}$ restores it to $1 - \beta^{(j)}$. With a shared norm, factor j uses $\beta^{(j)}/\bar s$, so its effective step size is $\beta^{(j)}s^{(j)}/\bar s$, which differs from $\beta^{(j)}$ by the relative error $|s^{(j)}/\bar s - 1|$. □

The factors of one token have different keys and hence different norms, so a norm shared by all factors fits none of them exactly, and each factor needs the squared norm of its own received key.

A.9 A Worked Example

The example step of this section has $\beta = 1.973$ and $s = 1.041$ on P1 with an INT4 key, so $\beta s = 2.054$. Table 3 follows it through several configurations of Table 1, each with its residual ratio.

Table 3: Effective step size of each configuration and the residual ratio $|1 - \beta s|$ it leaves at the example step ($\beta = 1.973$, $s = 1.041$). Weight-only and rotation recipes keep the key at 4 bits, so their effective step size is that of the unrepaired model up to the change of s they cause; $\mathbf{1}[\cdot]$ is the indicator.

Configuration	Effective step size	At the example step	$\lvert 1 - \beta s \rvert$
No fix	βs	2.0539	1.0539
Clip at 1.9	$\min(\beta, 1.9)\, s$	1.9779	0.9779
Renormalize	β	1.9730	0.9730
W6 weights	βs (key at 4 bits)	2.0539	1.0539
FP8 key	βs (s of the FP8 key)	1.9840	0.9840
QuaRot, SpinQuant	βs (key at 4 bits)	2.0470	1.0470
Skip if $\beta s > 2$	$\beta s\, \mathbf{1}[\beta s \le 2]$	0.0000	1.0000
NormFix	β	1.9730	0.9730

In the unquantized model the factor is $1 - \beta = -0.973$: the step is a near reflection that flips the residual and shrinks it to 97.3% of its norm. In the deployed model the effective step size is 2.054 and the factor is -1.054, so the same step flips the residual and enlarges it by 5.4%. Proposition 2(iii) predicts this from the margin: $m = 0.027$ gives the threshold $m/(2-m) = 0.0137$, and $s-1 = 0.041$ exceeds it. NormFix uses $\eta = 1.973/1.041 = 1.895$, so $\eta s = 1.973$ and the factor is -0.973 again. If the same key recurred 20 times without decay, the deployed residual would grow by $1.054^{20} = 2.86$, and under NormFix it would shrink to $0.973^{20} = 0.58$. Interpolating the recorded quantiles of s, a step size of 1.973 needs $s > 1.0137$ to overshoot, which happens on about half of the steps, while a step size of 1.9 would need $s > 1.053$, which happens on only a few percent of the recorded steps.

B Experimental Details

B.1 Models and Training

Class I contains DeltaNet (Yang et al., 2024) and Gated DeltaNet (Yang et al., 2025) with $\beta = \sigma(\cdot)$. Class II contains Gated DeltaNet with $\beta = 2\sigma(\cdot)$ (Grazzi et al., 2025) and DeltaProduct with $n_h \in \{1, 2, 3\}$ factors per token (Siems et al., 2025). All delta-rule models are trained from scratch with 340M parameters on 20B tokens of FineWeb-Edu (Penedo et al., 2024), with 32 layers, four heads per layer and head dimension $d = 128$ unless a scale axis of Appendix G changes them. A 1.3B DeltaNet is used in the step-size test of Appendix D.3. The quantized models are never retrained, except the QAT-W4 reference, which is retrained with 4-bit weights by quantization-aware training.

B.2 Quantization Recipes and Placements

The default recipe quantizes all linear weights to W4 with groups of 128 and the normalized key to INT4 with groups of 128 on P1. W4 per-channel uses one scale per output channel, W4A8 adds 8-bit activations, QuaRot+W4 and SpinQuant+W4 rotate the weights before 4-bit quantization (Ashkboos et al., 2024; Liu et al., 2025b), and the key-only recipes keep W4 weights and store the key in INT8 or in FP8 with the E4M3 format (Micikevicius et al., 2022). The baselines W6 weights and GPTQ or AWQ W4 (Frantar et al., 2023; Lin et al., 2024) keep the key at 4 bits, and QAT-W4 retrains the model with 4-bit weights (Liu et al., 2024a). Table 4 lists the range of β, the placement and the share of overshooting steps for every model under the default recipe, as recorded on the placement P1.

The placements differ in the order of normalizer and quantizer. On P1 the normalized key is quantized as an activation and used as it is; on P4 it is quantized on entry to the fused kernel in the kernel's own low-bit format. On P2 the key is quantized before the normalizer, and on P3 the kernel renormalizes the quantized key before use. On P2 and P3, $s = 1$ up to floating-point rounding, and on P1 and P4,

s carries the full rounding error. A recipe name such as W4 does not reveal the placement, which depends on the kernel and the serving code (Yang & Zhang, 2024). RWKV-7 and Mamba-2 are not listed, since neither has a scalar step size acting on a single key (Peng et al., 2025; Dao & Gu, 2024).

Table 4: Model summary under the default recipe: class, range of the step size β, placement and share of steps with $\beta s > 2$ (%) over the recorded steps.

Model	Class	Range of β	Placement	$\beta s{>}2$ (%)
DeltaNet	I	$(0, 1)$	P1	0.00
Gated DeltaNet	I	$(0, 1)$	P1	0.00
Gated DeltaNet, $\beta \in (0, 2)$	II	$(0, 2)$	P1	3.40
DeltaProduct $n_h{=}1$	II	$(0, 2)$	P1	3.41
DeltaProduct $n_h{=}2$	II	$(0, 2)$	P1	3.51
DeltaProduct $n_h{=}3$	II	$(0, 2)$	P1	3.39

B.3 Tasks and Metrics

Perplexity is measured on PG-19 (Rae et al., 2020) and Books3 (Gao et al., 2020) at 8k, 32k and 128k tokens, and the PPL columns report the geometric mean of the two corpora. RULER (Hsieh et al., 2024) is averaged over 4k, 16k, 64k and 128k tokens, and LongBench (Bai et al., 2024) averages its tasks. The state tasks are parity (Delétang et al., 2023), S_5 (Merrill et al., 2024) and Dyck languages; models are trained on them at 4k tokens and tested at up to 128k, and the main tables use 64k. The step-size test of Appendix D.3 uses associative recall (Arora et al., 2024) together with perplexity.

The share of overshooting steps is the fraction of recorded steps whose effective step size exceeds 2. The step-size error of a step is the relative deviation of its effective step size from the trained one, $|\gamma_t s_t - \beta_t|/\beta_t$ for a kernel that uses the step size γ_t, and the tables report its 99th percentile. In the ablations, the share of the gain kept is $100\,(\bar{y} - \bar{y}_0)/(\bar{y}_{\mathrm{NF}} - \bar{y}_0)$, where $\bar{y}$ is the mean of parity and S_5 accuracy at 64k tokens and $\bar{y}_0$ and $\bar{y}_{\mathrm{NF}}$ are its values without repair and with NormFix applied.

B.4 Recording, Timing and Statistics

For each model, placement and recipe we record 6144 steps from the four heads of six layers. The kernel writes β_t, s_t, $\|\boldsymbol{r}_t\|$ and $\|\boldsymbol{r}_t^+\|$ at each step; the identity of Lemma 1 is checked on the same records. Everything after the update, from the state error to the task scores, is rerun for every repair.

Kernel time is the median of five runs on one NVIDIA H100 GPU, per step and head. Throughput is the decode rate at batch size one, and memory is the peak allocated memory of the same runs. The intervals of Appendix E.2 are 95% confidence intervals of the paired difference over 4096 sequences.

C Recorded Statistics

This appendix lists the statistics of the recorded steps: the step sizes, the squared key norm s, their split over bins of β, the overshoots across models and placements, and the effective step size βs.

C.1 Step Sizes across Layers

The distribution of β depends on the class and hardly on the layer (Table 5). In class I the median lies between 0.575 and 0.606 in the six layer groups. In class II the median lies between 1.07 and 1.10, and 6.0% to 7.2% of the step sizes lie above 1.8 in every group (Figure 5a). The upper tail of β is therefore present throughout the depth of a class II model, and so are the steps that can overshoot.

Table 5: Median of the step size β and the share of β above 1.8 (%) in six layer groups.

