STORED SAMPLING NOISE CHANGES RECALL LEARNING IN ADAMW

Anonymous authors
Paper under double-blind review

ABSTRACT

Adam and AdamW keep running averages of past gradients and of their squares, the first and second moments, so the sampling noise of one minibatch is not applied once but again and again over later updates. Common accounts describe minibatch noise by its size or covariance at each step. We ask whether, with both fixed, it matters how the stored noise lines up with the noise that each step applies. Associative recall lets us answer this. Because a context's queries can be enumerated, the gradient of one sampled query splits exactly into their mean and a residual, and we can choose what the first moment stores while every update still uses the sampled gradient. In a small transformer that plain AdamW leaves on a plateau, removing the residual from the first moment alone let recall form at default momentum. We then built pairs of stored copies with the same per-step covariance. A copy that mostly opposes the residual let 18 of 24 seeds learn recall, and its sign-flipped counterpart let 5 of the same 24, a gap that held in a model twice as deep and when the second moment ignored the residual's sign. A copy that cancels the residual over later updates adds noise at every step, yet with a short first-moment memory and matched early step sizes it led to recall sooner than storing only the mean. In a language model trained with loss on an artificially small subset of positions, flipping the stored residual's sign still lowered held-out loss, but storing only the mean did better at matched update size. What AdamW stores, and not only how much noise each step carries, can decide whether a small transformer learns in-context retrieval within the training budget we give it.

1 INTRODUCTION

AdamW (Loshchilov & Hutter, 2019), like Adam (Kingma & Ba, 2015), stores running averages of the gradient and of its elementwise square. A minibatch gradient therefore acts in its own update, again through the first moment m over roughly the next $1/(1-\beta_1)$ updates, and through the second moment v, which rescales every update. Its sampling residual, the part that depends on which examples were drawn, is reused in the same way. Analyses usually describe minibatch noise by a noise scale (McCandlish et al., 2018; Smith & Le, 2018; Smith et al., 2018) or a covariance (Zhu et al., 2019; HaoChen et al., 2021), and batch size and clipping change how much of it there is (McCandlish et al., 2018; Zhang et al., 2020). Noise-scale analyses of momentum (Smith & Le, 2018) and stochastic differential equation (SDE) analyses of Adam (Malladi et al., 2022) include the memory, but only with m storing each gradient exactly as drawn. Variance-reduced and slow first moments move the stored mean together with its noise (Yuan et al., 2025; Pagliardini et al., 2025), and positive-negative momentum rescales the stored noise but not its expectation (Xie et al., 2021).

For stochastic gradient descent, how noise is correlated across steps is known to matter. Anticorrelated noise and random reshuffling lower the error that builds up over many steps without making any single step quieter (Koloskova et al., 2023; Mishchenko et al., 2020). Error feedback, similarly, delays a compression error instead of discarding it (Seide et al., 2014; Karimireddy et al., 2019). A sum of noise terms grows with its long-run variance, the per-step variance plus the covariances at all lags. If the same holds for the copy that AdamW stores, a copy that cancels the residual over time should behave like anticorrelated noise (Orvieto et al., 2022; Lucchi et al., 2022) injected through m.

But AdamW's second moment rescales each coordinate by the sampled gradient's square. Runs that store different noise also diverge quickly, and noise sometimes helps training leave sharp minima (Zhu

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.

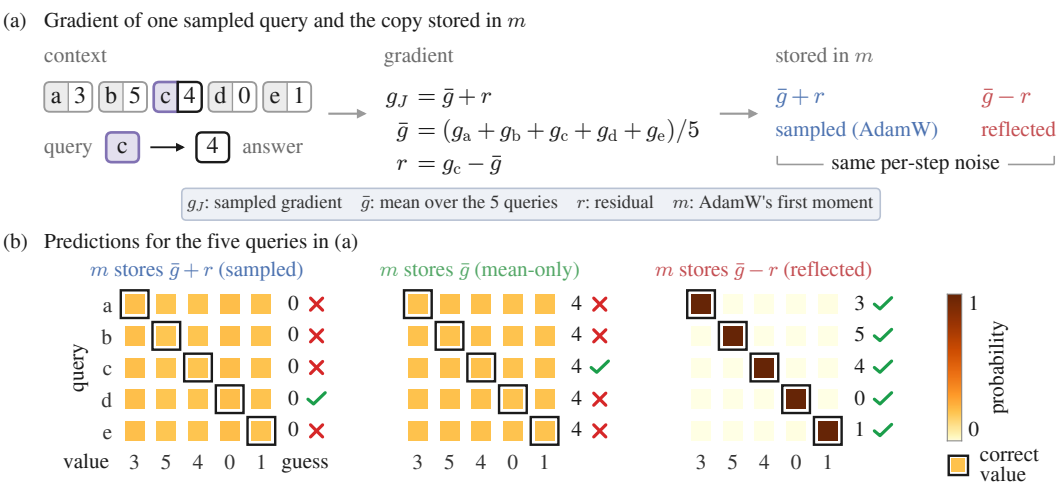


Fig. 1: AdamW's first moment and the sampling residual in associative recall. (a) A training context. The sampled query's gradient g_J is the mean $\bar{g}$ over the queries plus a residual r, and m stores $\bar{g} + r$, as in AdamW, or the reflected $\bar{g} - r$, with the same per-step covariance in the linear model. (b) Predictions for the five queries of (a) after 8,000 updates at $\beta_1 = 0.5$, in the seed whose reflected run acquired nearest the median time (outlined cells: correct values, checks: correct guesses).

et al., 2019). A copy that cancels the residual later could therefore slow learning rather than speed it. Our regime is far from smooth descent: recall in small transformers can stall on a plateau and then improve abruptly, and less gradient noise is known to shorten such plateaus (Kirsch et al., 2022; Sahasrabudhe, 2026). Instead of varying the amount of noise, as those studies do, we fix it and ask whether rearranging the stored noise over time moves the point where learning leaves the plateau. If only the amount of noise in each step mattered, such rearrangements would change nothing.

Answering this needs the exact mean of every sampled gradient, so that the stored noise can change at every update while its mean and covariance stay put. Earlier work follows one perturbed step, data reordered in a short window, or a transplanted first moment (Liu & Li, 2026; Sweeney, 2026; Xu, 2026), and to our knowledge none edits only the stored noise, throughout training, with mean and covariance fixed. Associative recall (Ba et al., 2016; Bietti et al., 2023; Arora et al., 2024), where a model returns the value its context pairs with a queried key, gives us that mean. A context has only five queries, so the gradient of one sampled query splits exactly into their mean $\bar{g}$ and a residual r with zero conditional mean. We can then store in m and v something other than the sampled gradient, which still drives the update (Fig. 1(a)). With one sampled query per context, a four-block transformer stayed on the plateau for all 8,000 updates. Averaging the five queries let the same model learn in-context retrieval, which induction-style circuits carry out in larger models (Olsson et al., 2022). We call this transition acquisition. Fig. 1(b) shows the outcome at $\beta_1 = 0.5$ on the context in (a). After 8,000 updates, storing the reflection $\bar{g} - r$ answered all five queries in seven of eight seeds, storing only the mean in one, and storing the sampled gradient in none. Runs still on the plateau spread their predictions almost evenly over the five values, so their correct answers are chance.