Class	Layer group	Median β	$\Pr[\beta > 1.8]$ (%)
I	1	0.602	0.0
I	2	0.575	0.0
I	3	0.606	0.0
I	4	0.596	0.0
I	5	0.599	0.0
I	6	0.588	0.0
II	1	1.072	6.9
II	2	1.070	7.2
II	3	1.081	6.7
II	4	1.081	6.4
II	5	1.095	6.0
II	6	1.102	6.1

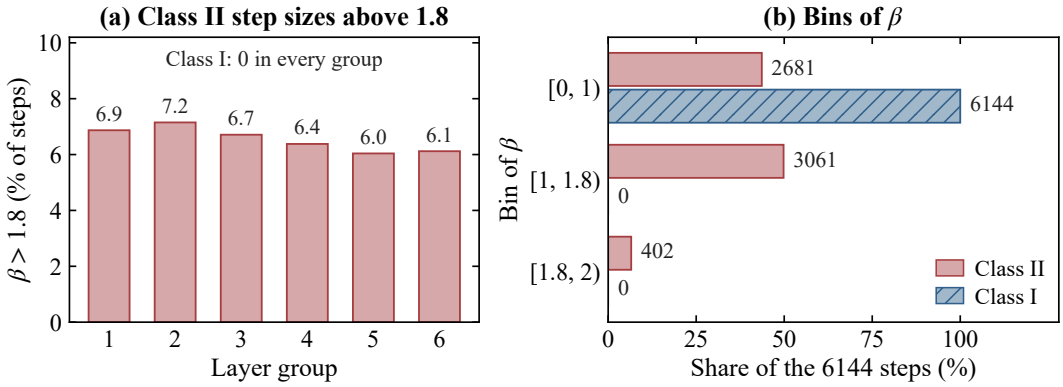

Figure 5: Step sizes. (a) Share of the class II step sizes above 1.8 in each layer group; in class I it is zero. (b) Share of the 6144 recorded step sizes of each class in the bins $[0, 1)$, $[1, 1.8)$ and $[1.8, 2)$, with the step counts.

C.2 Squared Key Norm

Table 6 gives the quantiles of s drawn in Figure 1b. On P2 and P3, where the normalizer runs last, all quantiles round to one. With INT4 keys the range from the 1st percentile to the maximum is 0.123 on P1 and 0.125 on P4, and the median is 1.013 and 1.010, so more than half of the steps have $s > 1$.

Table 6: Quantiles of the squared key norm s per placement with INT4 keys.

Placement	Key	1st	10th	50th	90th	99.9th	Max
P1	INT4	0.97106	0.98838	1.01269	1.03953	1.07276	1.09440
P2	INT4	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
P3	INT4	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
P4	INT4	0.95799	0.98564	1.01027	1.03682	1.07869	1.08282

C.3 Step-Size Bins

Table 7 splits every class II cell into the bins $[0, 1)$, $[1, 1.8)$ and $[1.8, 2)$ of β, which hold 2681, 3061 and 402 of the 6144 steps. Every overshooting step lies in the top bin. With INT4 keys, 209 of its 402 steps overshoot on P1 and 199 on P4; with FP8 keys, 35 and 40; with BF16 keys, 4 and 5 (Figure 6b); on P2 and P3, none. With INT4 keys the mean of s changes by at most 0.003 from bin to bin, small against a spread of about 0.02 and consistent with the independence assumption of Theorem 3. With FP8 keys on P1 the top bin has a lower mean of s, 0.992 against 0.999 and 1.002 in the other bins (Figure 6c). In class I all 6144 step sizes lie in the first bin, and none of them overshoots (Table 8).

Table 7: Class II mean of s and steps with $\beta s > 2$ in each bin of the step size β, per placement, with INT4 and FP8 keys; the three bins hold 2681, 3061 and 402 of the 6144 steps.

		$\beta \in [0, 1)$		$\beta \in [1, 1.8)$		$\beta \in [1.8, 2)$	
Placement	Key	Mean s	$\beta s{>}2$	Mean s	$\beta s{>}2$	Mean s	$\beta s{>}2$
P1	INT4	1.01510	0	1.01603	0	1.01307	209
P1	FP8	0.99941	0	1.00165	0	0.99204	35
P2	INT4	1.00000	0	1.00000	0	1.00000	0
P2	FP8	1.00000	0	1.00000	0	1.00000	0
P3	INT4	1.00000	0	1.00000	0	1.00000	0
P3	FP8	1.00000	0	1.00000	0	1.00000	0
P4	INT4	1.01501	0	1.01783	0	1.01662	199
P4	FP8	0.99992	0	1.00334	0	1.00096	40

Table 8: Class I: all 6144 step sizes lie in $[0, 1)$ and no step overshoots; mean of s per placement and key format.

Placement	Key	Mean s
P1	INT4	1.01065
P1	FP8	1.00001
P1	BF16	0.99999
P2	INT4	1.00000
P2	FP8	1.00000
P2	BF16	1.00000
P3	INT4	1.00000
P3	FP8	1.00000
P3	BF16	1.00000
P4	INT4	1.01251
P4	FP8	1.00479
P4	BF16	0.99992

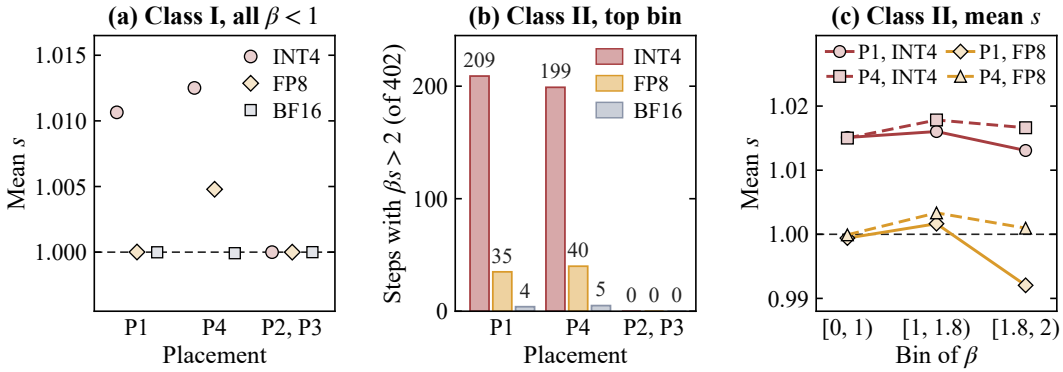

Figure 6: Step-size bins. (a) Mean of s over the class I steps per placement and key format. (b) Class II steps with $\beta s > 2$ among the 402 steps with $\beta \in [1.8, 2)$. (c) Mean of s in each bin of β for class II on P1 and P4 with INT4 and FP8 keys.

C.4 Statistics per Model and Placement

Table 9 gives the mean step size and the overshoot share of the six models on P1 and P4. The mean step size is 0.565 and 0.575 in class I and 1.066 to 1.092 in class II (Figure 7a). Overshooting steps need both a class II model and one of these two placements: 3.1% to 3.5% of the steps overshoot with INT4 keys and 0.5% to 0.7% with FP8 keys, and on P2 and P3 none of them does (Figure 7b).

Table 9: Mean step size $\bar{\beta}$ and share of steps with $\beta s > 2$ (%) on P1 and P4 with INT4 and FP8 keys, per model. On P2 and P3 no step overshoots.

		P1		P4	
Model	$\bar{\beta}$	INT4 key	FP8 key	INT4 key	FP8 key
DeltaNet (I)	0.575	0.0	0.0	0.0	0.0
Gated DeltaNet (I)	0.565	0.0	0.0	0.0	0.0
Gated DeltaNet, $\beta \in (0, 2)$ (II)	1.086	3.4	0.6	3.2	0.7
DeltaProduct n_h=1 (II)	1.069	3.4	0.7	3.1	0.5
DeltaProduct n_h=2 (II)	1.092	3.5	0.5	3.3	0.6
DeltaProduct n_h=3 (II)	1.066	3.4	0.6	3.2	0.5

Table 10: Share of steps with $\beta s > 2$ (%) in the default class II model per placement and key format, over all heads and in the worst single head, as drawn in Figure 7b.

Placement	Key	All heads	Worst head
P1	INT4	3.4	5.9
P1	FP8	0.6	1.6
P4	INT4	3.2	4.7
P4	FP8	0.7	1.2
P2	INT4	0.0	0.0
P2	FP8	0.0	0.0
P3	INT4	0.0	0.0
P3	FP8	0.0	0.0

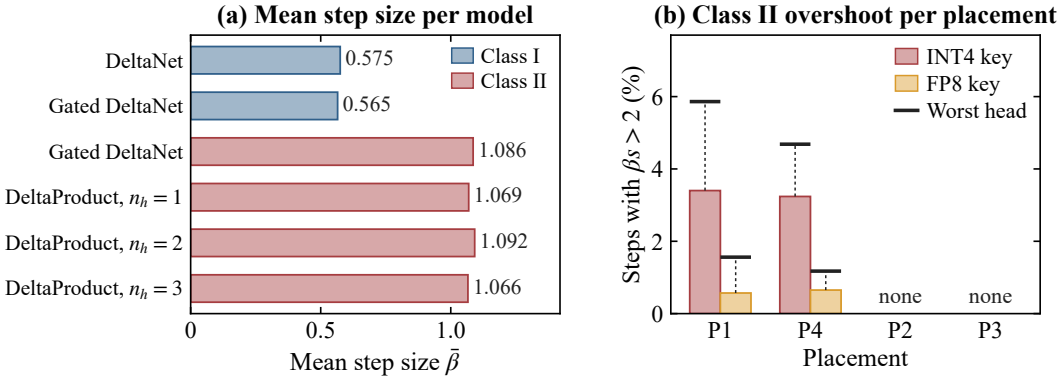

Figure 7: Statistics per model and placement. (a) Mean step size $\bar{\beta}$ of each model over the recorded steps, as in Table 9. (b) Share of steps with $\beta s > 2$ in the default class II model per placement and key format, with the worst single head marked; Table 10 lists the values.

C.5 Effective Step Size

Table 11 gives the quantiles of βs with INT4 keys for the three main models, and Figure 1a draws them, averaging the two class I models. In class I the maximum stays below 1.03. In class II the maximum is 2.131 and the 99th percentile is 2.034, so the top 1% of the steps all overshoot. Since the shift $\beta s - \beta = \beta(s - 1)$ is proportional to β, the largest step sizes receive the largest shifts.

Table 11: Quantiles of the effective step size βs and share of steps with $\beta s > 2$ (%) with INT4 keys on P1, per model.

Key	Model	1st	10th	25th	50th	75th	90th	99th	99.9th	Max	βs>2
INT4	DeltaNet (I)	0.1022	0.2674	0.4194	0.6088	0.7624	0.8720	0.9778	1.0012	1.0194	0.00
INT4	Gated DeltaNet (I)	0.1067	0.2565	0.4118	0.5859	0.7784	0.8567	0.9878	1.0001	1.0216	0.00
INT4	Gated DeltaNet, $\beta \in (0, 2)$ (II)	0.1356	0.4195	0.7121	1.1281	1.4181	1.7036	2.0342	2.0584	2.1310	3.40

D CONTROLLED EXPERIMENTS

These experiments change one quantity at a time, on the unquantized default class II model or on the class I models, and compare each result with the prediction that Section 3 makes for that quantity.

D.1 RADIAL AND TANGENTIAL SHIFTS

We add to every unit key a perturbation of norm a, either along the key (radial) or orthogonal to it (tangential), and record the shift Δs of the squared norm, the share of steps with $\beta s > 2$ and, for the radial case, the mean residual ratio $|1 - \beta s|$ (Table 12). Under a radial perturbation Δs is close to $2a + a^2$, and under a tangential one close to a^2, since a tangential error changes s only at second order. At $a = 0.045$ the radial perturbation makes 6.5% of the steps overshoot and the tangential one 0.8%, so the rate follows the change of s rather than the size of the key error, and for each kind the rate increases with Δs, as the nondecreasing tail function h of Theorem 3 requires (Figure 2a).

Table 12: Shifts of equal norm a along the unit key (radial) and orthogonal to it (tangential): measured Δs with the values $2a + a^2$ and a^2 for comparison, share of steps with $\beta s > 2$ (%), and the mean residual ratio $|1 - \beta s|$ of the radial case.

	Radial				Tangential		
a	Δs	$2a + a^2$	$\beta s{>}2$	Ratio	Δs	a^2	$\beta s{>}2$
0.004	0.00798	0.00802	2.7	0.403	0.00003	0.00002	0.0
0.009	0.01816	0.01808	4.5	0.411	0.00007	0.00008	0.1
0.017	0.03414	0.03429	6.0	0.420	0.00030	0.00029	0.2
0.028	0.05673	0.05678	6.3	0.427	0.00077	0.00078	0.4
0.045	0.09174	0.09202	6.5	0.452	0.00203	0.00202	0.8

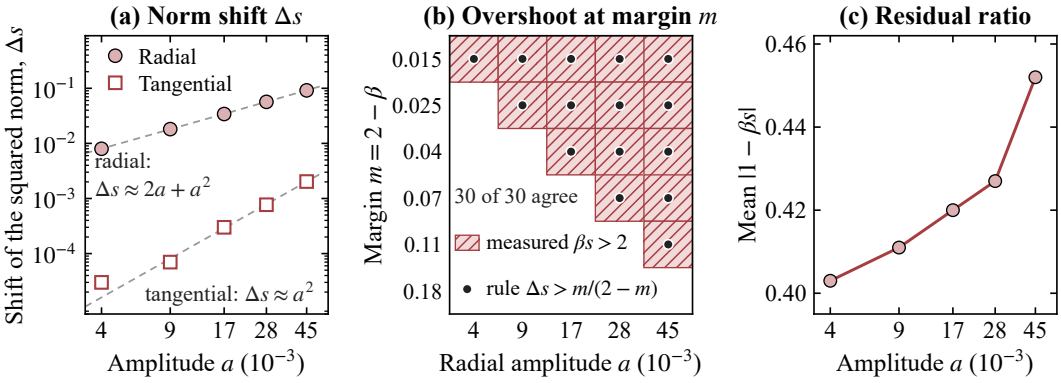

Figure 8: Shifts of the key norm. (a) Shift Δs of the squared norm under radial and tangential shifts of equal amplitude a, on logarithmic axes; the dashed lines are $2a + a^2$ and a^2. (b) Indicator of $\beta s > 2$ for a step size fixed at margin $m = 2 - \beta$ under each radial amplitude; hatched cells overshoot, and dots mark the cells where the rule $\Delta s > m/(2 - m)$ predicts an overshoot. (c) Mean residual ratio $|1 - \beta s|$ under the radial shifts.

D.2 TEST OF THE MARGIN CONDITION

Proposition 2(iii) is exact, so this grid only checks the kernel. We fix the step size at a margin m below 2, apply each of the five radial shifts and record whether the step overshoots (Table 13). With the radial Δs of Table 12, the rule $\Delta s > m/(2 - m)$ predicts the outcome of all 30 cells (Figure 8b).

Table 13: Indicator of $\beta s > 2$ for a step size fixed at margin $m = 2 - \beta$ under radial shifts of amplitude a, with the threshold $m/(2-m)$ of Proposition 2(iii).

		Radial amplitude a				
m	$m/(2-m)$	0.004	0.009	0.017	0.028	0.045
0.015	0.00756	1	1	1	1	1
0.025	0.01266	0	1	1	1	1
0.040	0.02041	0	0	1	1	1
0.070	0.03627	0	0	0	1	1
0.110	0.05820	0	0	0	0	1
0.180	0.09890	0	0	0	0	0

D.3 Step-Size Error in Class I

In class I no step overshoots, so this test isolates the first regime. We scale the key so that $|s-1|$ takes five values and measure recall and perplexity (Table 14). In all three models, recall falls and perplexity rises monotonically as the error grows. From $|s-1| = 0.005$ to 0.05, recall falls by 2.81, 4.35 and 3.29 points and perplexity rises by 0.59, 0.53 and 1.49 for DeltaNet, Gated DeltaNet and the 1.3B DeltaNet, respectively, with no overshooting step at any one of the five values of $|s-1|$.

Table 14: Class I response to a fixed step-size error $|s-1|$: associative recall (Rec, %) and perplexity (PPL) of DeltaNet, Gated DeltaNet and a 1.3B DeltaNet.