We use this split in three steps. First, we average one source at a time to find where query sampling hurts, and at the default β_1 the harm sits in the stored moments rather than in the update (Sec. 3). Second, we build stored residuals on which per-step covariance and the long-run sum disagree. A linear model with a frozen preconditioner shows what to vary. At $\beta_1 = 0.9$ the current update applies a tenth of a new residual, and the other nine tenths reach the parameters over later updates through m. Storing $-r$ in place of r turns the residual's net weight from 1 into -0.8, while the noise in each step keeps its covariance. We call the net weight the long-run gain κ (Sec. 2.4). Sign masks that flip the stored residual on complementary random coordinate sets give two copies with the same per-step covariance, one mostly opposing r and the other mostly agreeing with it, with gains of -0.35 and 0.55. With $\beta_1 = 0.5$, the reflection $\bar{g} - r$ has zero gain but adds more noise per step than storing only the mean, so a per-step account would favor the mean and a long-run account the reflection (Sec. 4). Third, we look for a correlate of failure in the first block's attention to the queried key. We also ask

whether the effect holds outside recall, in a language model with loss on only 64 of its 511 positions so that the all-position gradient is the sampled one's exact mean (Sec. 4.3 and Sec. 5).

This paper makes three contributions:

- A way to edit only the stored noise. Because the split is exact, the current update and each stored moment can receive their own copy of the residual, so what AdamW stores can change at every update while the update itself still uses the sampled gradient as drawn.
- Evidence that stored noise matters beyond its per-step covariance. Stored copies with the same per-step covariance led to recall in very different fractions of runs, and across sign masks acquisition fell as the long-run gain moved away from zero either way. In recall with a short first-moment memory and matched early step sizes, a copy cancelling the residual later beat storing only the mean, though the linear model gives it more noise per step.
- A first look at where the effect acts and where it stops. Masks learned like the half copies they contain, which in hindsight points to the noise inside the residual's main directions, and in one study failing runs kept attention to the query below uniform. In a language model trained with a subsampled loss, the stored residual's sign still mattered at equal per-step covariance, but storing only the mean beat reflection at matched step size.

2 Separating current and stored sources

2.1 Task and the conditional mean

Each context lists $n = 5$ distinct keys paired with distinct values from disjoint alphabets of size $N = 8$, followed by a query marker, a queried key, and a READ token where the model predicts the key's value (Fig. 1(a)). A minibatch holds 256 contexts, each with its own sampled query J. For parameters θ and a context x, the all-query mean $\bar{g}(x) = \mathbb{E}_J[g_J(x) \mid x]$ averages the n query gradients, a Rao-Blackwellized gradient (Liu et al., 2019), and $r = g_J - \bar{g}$ is the residual of query sampling. Let H collect the pre-step parameters, optimizer state, and minibatch. Given H, the residual has $\mathbb{E}[r \mid H] = 0$ and covariance Σ. Because $\bar{g}$ is conditional on the minibatch, variation between minibatches remains in all runs, and we manipulate only the query-sampling residual.

2.2 Separate sources for the update and the moments

Each update combines a current source u_t with the previously stored moments, which are then updated from their own sources s_t^m and s_t^v and affect only later updates (Appendix C). All sources use the same pre-step parameters and minibatch:

$$\begin{aligned} m_t^u &= \beta_1 m_{t-1} + (1-\beta_1)u_t, & v_t^u &= \beta_2 v_{t-1} + (1-\beta_2)u_t^{\odot 2}, \\ m_t &= \beta_1 m_{t-1} + (1-\beta_1)s_t^m, & v_t &= \beta_2 v_{t-1} + (1-\beta_2)(s_t^v)^{\odot 2}. \end{aligned} \tag{1}$$

The parameters are updated as $\theta_t = \theta_{t-1} - \eta_t\big(\widehat{m}_t^u/(\sqrt{\widehat{v}_t^u} + \epsilon) + \lambda\theta_{t-1}\big)$, with learning rate η_t, decoupled weight decay λ, stabilizer ϵ, and hats denoting bias correction. We call $\eta_t\|\widehat{m}_t^u/(\sqrt{\widehat{v}_t^u}+\epsilon)\|$ the update norm. By default the current source is the sampled query's gradient, $u_t = g_J$, and setting $s_t^m = s_t^v = g_J$ as well recovers plain AdamW, which we call sampled training throughout.

2.3 Stored sources

The stored first-moment sources keep $\bar{g}$ and vary the residual: sampled $\bar{g} + r$, mean-only $\bar{g}$, reflected $\bar{g} - r$, scaled $\bar{g} + cr$ (including the half copies $c = \pm\frac{1}{2}$), and sign-masked $\bar{g} + D_{k/8}r$. For the masks, each seed assigns every coordinate to one of eight bins uniformly at random, and the diagonal sign matrix $D_{k/8}$ is -1 on k of the bins and $+1$ on the rest, with $D_{(8-k)/8} = -D_{k/8}$ for $k \neq 4$. A fixed T gives Tr the conditional covariance $T\Sigma T^\top$, and reflection leaves it at Σ. A mask keeps every coordinate's magnitude, $|(Dr)_i| = |r_i|$, and changes the covariance to $D\Sigma D$, so complementary masks share the stored covariance and coordinate magnitudes. $D_{6/8}$ flips the stored residual on six of the eight bins and so mostly opposes r, and $D_{2/8}$ flips two and mostly agrees with it. The two differ in the sign of their cross-covariance with the current residual and in their odd moments. These identities, like later claims of equal covariance, compare sources at one common state H, not along

runs that have diverged. Some studies give v the matched source, the elementwise $\sqrt{\bar{g}^2 + r^2}$. Its square has the sampled square's conditional mean, $\bar{g}^2 + \mathbb{E}[r^2 \mid H]$, but ignores the sign of r.