	DeltaNet		Gated DeltaNet		DeltaNet 1.3B	
$\lvert s-1\rvert$	Rec	PPL	Rec	PPL	Rec	PPL
0.005	90.27	22.52	90.22	22.30	90.19	22.17
0.012	89.81	22.61	89.22	22.38	88.52	22.36
0.021	89.69	22.75	87.19	22.49	87.80	22.82
0.034	88.45	22.92	86.55	22.64	87.29	22.96
0.050	87.46	23.11	85.87	22.83	86.90	23.66

D.4 Predicted Overshoot Rate

Table 15 compares the rate that Theorem 3 predicts from the measured h and the recorded distribution of s with the measured rate. The two inputs come from different measurements: h from shifts of unquantized keys, which do not involve the quantizer, and s from the quantized record, which does not involve the step size. With INT4 keys the prediction is within 7% on both placements. With FP8 keys the rates are five to six times smaller and rest on the few steps with β within about 0.03 of 2, where h is sampled by the tangential shifts and the two smallest radial ones; the prediction is within 5% on P4 and 23% on P1. With FP8 keys on P1 the top bin of β has a lower mean of s, 0.992 against 0.999 and 1.002 in the other bins (Appendix C.3), so independence holds less well for that cell.

Table 15: Share of class II steps with $\beta s > 2$ (%) predicted by Theorem 3 and measured, and the relative difference of the prediction.

Placement	Key	Predicted	Measured	Relative difference
P1	INT4	3.17	3.40	-6.8%
P4	INT4	3.09	3.24	-4.6%
P1	FP8	0.70	0.57	$+22.9\%$
P4	FP8	0.68	0.65	$+4.5\%$

D.5 Overshoot Share and Accuracy Loss across Sequences

We group 768 unrepaired class II sequences into six bins by their share of overshooting steps and compare perplexity and state tracking with those of the lowest bin (Table 16). Both degrade steadily with the share: perplexity rises by up to 1.01 and state tracking falls by up to 7.8 points (Figure 9). The comparison is correlational; Appendices D.1 to D.3 test the mechanism by direct interventions.

Table 16: Unrepaired class II sequences binned by their share of steps with $\beta s > 2$: lower bound of the bin (%), change of perplexity and of state-tracking accuracy (points) relative to the first bin, and number of sequences.

Bin	Lower bound (%)	ΔPPL	ΔState	Sequences
1	0.3	0.000	0.00	119
2	1.0	0.376	−2.92	126
3	1.7	0.533	−4.07	132
4	2.4	0.688	−5.31	124
5	3.1	0.869	−6.52	137
6	3.8	1.008	−7.79	130

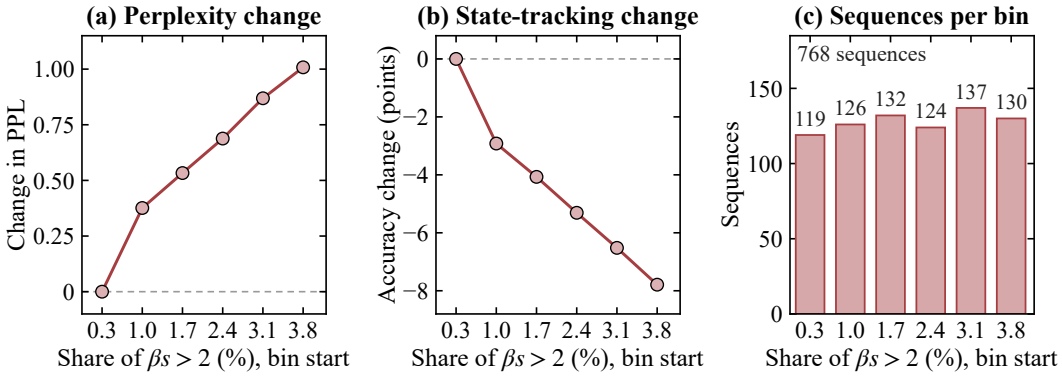

Figure 9: Unrepaired class II sequences binned by their share of steps with $\beta s > 2$. (a) Change of perplexity relative to the first bin. (b) Change of state-tracking accuracy relative to the first bin. (c) Number of sequences per bin.

E Full Results

E.1 Perplexity per Corpus

The PPL columns of Table 1 are geometric means over PG-19 and Books3, and Table 17 lists the two corpora separately. NormFix has lower perplexity than the unrepaired model on both corpora at every length, and at 128k tokens it is the lowest post-training row on Books3, and second lowest on PG-19.

Table 17: Perplexity per corpus at 8k, 32k and 128k tokens, default setting; shading as in Table 1.

	8k		32k		128k	
Method	PG-19	Books3	PG-19	Books3	PG-19	Books3
BF16 reference	14.65	15.80	17.73	19.53	21.01	21.78
No fix	15.95	17.86	20.09	21.39	25.35	26.96
Clip β at 1.9	15.80	17.61	19.99	22.32	25.35	26.75
Clip β at 1.0	16.03	17.77	20.30	22.14	25.40	28.17
Skip if $\beta s > 2$	15.33	17.00	19.97	21.80	24.16	26.73
Renormalize key	15.63	17.32	19.50	20.86	23.70	26.18
W6 weights	15.75	17.20	19.46	21.35	23.43	26.60
INT8 key	15.44	16.93	18.09	20.86	22.00	25.43
FP8 key	15.13	16.90	19.97	22.28	24.49	25.59
QuaRot + W4	15.21	17.10	19.51	21.81	24.22	27.10
SpinQuant + W4	15.99	17.77	20.33	22.16	24.16	27.33
GPTQ / AWQ W4	15.45	17.54	20.00	21.21	24.72	25.83
QAT-W4 (retrained)	15.28	16.51	18.33	19.81	22.64	24.31
NormFix	15.21	17.22	19.33	20.60	22.96	25.33

E.2 Paired Differences

Table 18 lists the paired differences on the state-task mean at 64k between NormFix and the rows that the main text compares it with. In both classes the interval against each of these rows lies above zero. Class I has no row for skipping: with no overshooting step, skipping would change nothing there.

Table 18: Paired differences on the mean of parity, S_5 and Dyck at 64k tokens (points), with 95% intervals over 4096 sequences.

Class	Comparison	Difference	Lower	Upper
II	NormFix − no fix	17.74	16.32	19.15
II	NormFix − skip	8.04	6.93	9.16
II	NormFix − renormalize	3.91	3.26	4.56
II	NormFix − FP8 key	3.37	2.71	4.03
II	NormFix − SpinQuant	10.01	9.04	10.98
I	NormFix − no fix	3.47	2.77	4.18
I	NormFix − renormalize	2.15	1.67	2.64
I	NormFix − FP8 key	1.03	0.55	1.51
I	NormFix − SpinQuant	1.50	0.94	2.06

E.3 Context Length

Table 19 gives the state-tracking accuracy behind Figure 3a. NormFix is the best of the four methods in all 15 cells, with a gain of 14.8 to 18.2 points over the unrepaired model at each length. Table 20 gives the perplexity of the same repairs at four lengths in both classes, and in class II NormFix has the lowest perplexity at 32k, 64k and 128k tokens, while skipping has the lowest at 8k (Figure 10c).

Table 19: State-tracking accuracy (%) of the unrepaired model and three repairs at five lengths, default setting; the best of the four is in bold.

Task	Length	No fix	Renormalize	Skip	NormFix
Parity	4k	71.5	82.1	82.0	**86.4**
Parity	16k	67.1	80.8	77.5	**84.6**
Parity	32k	66.1	81.3	75.7	**83.3**
Parity	64k	65.4	79.2	76.1	**83.3**
Parity	128k	64.7	79.5	73.2	**80.8**
S_5	4k	62.5	77.2	72.9	**79.5**
S_5	16k	61.8	75.8	71.1	**78.3**
S_5	32k	60.9	74.7	69.5	**77.2**
S_5	64k	59.0	73.7	68.3	**76.2**
S_5	128k	57.8	72.9	67.5	**75.3**
Dyck	4k	67.3	78.7	76.3	**82.1**
Dyck	16k	65.7	77.5	73.2	**80.8**
Dyck	32k	65.6	74.7	72.7	**80.8**
Dyck	64k	62.9	75.9	72.0	**81.1**
Dyck	128k	62.3	76.2	71.4	**79.6**

Table 20: Perplexity of the unrepaired model and three repairs at four lengths for the two class I models and the default class II model; skipping is shown for class II only.