2.4 THE LONG-RUN GAIN

To predict which stored residuals should matter, we use a linear model of how noise propagates, which says nothing about learning itself. It freezes the preconditioner and learning rate and ignores bias correction, weight decay, and feedback on later gradients. A residual r drawn at one update enters that update's first-moment term m_t^u through u_t with weight $a = 1 - \beta_1$, and its stored copy Tr enters $\ell \geq 1$ updates later with weight $a\beta_1^\ell$. Summed over updates, the residual contributes $(aI + \beta_1 T)r$, whose component along r is

$$\kappa(T) = (1 - \beta_1) + \beta_1 q_T, \qquad q_T = \frac{\langle r, Tr \rangle}{\|r\|^2}. \tag{2}$$

We call κ the *long-run gain* and q_T the stored copy's alignment with r. The same quantity is the zero-frequency response of the momentum filter (Li et al., 2025) applied to the residual alone and the frozen-preconditioner case of AdaShift's net update factor (Zhou et al., 2019). In heavy-ball form it gives the $1/(1-\beta)$ in the momentum noise scale (Smith & Le, 2018). At $\beta_1 = 0.5$, for example, a residual enters the current update with weight 0.5 and a reflected copy enters later updates with total weight -0.5, so the two cancel and $\kappa = 0$. With a scaled copy $T = cI$ the whole contribution is κr, and at $\beta_1 = 0.9$ reflection has $\kappa = -0.8$ and the half copies $c = \mp\frac{1}{2}$ have $\kappa = -0.35$ and 0.55.

By analogy with anticorrelated noise, runs whose stored residual has a smaller $|\kappa|$ should learn faster. On a plateau the signal toward the solution is weak, and noisy gradients can keep training there (Kirsch et al., 2022). A copy that cancels the residual over time leaves less noise to build up against that weak signal. We drew this account post hoc from earlier runs, some with scaled copies, and tested it later in Sec. 4 and with the unclipped independent draw of Sec. 3, which it failed.

The sign of the stored copy does not change the per-step, or lag-0, covariance of m_t^u. For uncorrelated residuals with fixed covariance, its residual part has stationary covariance $a^2\Sigma + a^2\beta_1^2 T\Sigma T^\top/(1-\beta_1^2)$, the same for T and $-T$. At $\beta_1 = 0.5$, sampled and reflected residuals both give $\frac{1}{3}\Sigma$, more than the $\frac{1}{4}\Sigma$ of a mean-only m, yet their long-run gains are 1 and 0, while mean-only m has 0.5.

Copies that are not multiples of r, such as masks, also reapply a part orthogonal to r, which κ ignores. Gradients concentrate in a few directions (Gur-Ari et al., 2018), so the subspace power $\sqrt{P_S}$ compares a residual's summed contribution inside its top-8 principal subspace S with that of sampled training (Appendix D). It equals $|\kappa|$ for scaled copies and nearly so for masks. The measures disagree for an independent second draw. Its copy is uncorrelated with r, so $\kappa = 0.1$ at $\beta_1 = 0.9$, but keeps nearly all its noise in S ($\sqrt{P_S} \approx 0.91$). We introduced $\sqrt{P_S}$ after a clipped version failed.

2.5 MODEL, TRAINING, AND OUTCOMES

Our reference setting trains a four-block, width-160 decoder (the reference shape, 1.2M parameters) with AdamW at $(\beta_1, \beta_2) = (0.9, 0.95)$ for 8,000 updates (Appendix B). Some studies clip each source to global norm 1.0 on its own, and the exact split then holds only before clipping. A run *acquires* recall at the first evaluation where its accuracy on a fixed set of 512 contexts reaches 0.95. Runs that never acquire are right-censored, and we report Kaplan–Meier (KM) medians (Kaplan & Meier, 1958). A run's AUC is the area under its accuracy curve divided by the horizon. Since accuracy jumps at acquisition, the AUC mostly measures how early a run acquires. Conditions within a study share seeds, eight unless noted. Key comparisons use exact one-sided paired sign tests on the AUC in recall and on held-out loss or onset time in the language model, adjusted within each study by Holm or a fixed testing order. Appendix A maps each result to the appendix with its tests.

3 WHICH SOURCE CARRIES THE RESIDUAL'S EFFECT

We first ask through which of the three sources the sampling residual does its harm. Table 1 averages one source at a time at the default $\beta_1 = 0.9$. Giving the stored first moment the mean of the five queries was enough for recall to form in every run, whereas averaging only v let one run of eight acquire (Table 1(a)). A residual with scrambled signs and positions gave the same split, acquiring in

Table 1: Runs of eight that acquire recall by update 8,000, by the source of the current update and of the stored moments. The centered update has the all-query update as its mean and twice the sampled update's covariance. Other sources are each clipped to norm 1.0, but centered rows are not.

(a) Sampled update

Stored m	Stored v	Acquired
sampled	sampled	0/8
mean-only	sampled	8/8
sampled	mean-only	1/8
mean-only	mean-only	8/8

(b) All-query and centered updates

Current update	Stored m and v	Acquired
all-query	sampled	0/8
all-query	mean-only	8/8
centered	sampled	0/8
centered	mean-only	8/8

every run in m and in none in v alone. The current update mattered little. An all-query update did not help while the stored moments stayed sampled, and a sampled update learned once they were averaged (Table 1(b)). The centered update, which has the all-query update as its mean but twice the sampled covariance, split the same way, learning over averaged moments and failing over sampled ones. At this β_1, query sampling hurt through what AdamW stored, not what it applied.

The linear model of Sec. 2.4 gives a simple reason. The stored moments carry a fraction β_1 of each residual's summed weight, and averaging m lowers κ from 1 to 0.1. At $\beta_1 = 0.95$, mean-only m again acquired in every run and sampled training in none. At $\beta_1 = 0.3$ the balance tips the other way, since the current update then carries most of the weight. There, mean-only m acquired in one of eight runs and all-query training in all eight, which fits the long-run gain better than an account in which only the noise stored in m matters. A copy with random signs, which acquired in every run at $\beta_1 = 0.9$, never acquired at 0.3. Sampled training's failure is, at the least, a long delay. Over 32,000 updates only two of its eight runs acquired, at updates 15,750 and 24,250, while every all-query run had already acquired by update 750, a ratio of median times of more than 20 (Appendix E.5).

Averaging removes the residual from m, so it lowers the amount of stored noise and κ together. Copies that keep the residual's size but scramble its direction pull the two apart (Appendix E.2). We flipped each coordinate's sign at random (sign-mixed), also permuted coordinates in each tensor (sign-permuted), or drew a random direction with the residual's norm. All three have $\kappa = 0.1$ in expectation and acquired in all eight runs, and the sign-mixed copy had a larger AUC than sampled training in every pair (Holm $p = 0.016$). Reflection keeps even the residual's covariance, and with its gain of -0.8 it failed in every run. Averaging m also more than halved the early update norm, but halving or doubling the learning rate did not rescue sampled training, while sign-permuted m still acquired in every run at both rates. These copies change the covariance, though, so they cannot separate alignment from the shape of the stored noise, which the constructions of Sec. 4 hold fixed.

A small long-run gain was not sufficient, however. Storing the gradient of a second, independently drawn query gives m a copy with the residual's covariance and the κ of mean-only m. This copy never led to acquisition, with or without clipping, and mean-only m acquired in every paired run (Appendix E.2). Without clipping, its AUC fell below that of mean-only m in all eight pairs, by 0.72 on average. Unlike the sign-scrambled copies, it keeps its noise in the few directions the residual occupies, and as Sec. 2.4 noted, the subspace power fits this failure post hoc and κ alone does not.