Model	Length	No fix	Renormalize	Skip	NormFix
DeltaNet (I)	8k	15.37	16.18		15.44
DeltaNet (I)	32k	19.97	19.43		18.99
DeltaNet (I)	64k	22.19	21.15		21.29
DeltaNet (I)	128k	24.12	23.34		23.35
Gated DeltaNet (I)	8k	16.49	16.48		15.58
Gated DeltaNet (I)	32k	19.49	19.86		19.42
Gated DeltaNet (I)	64k	21.69	21.29		21.00
Gated DeltaNet (I)	128k	24.15	23.68		23.42
Gated DeltaNet, $\beta \in (0, 2)$ (II)	8k	16.88	16.45	16.14	16.18
Gated DeltaNet, $\beta \in (0, 2)$ (II)	32k	20.73	20.17	20.86	19.95
Gated DeltaNet, $\beta \in (0, 2)$ (II)	64k	24.68	22.90	23.34	21.96
Gated DeltaNet, $\beta \in (0, 2)$ (II)	128k	26.14	24.91	25.41	24.12

E.4 Step Statistics of All Configurations

Figure 10 draws the step-size columns of Table 1. Only NormFix and renormalization remove both the overshoots and the step-size error. Clipping at 1.0 and skipping remove the overshoots but change the step a lot, and the 8-bit keys shrink both quantities without removing either of them completely.

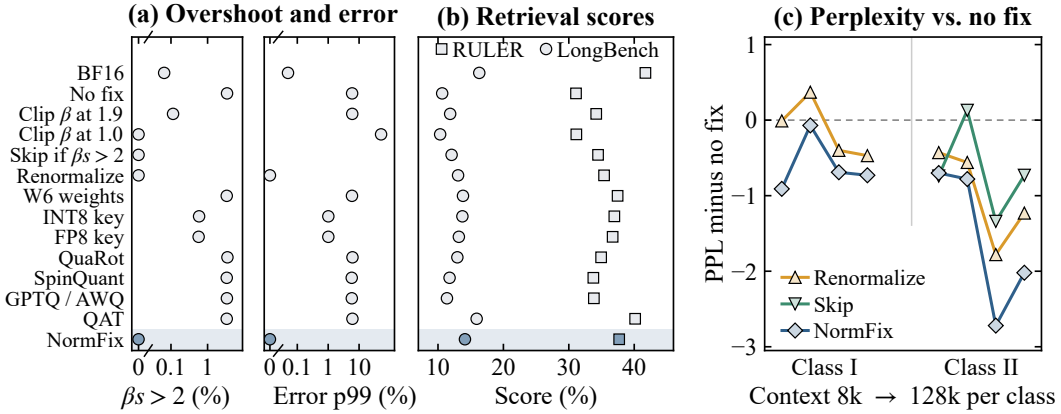

Figure 10: Step-size and task columns of the class II rows of Table 1, with NormFix shaded. (a) Share of steps with $\beta s > 2$ and 99th percentile of the step-size error (%; no value for skipping); zero is drawn at the left edge of each logarithmic axis. (b) RULER and LongBench. (c) Perplexity of renormalization, skipping and NormFix minus that of the unrepaired model from 8k to 128k tokens (Table 20), for Gated DeltaNet in class I and the default model in class II; skipping is drawn for class II only.

E.5 State Error

Table 21 gives the state error behind Figure 1c. From 1k to 128k tokens it grows by a factor of 11 to 13 in both classes, which corresponds to $L^{0.51}$, and class II carries 2.4 to 2.8 times the error of class I. Table 22 lists the 40 sequences of Figure 4a and b: NormFix lowers the state error of every sequence, and the ratio of the error after the repair to the error before it lies between 0.227 and 0.486. Table 23 gives the changes of Figure 4c. In class I on P1, NormFix lowers perplexity by 0.73 to 0.77 and raises RULER and state tracking by 1.9 to 3.5 points, while on P2 and P3, where $s = 1$, all changes stay below 0.0002 in perplexity and 0.01 points in accuracy, two orders of magnitude below those on P1.

Table 21: State error left by a radial perturbation of every key, for four perturbation sizes (1 is the smallest) and six context lengths, without repair.

Class	Size	1k	4k	16k	32k	64k	128k
I	1	0.00138	0.00290	0.00588	0.00760	0.01211	0.01695
I	2	0.00282	0.00561	0.01166	0.01719	0.02342	0.03351
I	3	0.00442	0.00874	0.01715	0.02391	0.03576	0.05066
I	4	0.00567	0.01136	0.02213	0.03333	0.04572	0.06970
II	1	0.00368	0.00714	0.01636	0.02165	0.02973	0.04292
II	2	0.00697	0.01530	0.03096	0.04165	0.06359	0.09011
II	3	0.01108	0.02228	0.04518	0.06435	0.09614	0.13150
II	4	0.01502	0.03082	0.05785	0.08774	0.12543	0.17722

Table 22: State error of the 40 class II sequences of Figure 4 without repair (before) and with NormFix (after), and their ratio.

#	Before	After	Ratio	#	Before	After	Ratio
1	0.1686	0.0513	0.304	21	0.1414	0.0473	0.334
2	0.1795	0.0701	0.391	22	0.1717	0.0810	0.472
3	0.2086	0.0643	0.308	23	0.1003	0.0323	0.322
4	0.0749	0.0170	0.227	24	0.0937	0.0321	0.342
5	0.1733	0.0407	0.235	25	0.1620	0.0373	0.230
6	0.1510	0.0572	0.379	26	0.1383	0.0415	0.300
7	0.0487	0.0204	0.420	27	0.0364	0.0123	0.339
8	0.2109	0.0623	0.295	28	0.1322	0.0373	0.282
9	0.1983	0.0829	0.418	29	0.0418	0.0190	0.453
10	0.0407	0.0152	0.372	30	0.1260	0.0601	0.477
11	0.1113	0.0290	0.261	31	0.0577	0.0150	0.260
12	0.0331	0.0150	0.454	32	0.1872	0.0433	0.231
13	0.1413	0.0368	0.260	33	0.1760	0.0768	0.436
14	0.0750	0.0176	0.235	34	0.1067	0.0348	0.327
15	0.0528	0.0173	0.328	35	0.2059	0.0578	0.281
16	0.1189	0.0527	0.443	36	0.2079	0.0678	0.326
17	0.2124	0.1032	0.486	37	0.2117	0.0891	0.421
18	0.1166	0.0451	0.387	38	0.1442	0.0622	0.431
19	0.1157	0.0510	0.441	39	0.0515	0.0187	0.362
20	0.0606	0.0209	0.345	40	0.1992	0.0815	0.409

Table 23: Change caused by NormFix: decrease of perplexity at 128k tokens and gain in RULER and state tracking (points), in class I on P1, where $s \neq 1$, and on P2 and P3, where $s = 1$.

Setting	PPL decrease	RULER gain	State gain
DeltaNet (I), P1	0.77	1.86	3.30
Gated DeltaNet (I), P1	0.73	3.01	3.47
P2	0.00018	0.008	0.003
P3	−0.00004	0.007	−0.009

F Ablation Studies

This appendix gives the full sweeps behind Table 2, all in the default setting. Table 27 repeats every ablation with its perplexity at 128k tokens. The sweeps below also report the share of steps with $\beta s > 2$ in percent, the mean absolute step-size error and the mean single-step residual ratio $|1 - \beta s|$.

F.1 Clipping the Step Size

A cap on β removes overshoots only by shrinking the step (Table 24). A cap at 1.9 removes 97% of the overshooting steps and gains 6.6 points of parity over no clipping. A cap at 1.5 removes all of them but gains only 3.4 points, and a cap at 1.0 ends below no clipping on both tasks: it multiplies the mean step-size error by 11.6 and lowers the mean residual ratio from 0.425 to 0.169, as the

reflections turn into near-complete erasures (Figure 11). Clipping helps only while it keeps most of the reflection, and every cap in this sweep stays well below the full gain of rescaling the step size.