4 Stored residuals with equal covariance

Sample count could explain the effect of averaging on its own, since mean-only m sees all five queries of each context. Sampled training with 1,280 contexts per update sees as many context-query pairs as mean-only m at 256, and all eight of its runs acquired, with the same KM median (Appendix E.3). Both constructions of Sec. 1 use all five queries, so sample count cannot separate them. If per-step covariance decided, the masks would tie and reflection at $\beta_1 = 0.5$ would lag mean-only m.

4.1 Complementary sign masks

The mask study stored $\bar{g} + D_{k/8}r$ in m for $k = 1, \ldots, 7$, without clipping and on new seeds, and tested the masks before the half copies (Appendix F). Its central pair, $D_{2/8}$ and $D_{6/8} = -D_{2/8}$, has

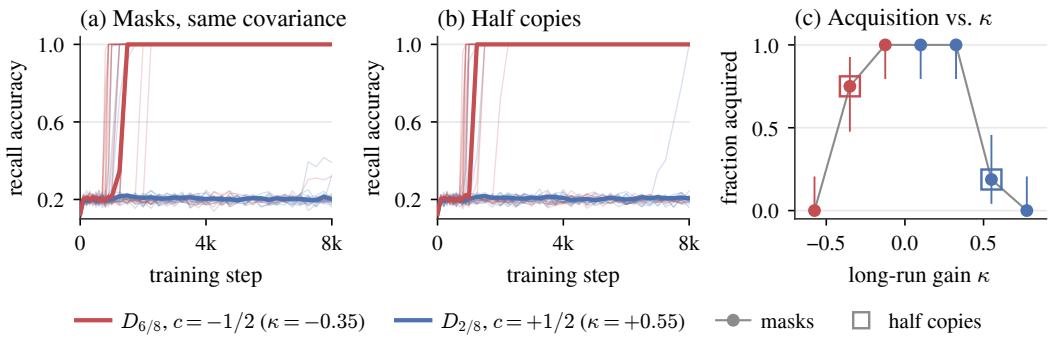


Fig. 2: Sign masks and half copies in recall at $\beta_1 = 0.9$. (a) Complementary masks $D_{6/8}$ and $D_{2/8}$, with equal stored covariance. (b) Half copies $c = \mp\frac{1}{2}$, with the masks' long-run gains. Thin lines are seeds and thick lines medians. (c) Fraction of runs acquiring by update 8,000 against κ on 16 shared seeds, with 95% Clopper–Pearson intervals (Clopper & Pearson, 1934). Squares mark half copies.

equal stored covariance and decomposes as

$$\bar{g} + D_{2/8}r = \bar{g} + \tfrac{1}{2}r + (D_{2/8} - \tfrac{1}{2}I)r, \qquad \bar{g} + D_{6/8}r = \bar{g} - \tfrac{1}{2}r - (D_{2/8} - \tfrac{1}{2}I)r, \tag{3}$$

so each mask is a half copy plus a remainder that is orthogonal to r in expectation. Running the half copies, which share the masks' gains but lack the remainder, tells us whether it matters. An earlier study had found $c = -\frac{1}{2}$ acquiring in 7 of 8 runs and $c = \frac{1}{2}$ in 1 of 8 (Appendix E.2).

$D_{6/8}$, whose copy mostly opposes r ($\kappa = -0.35$), acquired in 18 of its 24 runs, and $D_{2/8}$ ($\kappa = 0.55$) in 5 of its 24 (Fig. 2(a)). $D_{6/8}$ had the larger AUC in 21 of 24 pairs ($p = 1.4 \times 10^{-4}$), and the half copies came out in the same order (Fig. 2(b)). Across all seven masks, acquisition fell as $|\kappa|$ grew (Fig. 2(c)). Opposition alone did not help. The mask that opposes r most, $D_{7/8}$ ($\kappa = -0.575$), never acquired, nor did reflection at $\beta_1 = 0.9$. Among masks, the gain's size mattered more than its sign.

Although in the linear model a mask's update carries 2.55 times a half copy's per-step variance, on 16 shared seeds, masks and half copies with equal κ acquired equally often (12 at $\kappa = -0.35$, 3 at 0.55), so the remainder seemed irrelevant. The remainder has random signs. On frozen states from an earlier study of scaled copies, a random-sign copy put a median 0.4% as much energy into S as the residual itself (post hoc, Appendix D.3). Most of the masks' extra noise thus probably lands outside S.

One confound is the pairing of m with v. When v stores the sampled square $(\bar{g} + r)^2$, the cross-moment between the inputs of m and v flips sign between complementary masks. The matched v ignores the sign of r and so removes the mean part of that pairing, the cross term $2\bar{g}r$. With the matched v in every condition and otherwise the same recipe, $D_{6/8}$ had the larger AUC in all 16 pairs on new seeds ($p = 1.5 \times 10^{-5}$), and 13 of its 16 runs acquired, against 2 for $D_{2/8}$. An odd-moment coupling through $\mathbb{E}[r^3 \mid H]$ remains, and we did not test whether it affects acquisition.

The contrast held at eight blocks, where sampled training failed and v again stored the sampled square. An eight-seed test pointed the same way without rejecting (Appendix I.2). On 16 further seeds, 11 $D_{6/8}$ runs acquired and one $D_{2/8}$ run did, and $D_{6/8}$ had the larger AUC in 13 pairs ($p = 0.011$).

4.2 Reflection at zero long-run gain

Reflection is where the two views disagree most. At $\beta_1 = 0.5$ the reflected residual $\bar{g} - r$ has zero long-run gain (Sec. 2.4). In the linear model it behaves like an antithetic pair spread over updates (Liu & Xu (2018) draw such pairs within one minibatch) and like anticorrelated noise injection (Orvieto et al., 2022). Its per-step covariance, though, equals that of sampled training and exceeds that of mean-only m. An account in which per-step noise delays plateau escape (Sahasrabudhe, 2026) would therefore put reflection behind mean-only m, and the long-run gain puts it first of the three.

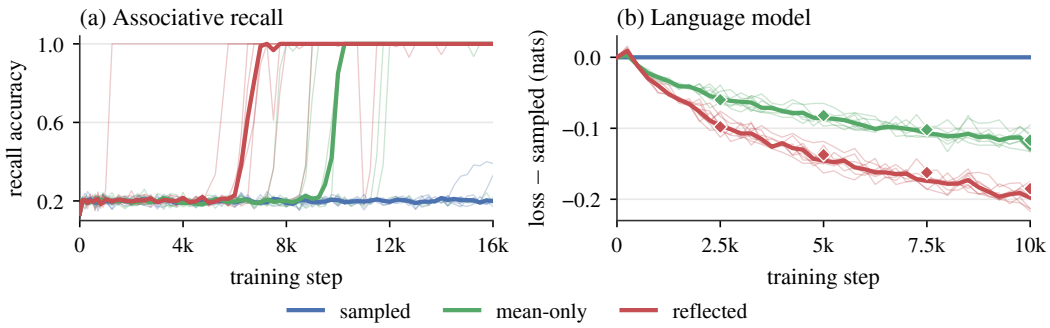


Fig. 3: Sampled, mean-only, and reflected m at $\beta_1 = 0.5$. Thin lines are seeds and thick lines medians. (a) Recall accuracy with sampled v, and reflection at 0.84 times the learning rate. (b) Language-model held-out loss minus that of sampled m, with the matched v, on a diagnostic set (lines) and the primary set (diamonds). The language-model run at matched update norm appears only in Table 2(b).