Table 24: Sweep of the cap on the step size, default setting: share of steps with $\beta s > 2$ (%), mean absolute step-size error, mean residual ratio $|1 - \beta s|$, parity and S_5 at 64k tokens (%) and perplexity at 128k tokens.

Setting	$\beta s{>}2$	Step-size error	Residual ratio	Parity	S_5	PPL
Cap at 1.0	0.00	0.232	0.169	63.72	56.44	26.75
Cap at 1.5	0.00	0.058	0.361	68.80	62.08	26.45
Cap at 1.9	0.11	0.022	0.400	71.96	63.85	26.04
No clipping (cap at 2)	3.40	0.020	0.425	65.40	58.99	26.14

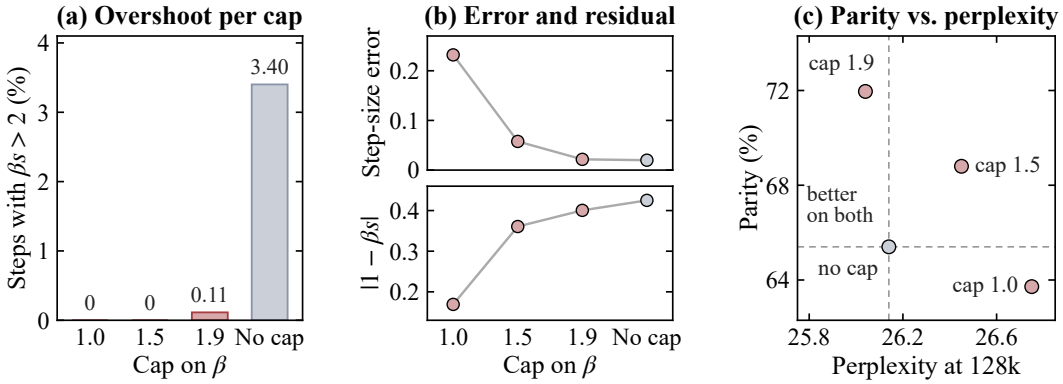

Figure 11: Clipping the step size. (a) Share of steps with $\beta s > 2$ under each cap. (b) Mean absolute step-size error and mean residual ratio $|1 - \beta s|$. (c) Parity against perplexity for each cap, with lines through the unclipped model.

F.2 Applying NormFix above a Threshold

Applying NormFix only to steps whose effective step size exceeds a threshold removes every overshoot at any threshold, since a step that would overshoot lies above it (Table 25). The threshold only changes the error of the steps it leaves alone. The loosest threshold, $\beta s > 2$, leaves a mean absolute step-size error of 0.019 and gives up 1.94 points of parity and 1.88 of S_5 (Figure 12). A threshold still needs s at every step to decide which steps to repair, so NormFix repairs every step.

Table 25: NormFix applied only to steps with βs above a threshold, default setting; columns as in Table 24.

Setting	$\beta s{>}2$	Step-size error	Residual ratio	Parity	S_5	PPL
Every step	0.00	0.000	0.401	83.29	76.16	24.12
Only $\beta s > 1$	0.00	0.005	0.400	82.72	75.67	24.55
Only $\beta s > 1.9$	0.00	0.018	0.404	82.00	74.39	24.90
Only $\beta s > 2$	0.00	0.019	0.405	81.35	74.28	25.28

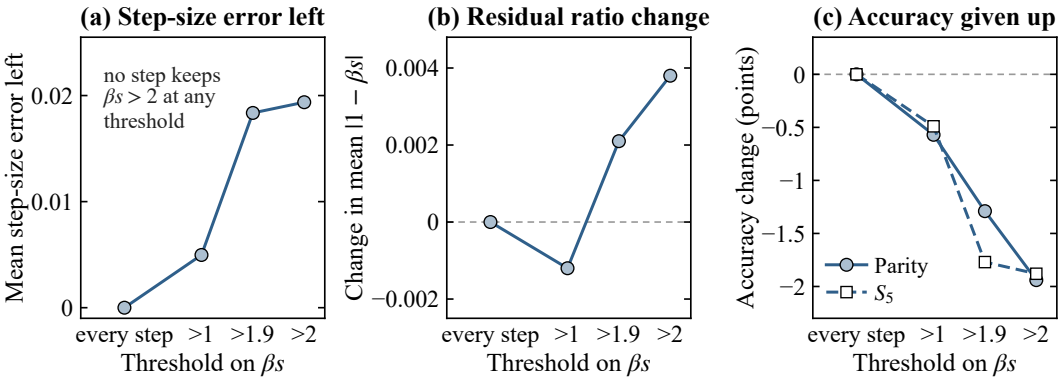

Figure 12: Threshold on βs. (a) Mean absolute step-size error left on the untreated steps; no step exceeds 2 at any threshold. (b) Change of the mean residual ratio against the repair on every step. (c) Parity and S_5 accuracy given up against the repair on every step.

F.3 Repaired Layers

Table 26 repairs the first n of the 32 layers. The share of overshooting steps falls as more layers are repaired and reaches zero only when all 32 are. The parity gain rises steadily: 4, 8, 16 and 24 layers reach 40%, 64%, 80% and 92% of the final parity gain (Figure 13). Early layers contribute more gain per repaired layer, and every later block still adds to it, so we apply NormFix to all of the 32 layers.

Table 26: Repairing the first n of 32 layers, default setting; columns as in Table 24.

Layers	$\beta s>2$	Step-size error	Residual ratio	Parity	S_5	PPL
0 of 32	3.4	0.020	0.425	65.4	59.0	26.14
4 of 32	2.9	0.017	0.412	72.6	64.2	25.78
8 of 32	2.5	0.014	0.410	76.9	68.1	25.20
16 of 32	1.9	0.011	0.401	79.8	71.7	24.72
24 of 32	0.6	0.004	0.382	81.8	74.0	24.36
32 of 32	0.0	0.000	0.401	83.3	76.2	24.12

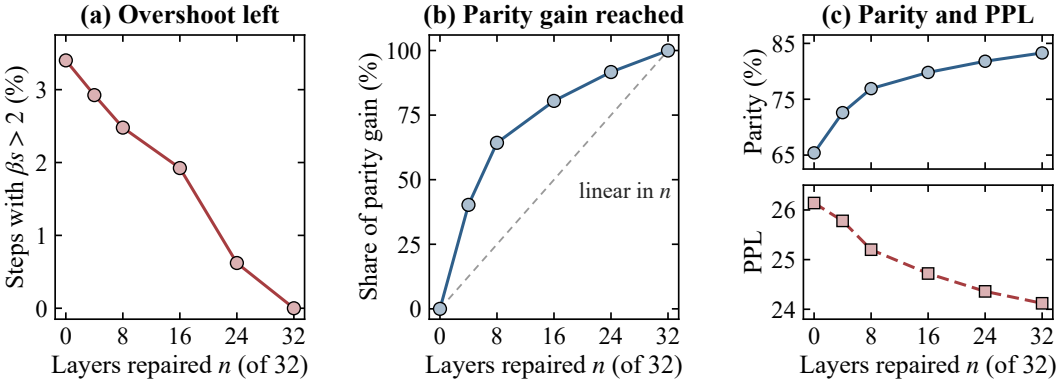

Figure 13: Repairing the first n of 32 layers. (a) Share of steps with $\beta s > 2$ that remains. (b) Share of the final parity gain reached; the dashed line is a gain that grows linearly with n. (c) Parity and perplexity along the sweep.

F.4 Exempting Heads from the Repair

Figure 14 follows the mean squared state error over a 64-step test sequence that starts from an outdated state, with and without the repair. The transient ends within four steps, after which both kernels reach the same error level. Over the test sequence, the ratio of repaired to unrepaired squared error lies between 0.988 and 1.073 across the blocks of class, placement and key format, and it

rounds to one on P2 and P3, where $s = 1$. Over such a short sequence the effect of the repair has no consistent sign, and in some blocks the repair raises the error. Row 13 of Table 2 leaves unrepaired the heads on which the unrepaired kernel had the lower error on this test, and loses 6% of the gain.