In our first reflection study, all eight reflected runs acquired, against none with sampled training and one with mean-only m. The reflected runs also took early steps about 19% larger, so we retrained them with the learning rate scaled by 0.84, which matched their early update norms to those of sampled training. Within 16,000 updates every rescaled run acquired (KM median 6,750). Seven mean-only runs acquired (median 10,250), and no sampled run did. Reflection had the larger AUC against both in every pair ($p = 0.012$, Fig. 3(a)). In recall, then, a stored copy that adds noise at every step but cancels over time beat a stored mean that adds less noise, at matched early step size.

It also held with the matched v in all three conditions (Appendix G.2). With the rescaled learning rate and 12,000 updates, every reflected run acquired by update 8,000 and no sampled run did ($p = 0.0039$ on AUC). Against mean-only m the difference was mostly one of speed, with KM medians of 4,500 and 9,250 updates and a larger reflected AUC in all eight pairs. In the measured windows before acquisition, reflected update norms stayed within 3% of mean-only m's under either v (post hoc).

4.3 Attention to the query

The READ position must pick up the queried key before the model can look up its value, so in a study of several stored copies at $\beta_1 = 0.9$ we measured how much the first block attends to it (Appendix J). There, the first block's attention from the READ token to the queried key fell below the uniform 1/13 in every condition over the first 600 updates, and further in runs storing the sampled or reflected residual. The four conditions that learned acquired by 900. By update 1,000 their attention had risen to 0.27–0.28, and it stayed near 0.03 in sampled and reflected runs. A bias toward the query raised sampled training's later AUC in most seeds (not significant jointly), and a clamp helped only while it stayed on (Appendix J.4). We treat low attention as a correlate of failure, not a mechanism.

5 Stored noise in a language model

We can ask the same question of a language model, but only with an artificial protocol: an exact mean requires subsampled supervision. The sampled loss covers 64 of each sequence's 511 positions, and the all-position gradient, which averages all 511, is its conditional mean. We trained a four-block decoder with 29M parameters from scratch on WikiText-103 (Merity et al., 2016) at a constant learning rate, without clipping unless noted (Appendix H). The subsampling handicaps sampled training by design, since each sampled update is supervised on about an eighth of the positions. Unlike in recall, mean-only m helped in this protocol only when v was averaged or matched as well. Over a sampled v, by contrast, it ended above sampled training, with loss spikes, and averaging both moments, which keeps the residual out of storage entirely, gave the lowest loss (Table 2(a)).

Reflection. As in recall, the current update stays sampled and only the stored sources change. At $\beta_1 = 0.5$ with the matched v, reflected m ended below both mean-only and sampled m in all eight

Table 2: Held-out loss (nats, median of eight seeds) after 10,000 sampled updates of the language model. Δ is the median paired difference from the first row, with the same sign in every seed.

(a) $\beta_1 = 0.9$

Stored m	Stored v	NLL	Δ
sampled	sampled	5.166	–
mean-only	mean-only	4.605	-0.563
mean-only	sampled	5.461	$+0.299$
mean-only	matched	4.889	-0.277

(b) $\beta_1 = 0.5$, *matched* v

Stored m	NLL	Δ
sampled	5.096	–
mean-only	4.981	-0.117
reflected	4.910	-0.185
reflected, 0.84η	5.016	-0.077

seeds, matching the order of their long-run gains, 0, 0.5, and 1 (Table 2(b), Fig. 3(b)). Reflected m again took larger steps, and the same 0.84 rescaling matched its update norm to mean-only m's, leaving it at 0.92 times sampled m's (Appendix H.4). At equal per-step covariance, the rerun still ended below sampled m in every seed ($p = 0.0078$). Against mean-only m it ended 0.033 nats higher (median). At matched step size, per-step noise predicts which ends lower, and κ does not.

Prefix-matching heads. Prefix-matching, or induction-style, heads (Olsson et al., 2022) attend from a repeated token to the token that followed its earlier occurrence. We call onset the first of four consecutive evaluations at which this attention score is at least 0.10. Ablating the top two such heads raised the loss more on repeated than on other text, relative to random head pairs, in 8/8 all-position and 7/8 sampled-moment models, as copying heads should (Appendix J.5). Sampled moments delayed onset under a sampled update in every pair of two 100,000-update replications, one on WikiText-103 and one, with clipping, on FineWeb-Edu (Penedo et al., 2024) (Appendix H.2). Batch and context size predict when such heads form in pretraining (Aoyama et al., 2026), but here onset tracked the extra positions only when they entered both stored moments. In a clipped factorial of 150,000 updates, changing the update's source left the onset where it was, and sampling both stored moments delayed it about twofold under either update (8/8 and 7/8 pairs). An eight-block run that varied only the moments' source repeated this (8/8). Compared at equal attained loss, however, all-position moments had the higher attention score in only one of the eight WikiText-103 pairs.

6 Related work

Amount and structure of the noise. The gradient noise scale reduces minibatch noise to one number (McCandlish et al., 2018), and SGD's noise scale grows as $1/(1-\beta_1)$ with momentum (Smith & Le, 2018; Smith et al., 2018). Anisotropy (Zhu et al., 2019; HaoChen et al., 2021), gradient sign and relative variance (Balles & Hennig, 2018), and heavy tails in transformers (Zhang et al., 2020) show magnitude is not enough. Noise with SGD's covariance but another distribution also regularizes (Wu et al., 2020). Minibatch noise changes how Adam's β_2 shapes its implicit bias (Cattaneo & Shigida, 2026). SDEs approximate adaptive methods (Malladi et al., 2022), and in Langevin analyses key properties depend on the noise only through its second moment (Cheng et al., 2020).

Noise that is correlated over time. With linearly correlated noise, the correlation across gradient-descent steps enters the error bound, and in the convex case anticorrelated noise gives a smaller error term than independent noise with the same bound on each draw (Koloskova et al., 2023). Injecting anticorrelated noise leads gradient descent to flatter minima (Orvieto et al., 2022), and Lucchi et al. (2022) generalize this injection to fractional noise with correlated increments. Differentially private training correlates its noise to cancel in the iterates (Kairouz et al., 2021; Denisov et al., 2022; Choquette-Choo et al., 2024). Random reshuffling converges faster than sampling with replacement because each epoch's gradient errors sum to zero at fixed parameters (Gürbüzbalaban et al., 2021; Mishchenko et al., 2020), and its anticorrelated noise lowers the weight variance of momentum SGD in flat directions (Kühn & Rosenow, 2025). Momentum lowers per-step variance but, to first order in the learning rate, not the variance accumulated along the trajectory (Gitman et al., 2019; Wang et al., 2024). In these works the correlation comes from the data order, from injected noise, or from momentum itself. We keep each sampled gradient as drawn and change only AdamW's stored copy.