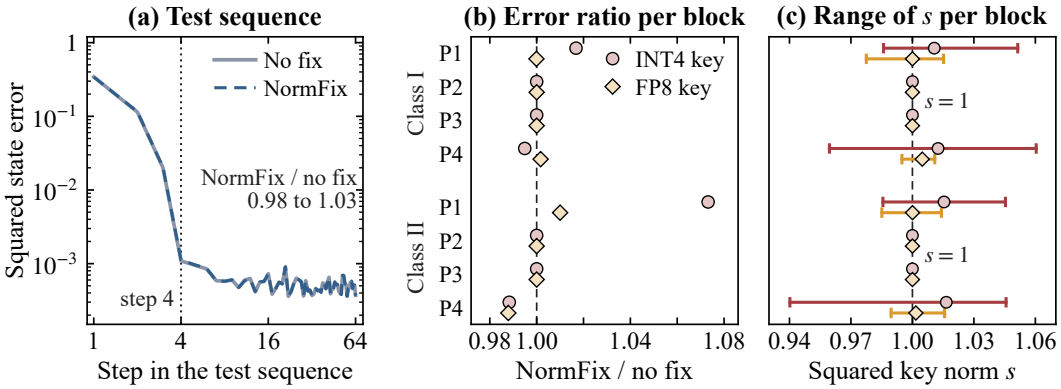

Figure 14: Test sequences that start from an outdated state. (a) Mean squared state error at each of the 64 steps without and with NormFix; the dotted line marks step 4, and the ratio of the two errors stays between 0.98 and 1.03 at every step. (b) Ratio of the squared error with NormFix to that without over the test sequence, per block of class and placement, for INT4 and FP8 keys. (c) Range and mean of s in each block.

Table 27: All ablations of Table 2 with perplexity at 128k tokens, default setting; rows and shading as in Table 2. Kept: share of the NormFix gain on the mean of parity and S_5 (%).

#	Variant	Parity	S_5	PPL	Kept
0	**NormFix**	83.3	76.2	24.12	100.0
1	Skip	76.1	68.3	25.41	57.0
2	Clip at 1.9	72.0	63.9	26.04	32.6
3	Renormalize	79.2	73.7	24.91	81.3
4	Only $\beta s > 2$	81.3	74.3	25.28	89.1
5	Only $\beta s > 1.9$	82.0	74.4	24.90	91.3
6	Only $\beta s > 1$	82.7	75.7	24.55	97.0
7	First 8 layers	76.9	68.1	25.20	58.8
8	First 16 layers	79.8	71.7	24.72	77.3
9	First 24 layers	81.8	74.0	24.36	89.6
10	s before quantization	65.4	59.0	26.14	0.0
11	Per-layer mean s	82.1	75.0	24.28	93.4
12	16-bit accumulation	82.2	75.0	24.08	93.8
13	Head exemption	82.0	75.4	24.31	94.4
	No fix	65.4	59.0	26.14	0.0

G Results on Other Models and Settings

Table 28 lists the results with NormFix on the model, scale and recipe axes. On the context axis, perplexity, RULER and the state tasks are measured at the row's own length; on the other axes, perplexity is at 128k tokens, RULER averages all lengths and the state tasks are at 64k tokens. Shaded rows mark the default setting and, except on the context axis, repeat the NormFix row of Table 1.

Table 29 gives the gain of NormFix over the unrepaired model per model and task. Including RULER, the two class I models gain 1.9 to 4.4 points and the four class II models 5.8 to 18.3 points, and on every task the 95% intervals of the two class means are disjoint (Figure 3b). The range of β decides which regime a model can reach, and the gain of NormFix follows the same split in all six models.

Table 31 gives the three scale axes: the share of overshooting steps before the repair stays between 3.3% and 3.5%, and the gain on the state tasks stays between 15.6 and 18.2 points. The context length moves the gain the most, from 15.6 points at 4k tokens to 17.7 points at 64k tokens (Figure 15).

Table 28: Results with NormFix on the model, scale and recipe axes: perplexity at 128k tokens, RULER averaged over lengths and the mean of the three state tasks at 64k tokens; on the context axis all three are measured at the row's length. The default setting is shaded.

Axis	Setting	PPL	RULER	State
Model	DeltaNet (I)	23.35	44.44	93.50
	Gated DeltaNet (I)	23.42	43.10	92.30
	Gated DeltaNet, $\beta \in (0, 2)$ (II)	24.12	37.70	80.17
	DeltaProduct n_h=1 (II)	24.35	37.41	79.88
	DeltaProduct n_h=2 (II)	24.33	36.97	81.59
	DeltaProduct n_h=3 (II)	23.22	37.65	80.16
Context	4k	14.18	39.43	82.68
	16k	17.96	37.35	81.24
	64k	21.96	37.91	80.17
	128k	24.12	36.11	78.58
d	64	25.19	36.10	78.58
	128	24.12	37.70	80.17
	256	23.35	38.51	81.31
Layers	16	25.39	36.49	79.48
	24	23.69	35.71	81.79
	32	24.12	37.70	80.17
Recipe	W4 per-channel	23.21	35.13	78.00
	W4A8	24.42	38.50	82.42
	QuaRot + W4	24.18	38.13	80.86
	SpinQuant + W4	23.96	36.96	78.95

Table 29: Gain of NormFix over the unrepaired model per model and task (points). Table 30 gives the class means.

Model	Parity	S_5	Dyck	RULER
DeltaNet (I)	3.15	4.28	2.48	1.86
Gated DeltaNet (I)	2.77	3.21	4.44	3.01
Gated DeltaNet, $\beta \in (0, 2)$ (II)	17.89	17.17	18.15	6.60
DeltaProduct n_h=1 (II)	17.29	17.87	17.65	5.99
DeltaProduct n_h=2 (II)	17.45	18.27	17.37	6.33
DeltaProduct n_h=3 (II)	17.07	18.09	17.35	5.79

Table 30: Class means of the gain of NormFix per task (points) with their 95% intervals, drawn as bars in Figure 3b.

	Class I		Class II	
Task	Gain	95% interval	Gain	95% interval
Parity	2.96	[2.33, 3.60]	17.43	[16.27, 18.59]
S_5	3.75	[3.04, 4.46]	17.85	[16.56, 19.14]
Dyck	3.46	[2.86, 4.06]	17.63	[16.36, 18.90]
RULER	2.44	[1.87, 3.00]	6.18	[5.37, 6.99]

Table 31: Scale axes: share of steps with $\beta s > 2$ before the repair (%), mean state-tracking accuracy without repair and with NormFix (at 64k tokens, or at the row's length on the context axis), and the gain (points).

Axis	Setting	βs>2 before	No fix	NormFix	Gain
Context	4k	3.5	67.11	82.68	15.57
	16k	3.5	64.85	81.24	16.39
	64k	3.4	62.44	80.17	17.74
	128k	3.4	61.57	78.58	17.01
d	64	3.3	60.35	78.58	18.23
	128	3.4	62.44	80.17	17.74
	256	3.4	63.22	81.31	18.09
Layers	16	3.4	61.73	79.48	17.75
	24	3.4	63.92	81.79	17.87
	32	3.4	62.44	80.17	17.74

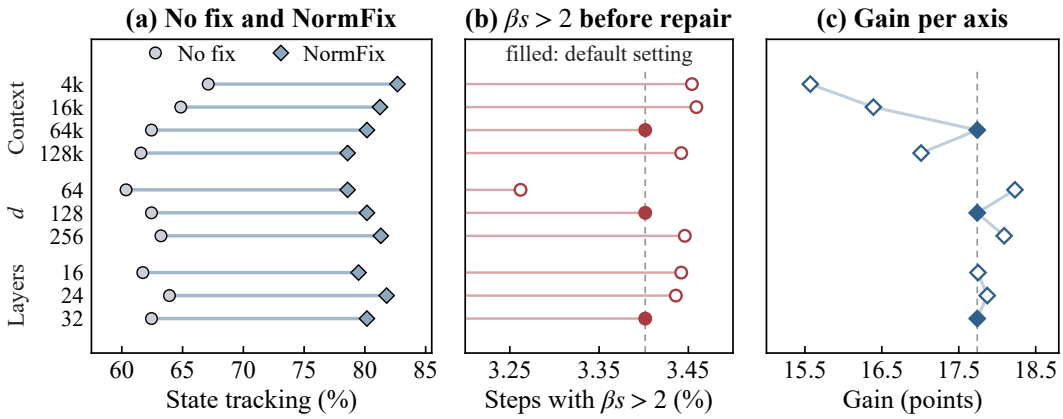

Figure 15: Scale axes. (a) Mean state-tracking accuracy without repair and with NormFix at each setting. (b) Share of steps with $\beta s > 2$ before the repair. (c) Gain along each axis. In (b) and (c), filled markers and dashed lines mark the default setting.