Optimizers that edit the memory. Prior designs reweight the current gradient against momentum or negate momentum (Ma & Yarats, 2019; Gidel et al., 2019), build v from a shifted gradient (Zhou et al., 2019), store a variance-reduced first moment (Cutkosky & Orabona, 2019; Yuan et al., 2025), add a slow component (Pagliardini et al., 2025; Lee et al., 2024), or mask coordinates that oppose the current gradient (Liang et al., 2026). SAG, SAGA, and SVRG use past gradients as control variates (Schmidt et al., 2017; Defazio et al., 2014; Johnson & Zhang, 2013). We instead change only the stored residual, around the exact mean, at every update of a full run. MARS computes v from its variance-reduced gradient and calls the alternative a mismatch (Yuan et al., 2025), and ADOPT removes the current gradient from v (Taniguchi et al., 2024). Our matched v applies the principle of batch-size-invariant Adam, which averages squared microbatch gradients in v (Wang & Aitchison, 2024), to a single draw. Adam works better when the squared mean of per-example gradients dominates its preconditioner (Roulet & Agarwala, 2026). Closest to our work, Liu & Li (2026) track one perturbed step written into the parameters, m, or v over a few dozen updates. They find that a deviation in m peaks later than one in the parameters and accumulates a substantially larger response, and that one in v alone stays small. In concurrent state-surgery experiments, Xu (2026) shows that AdamW's first moment carries a transferred perturbation whose behavioral effect later training decides. Sweeney (2026) derives the variance over data orderings within a fine-tuning window from AdamW's impulse weights under a frozen preconditioner, and Guo (2026) finds that such delayed transport through AdamW's moments changes short-horizon training decisions.

Recall formation. Recall and abrupt learning have been tied to the training distribution (Bietti et al., 2023; Reddy, 2024; Chan et al., 2022), associative-memory dynamics (Cabannes et al., 2024), and induction-head subcomputations (Olsson et al., 2022; Singh et al., 2024), and plateaus to softmax and sparse-attention dynamics (Hoffmann et al., 2024; Zucchet et al., 2025; Gopalani & Hu, 2025). Gradient noise can lengthen such plateaus, since larger batches shorten them (Kirsch et al., 2022) and smaller batches or larger learning rates delay escape from them (Sahasrabudhe, 2026).

7 Scope and limits

The mask contrast replicated at eight blocks but vanished at widths 320 and 640 and shrank to a speed difference at $\beta_2 = 0.999$, settings where every mask run acquired. Sampled training's failure itself depends on the recipe. Query-key normalization (Henry et al., 2020; Wortsman et al., 2024) and Lion (Chen et al., 2023) removed it within 8,000 updates and Muon (Jordan et al., 2024) later, and $\beta_2 = 0.999$ and a cosine schedule did not (Appendix I). The long-run gain alone does not predict acquisition. Neither it nor the subspace power, introduced after seeing outcomes, accounts for v. In the language model, reflection shows an effect of the stored residual, not a better optimizer.

8 Conclusion

When we edited only what AdamW stores and left the applied noise as drawn, how the stored noise was arranged over time could decide whether a small transformer learned in-context retrieval within its training budget. Complementary masks shared the batch size, learning rate, and per-step noise covariance, and still led to recall in very different fractions of seeds. Across masks, recall became more likely as the long-run gain approached zero from either side. Summaries of minibatch noise by per-step size or covariance can therefore miss a variable that AdamW's memory sets, namely how each stored copy lines up with the residual it copies. The language model shows that this variable does not act alone, as per-step noise predicted the comparison with a stored mean at matched step size. In one study, failing runs also kept the first block's attention to the query below uniform. We cannot yet explain this correlate, and whether the accumulated noise holds it down is untested.

Optimizers that edit the memory (Sec. 6) can change this alignment along with the amount of stored noise, and where the mean is known our split can tell the two apart. It needs the exact mean of the sampled gradient, which associative recall and subsampled supervision provide. A second minibatch is the obvious substitute in ordinary pretraining, but in recall a stored independent draw blocked acquisition despite its small gain (Sec. 3), so how noisy a usable substitute may be is open.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing writing, retrieval and discovery, draft sections of the paper, generating synthetic datasets, drawing and editing the figures, mathematical claims, and implementing and debugging codes. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Tatsuya Aoyama, Ethan Gotlieb Wilcox, and Nathan Schneider. Predicting the emergence of induction heads in language model pretraining. In *International Conference on Machine Learning*, 2026. arXiv:2511.16893.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Y. Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, volume 2024, pp. 15664–15730, 2024.

Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, volume 29, 2016. arXiv:1610.06258.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 10–15 Jul 2018.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1560–1588, 2023. arXiv:2306.00802.

Vivien Cabannes, Berfin Şimşek, and Alberto Bietti. Learning associative memories with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5114–5134. PMLR, 2024. URL https://proceedings.mlr.press/v235/cabannes24a.html.

Matias D. Cattaneo and Boris Shigida. The effect of mini-batch noise on the implicit bias of Adam. arXiv preprint arXiv:2602.01642, 2026.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891. Curran Associates, Inc., 2022. doi: 10.52202/068431-1371. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/77c6ccacfd9962e2307fc64680fc5ace-Paper-Conference.pdf.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49205–49233. Curran Associates, Inc., 2023. doi: 10.52202/075280-2140. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a39b4925e35cf447ccba8757137d84f-Abstract.html.

Xiang Cheng, Dong Yin, Peter L. Bartlett, and Michael I. Jordan. Stochastic gradient and Langevin processes. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Dami Choi, Alexandre Passos, Christopher J. Shallue, and George E. Dahl. Faster neural network training with data echoing. arXiv preprint arXiv:1907.05550, 2019.

Christopher Choquette-Choo, Krishnamurthy Dvijotham, Krishna Pillutla, Arun Ganesh, Thomas Steinke, and Abhradeep Guha Thakurta. Correlated noise provably beats independent noise for differentially private learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 4616–4676, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1384616b65241ef17aad3f3bde8fd623-Paper-Conference.pdf.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, volume 27, pp. 1646–1654, 2014.

Sergey Denisov, H. Brendan McMahan, John Rush, Adam Smith, and Abhradeep Guha Thakurta. Improved differential privacy for SGD via optimal private linear operators on adaptive streams. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5910–5924. Curran Associates, Inc., 2022. doi: 10.52202/068431-0428. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/271ec4d1a9ff5e6b81a6e21d38b1ba96-Paper-Conference.pdf.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1802–1811. PMLR, 2019. URL https://proceedings.mlr.press/v89/gidel19a.html.