H KERNEL IMPLEMENTATION AND COST

Algorithm 1 shows one step of the fused delta-rule kernel of one head with NormFix. Lines 1 and 2 are new, and line 3 is the usual update with η_t in place of β_t. For DeltaProduct the three lines run once per factor with that factor's own key and step size. The renormalization kernel replaces line 2 by a scaling of the key, $\hat{\boldsymbol{k}}_t \leftarrow \hat{\boldsymbol{k}}_t/\sqrt{s_t}$, which costs a second pass over the key and a write-back.

Algorithm 1 NormFix inside a fused delta-rule kernel, one head, one step.

Require: state $\boldsymbol{S}$, decay α_t, step size β_t, received key $\hat{\boldsymbol{k}}_t$, value $\boldsymbol{v}_t$

1: $s_t \leftarrow \sum_{i=1}^{d} \hat{k}_{t,i}^2$ ▷ new, FP32 accumulation

2: $\eta_t \leftarrow \beta_t / s_t$ ▷ new

3: $\boldsymbol{r}_t \leftarrow \alpha_t \boldsymbol{S}\hat{\boldsymbol{k}}_t - \boldsymbol{v}_t$; $\quad \boldsymbol{S} \leftarrow \alpha_t \boldsymbol{S} - \eta_t \boldsymbol{r}_t \hat{\boldsymbol{k}}_t^\top$ ▷ as before, with η_t

Ensure: $\boldsymbol{S}$

Table 32 counts the extra operations of the two repairs per step and head at $d = 128$. NormFix adds the 128 multiplications and 127 additions of the reduction and two registers, and it moves no extra data. Renormalization adds 128 more multiplications, five more registers and 513 bytes of traffic. In the measured kernels, NormFix adds 0.27 μs per step and head in both Gated DeltaNet models, while renormalization adds 0.55 μs in class II and 0.63 μs in class I, 2.0 and 2.3 times as much (Figure 16).

Table 32: Extra multiplications, additions, registers and bytes of memory traffic per step and head at $d = 128$, and kernel time (μs) and decode throughput (tok/s) in the default setting.

Repair	Mul	Add	Regs	Bytes	μs	tok/s
No fix	0	0	0	0	7.46	742.9
Renormalize key	256	127	7	513	8.01	703.5
NormFix	128	127	2	0	7.73	722.0

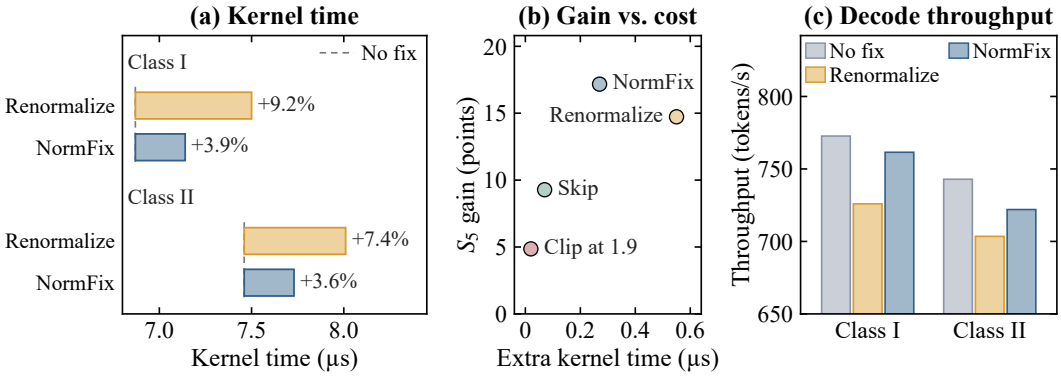

Figure 16: Cost for the Gated DeltaNet models of Table 1. (a) Kernel time per step and head of renormalization and NormFix in each class, with the overhead against the unrepaired kernel (dashed). (b) Gain on S_5 against the extra kernel time in the default setting. (c) Decode throughput without repair, with renormalization and with NormFix; the axis starts at 650 tokens/s.

I NOTATION

Table 33 lists the main symbols of the paper in the order of the sections where they are defined.

Table 33: Notation.

Symbol	Meaning	Defined in
L	context length in tokens	Figure 1
t	step (token) index	Section 3.1
$\boldsymbol{k}_t, \boldsymbol{v}_t, \boldsymbol{q}_t \in \mathbb{R}^d$	key (unit norm), value and query of step t; d is the head dimension	Section 3.1
β_t, σ	step size; $\sigma(\cdot) \in (0,1)$ in class I and $2\sigma(\cdot) \in (0,2)$ in class II, with σ the logistic sigmoid	Section 3.1
$\alpha_t \in (0,1]$	decay of the gated models	Section 3.1
$\boldsymbol{S}_t \in \mathbb{R}^{d\times d}$, $\boldsymbol{o}_t$, $\boldsymbol{I}$	state of one head, its output and the identity	Section 3.1
n_h	number of factors per token in DeltaProduct	Section 3.1
$\hat{\boldsymbol{k}}_t$	key the kernel receives	Section 3.1
$s_t = \|\hat{\boldsymbol{k}}_t\|_2^2$	squared norm of the received key	Section 3.1
$\beta_t s_t$	effective step size	Section 3.1
P1 to P4	placements of the quantizer relative to the normalizer	Section 3.1
$\boldsymbol{r}_t$, $\boldsymbol{r}_t^+$	residual before and after the update, read with $\hat{\boldsymbol{k}}_t$	Section 3.2
$\varepsilon_t = s_t - 1$	deviation of the squared norm from one	Section 3.2
$m_t = 2 - \beta_t$	margin of the step size below 2	Section 3.2
$\bar{G}$, h	tail of β_t and the function $h(\delta) = \bar{G}(2/(1+\delta))$	Section 3.3
f, $f(2^-)$, ℓ	density of β_t near 2, its left limit at 2 and its Lipschitz constant	Section 3.3
$s_{\max}$, R	upper bound of s_t and remainder of the first-order form	Section 3.3
δ	shift of the squared norm, or the level in the bound of Theorem 3	Section 3.3
a, Δs	amplitude of a radial or tangential shift of the key and the resulting shift of s	Section 3.3
$\eta_t = \beta_t / s_t$	step size used by NormFix	Section 4.1
$\boldsymbol{S}_t^\star$, $\boldsymbol{E}_t$	state of a kernel with the same α_t, β_t, $\boldsymbol{v}_t$ and the unquantized key $\boldsymbol{k}_t$, and error of the deployed state	Section 4.2
γ_t, $\boldsymbol{A}_t$, $\boldsymbol{F}_t$	step size used by a kernel, error transition matrix and local error	Section 4.2
$\boldsymbol{A}_t^{\mathrm{NF}}$, $\boldsymbol{F}_t^{\mathrm{NF}}$	transition matrix and local error under NormFix	Section 4.2
c_t, g_t, $\bar{\boldsymbol{k}}_t$	scales of the key and of the step size in a scalar repair, and the direction $\hat{\boldsymbol{k}}_t/\sqrt{s_t}$	Section 4.3
$\boldsymbol{\epsilon}$, ϵ_i	quantization error of the key, $\hat{\boldsymbol{k}} = \boldsymbol{k} + \boldsymbol{\epsilon}$, and its coordinates	Appendix A
μ, J	step size and loss of a least-mean-squares update	Appendix A.1
$w(\delta)$	length of the interval of step sizes that overshoot	Appendix A.4
$\boldsymbol{A}_t^{\mathrm{raw}}$, $\boldsymbol{F}_t^{\mathrm{raw}}$	transition matrix and local error of the unrepaired kernel	Appendix A.5
ρ	bound on the root-mean-square local error	Appendix A.5
τ, σ_i^2	grid step and error variance of coordinate i	Appendix A.7
$\bar{\varepsilon}$, σ_ε, Φ, φ	mean and standard deviation of $s-1$, standard normal distribution and density	Appendix A.7
$s_t^{(j)}$, $\eta_t^{(j)}$, $\bar{s}_t$	norm and step size of factor j, and a norm shared by all factors	Appendix A.8