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pulkit Gopalani and Wei Hu. What happens during the loss plateau? Understanding abrupt learning in Transformers. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 54928–54960. Curran Associates, Inc., 2025. doi: 10.52202/085713-1838. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4f06c73c45d2625f0e687f7e6a206332-Abstract-Conference.html.

Jinhui Guo. Delayed optimizer-state transport shapes short-horizon training decisions. *arXiv preprint arXiv:2608.24593*, 2026.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. arXiv preprint arXiv:1812.04754, 2018.

M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, Mar 2021. doi: 10.1007/s10107-019-01440-w.

Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2315–2357, 2021.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020. doi: 10.18653/v1/2020.findings-emnlp.379.

David T. Hoffmann, Simon Schrodi, Jelena Bratulić, Nadine Behrmann, Volker Fischer, and Thomas Brox. Eureka-moments in transformers: Multi-step tasks reveal softmax induced optimization problems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 18409–18438, 2024. URL https://proceedings.mlr.press/v235/hoffmann24a.html.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, pp. 315–323, 2013.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5213–5225. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kairouz21b.html.

E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. doi: 10.1080/01621459.1958.10501452.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. arXiv preprint arXiv:2212.04458, 2022. NeurIPS 2022 Workshop on Meta-Learning.

Anastasiia Koloskova, Ryan McKenna, Zachary Charles, John Rush, and H. Brendan McMahan. Gradient descent with linearly correlated noise: Theory and applications to differential privacy. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 35761–35773. Curran Associates, Inc., 2023. doi: 10.52202/075280-1552. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/70255afc962aca0930327c090eb7d8c5-Paper-Conference.pdf.

Marcel Kühn and Bernd Rosenow. Anti-correlated noise in epoch-based stochastic gradient descent: implications for weight variances in flat directions. *Machine Learning: Science and Technology*, 6 (4):045003, 2025.

Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0948. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/350e718ff74062b4bac2c6ffd9e1ac66-Abstract-Conference.html.

Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients. arXiv preprint arXiv:2405.20233, 2024.

Xianliang Li, Jun Luo, Zhiwei Zheng, Hanxiao Wang, Li Luo, Lingkun Wen, Linlong Wu, and Sheng Xu. On the performance analysis of momentum method: A frequency domain perspective. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 47713–47734, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/76c64372d25513fed60a9e4176a62481-Paper-Conference.pdf.

Kaizhao Liang, Lizhang Chen, Bo Liu, and Qiang Liu. Cautious optimizers: Improving training with one line of code. In *International Conference on Learning Representations*, 2026.

Jingchang Liu and Linli Xu. Accelerating stochastic gradient descent using antithetic sampling. *arXiv preprint arXiv:1810.03124*, 2018.

Kang Liu and Suyan Li. Finite-horizon input-output dynamics of minibatch perturbations in AdamW. arXiv preprint arXiv:2608.19762, 2026.

Runjing Liu, Jeffrey Regier, Nilesh Tripuraneni, Michael I. Jordan, and Jon McAuliffe. Rao-blackwellized stochastic gradients for discrete distributions. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. arXiv:1711.05101.

Aurelien Lucchi, Frank Proske, Antonio Orvieto, Francis Bach, and Hans Kersting. On the theoretical properties of noise correlation in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. arXiv:2209.09162.

Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and Adam for deep learning. In *International Conference on Learning Representations*, 2019. arXiv:1810.06801.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7697–7711. Curran Associates, Inc., 2022. doi: 10.52202/068431-0559. arXiv:2205.10287.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. arXiv preprint arXiv:1812.06162, 2018.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtarik. Random reshuffling: Simple analysis with vast improvements. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17309–17320. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8cc6e90ccbff44c9cee23611711cdc4-Paper.pdf.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Antonio Orvieto, Hans Kersting, Frank Proske, Francis Bach, and Aurelien Lucchi. Anticorrelated noise injection for improved generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17094–17116. PMLR, 17–23 Jul 2022. arXiv:2202.02831.

Matteo Pagliardini, Pierre Ablin, and David Grangier. The AdEMAMix optimizer: Better, faster, older. In *International Conference on Learning Representations*, volume 2025, pp. 64715–64757, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a2cf225ba392627529efef14dc857e22-Paper-Conference.pdf.

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, Datasets and Benchmarks Track*, 2024.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, volume 2024, pp. 19226–19239, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/540a3bf4863672cdaebfed69fa2ce335-Paper-Conference.pdf.

Vincent Roulet and Atish Agarwala. Per-example gradients: a new frontier for understanding and improving optimizers. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YB6U1yWcGy.

Mihir Sahasrabudhe. Marginals before conditionals. arXiv preprint arXiv:2603.10074, 2026.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1–2):83–112, 2017.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Interspeech 2014*, pp. 1058–1062, 2014. doi: 10.21437/Interspeech.2014-274.

Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45637–45662. PMLR, 2024. URL https://proceedings.mlr.press/v235/singh24c.html.

Samuel L. Smith and Quoc V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJij4yg0Z.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1Yy1BxCZ.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations*, 2025. arXiv:2405.16002.

John Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise. arXiv preprint arXiv:2606.29554, 2026.

Shohei Taniguchi, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. ADOPT: Modified Adam can converge with any β_2 with the optimal rate. In *Advances in Neural Information Processing Systems*, 2024.

Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 13514–13557, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3a6935d11910d6f9142b0a1e36fc6753-Paper-Conference.pdf.

Xi Wang and Laurence Aitchison. Batch size invariant Adam. arXiv preprint arXiv:2402.18824, 2024.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alexander Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale Transformer training instabilities. In *International Conference on Learning Representations*, volume 2024, pp. 49844–49869, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d848cb2c84f0bba7f1f73cf232734c40-Paper-Conference.pdf.

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10367–10376, 2020.

Zeke Xie, Li Yuan, Zhanxing Zhu, and Masashi Sugiyama. Positive-negative momentum: Manipulating stochastic gradient noise to improve generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11448–11458. PMLR, 2021.

Qinyang Xu. Stored in optimizer state, valued by later training: A causal account of subliminal trait transfer. *arXiv preprint arXiv:2608.20442*, 2026.

Huizhuo Yuan, Yifeng Liu, Shuang Wu, Xun Zhou, and Quanquan Gu. MARS: Unleashing the power of variance reduction for training large models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73553–73587. PMLR, 2025. URL https://proceedings.mlr.press/v267/yuan25f.html.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. AdaShift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations*, 2019.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7654–7663. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhu19e.html.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11127–11135, June 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zou_A_Sufficient_Condition_for_Convergences_of_Adam_and_RMSProp_CVPR_2019_paper.html.

Nicolas Zucchet, Francesco D'Angelo, Andrew Lampinen, and Stephanie Chan. The emergence of sparse attention: impact of data distribution and benefits of repetition. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 35121–35158. Curran Associates, Inc., 2025. doi: 10.52202/085713-1179. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/323b34ddc3b1787a685f4664d8dbb76c-Abstract-Conference.html.

A GUIDE TO THE APPENDIX

Table 3 maps each main-text claim to its direct tests and their limits, and Table 4 lists the notation.

Table 3: Evidence map. Main-text claims, their direct tests with locations, and their limits. No main-text claim rests on the exploratory results in the last row or on Appendices D.4 and J, apart from the head ablations of Appendix J.5.

Claim (main text)	Direct tests (location)	Limits
Averaging the stored m, not v, let runs acquire at $\beta_1 = 0.9$ (Sec. 3)	allocation, crossed and centered updates (Table 1, Appendix E.1); decay study (Appendix E.4)	depends on the protocol (shared query index: 2/8; per-source clipping); not for every v source (Appendix H.3)
Stored residuals with equal conditional covariance and coordinate magnitudes gave different outcomes (Sec. 4.1)	complementary masks, 24 seeds (Appendix F); masks with matched v, 16 new seeds (Appendix F.1)	reference shape and eight blocks (16-seed replication), speed only at $\beta_2 = 0.999$, and the failing $\lvert\kappa\rvert$ moved with shape (Appendix I.2); the odd-moment coupling is kept and its effect untested
At $\beta_1 = 0.5$, reflected m (zero long-run gain) had a larger AUC than sampled and mean-only m (Sec. 4.2)	zero-gain study (Appendix G); follow-up with early step size matched (Appendix G.1); follow-up with matched v (Appendix G.2)	sampled v in the first two studies (the matched v keeps the odd-moment pairing); step size matched to sampled training over updates 1–1,000 and to mean-only m in later windows (post hoc); against mean-only m mainly a difference in speed; zero gain did not determine when runs acquired
Acquisition fell as $\lvert\kappa\rvert$ grew across masks (Sec. 4.1, descriptive) and as $\sqrt{P_S}$ grew across recall conditions (Sec. 2.4 and Sec. 7, post hoc)	seven-mask sweep (Appendix F); frozen-state retention and post hoc placement (Appendices D.1 to D.3); unclipped independent draw (Appendix E.2)	descriptive; $\lvert\kappa\rvert$ not separated from κ^2 and failed for the independent draw; $\sqrt{P_S}$ defined post hoc; retention measured on scaled-copy states, not mask runs, and (A5) untested; exceptions at $N = 16$ and in multi-query recall (Appendices I.2 and I.3)
In a language model, with matched v reflected m lowered the loss, sampled moments delayed prefix-matching onset, and the effect of mean-only m depended on the v source (Sec. 5)	reflection and pairing studies (Appendices H.3 and H.4); onset replications, one clipped (Appendix H.2); head ablations (Appendix J.5)	artificially subsampled supervision; no all-position condition at $\beta_1 = 0.5$; at matched update norm, reflection trails mean-only m; onset lead reverses at equal loss; the clipped protocol differs (Appendix H.2)
Exploratory (Sec. 7)	query order (Appendix I.4); partial averaging (Appendix D.4); other studies (Appendix J)	partial averaging does not separate the long-run gain from per-step variance; query order null near the floor

B EXPERIMENTAL SETUP

B.1 TASK, MODEL, AND RECIPES

Each recall sequence has 13 tokens (Sec. 2.1), and all conditions on a seed see the same contexts and queries. The sampled gradient g_J draws one query per context, uniformly and independently, and is computed with $\bar{g}$ at the same parameters in FP32 without TF32. Both process 1,280 sequences per update, but g_J repeats each of its 256 (context, query) pairs five times and $\bar{g}$ holds 1,280 distinct pairs.

Each study was evaluated on one fixed set of 512 contexts with one query each. Most studies evaluated every 100 updates up to 1,000 and every 250 thereafter; the Q/K study evaluated every 50 updates up to 1,500 and every 100 thereafter. The runs of Table 1 used coarser grids, and its crossed and centered runs could stop 1,000 updates after acquisition.

Table 4: Notation. Hats on m and v denote bias correction, $D_{(8-k)/8} = -D_{k/8}$ for $k \neq 4$, and p is always a p-value. LM: language model.

Symbol	Meaning	Symbol	Meaning
n	pairs per recall context (5)	θ_t	parameters after update t
N	key and value alphabet size (8)	$m_t,\ v_t$	stored first and second moments
K	pairs per multi-query sequence	$m_t^u,\ v_t^u$	moments applied at update t
B	contexts per update (default 256)	$\beta_1,\ \beta_2$	decay rates of m and v
x	recall context (LM: sequence)	a	$1-\beta_1$
$J,\ J'$	sampled query; independent draw	$\eta_t,\ \lambda,\ \epsilon$	learning rate, weight decay, Adam's ϵ
n'	queries in a subset $\mathcal{S}$	Λ	frozen preconditioner $\mathrm{diag}(1/(\sqrt{v}+\epsilon))$
H	pre-step parameters, state, minibatch	d_t	Adam direction $\widehat{m}_t^u/(\sqrt{\widehat{v}_t^u}+\epsilon)$
g_J	sampled gradient, one query per context	C_t	coherent-pair ratio bound
$g_j,\ g_{\mathcal{S}}$	gradient of query j; mean over $\mathcal{S}$	q_T	overlap $\langle r, Tr\rangle/\|r\|^2$
$\bar{g}$	all-query (LM: all-position) mean	$\kappa(T),\ \kappa$	long-run gain $a+\beta_1 q_T$
$r,\ r'$	residual $g_J-\bar{g}$; that of J'	A_T	total residual weight $aI+\beta_1 T$
$\Sigma,\ \Sigma_x$	$\mathrm{Cov}(r \mid H)$; within context x	P	long-run power (sampled training: 1)
$\alpha(\bar{g})$	common gradient scale (LM)	$S,\ \Pi_S$	gradient subspace; its projector
$w(J)$	per-source clipping factor	$q_S,\ \rho_S$	overlap and retention in S
u_t	current source of update t	$\kappa_S,\ P_S$	gain $a+\beta_1 q_S$ and power in S
$s_t^m,\ s_t^v$	stored sources of m and v	$z_t,\ \gamma(\ell)$	numerator residual; lag-ℓ covariance
$T,\ \zeta$	stored copy Tr; any stored residual	$E_t,\ M_t$	residual $m_t - M_t$; first moment of the $\bar{g}_t$
c	scaled copy $\bar{g}+cr$	$\widehat{S}(t)$	Kaplan–Meier estimate
D	fixed signs $\mathrm{diag}(d_i)$, fair unless stated otherwise	AUC	trapezoidal area under accuracy over the grid, divided by the horizon
$D_{k/8}$	k of 8 bins negative, $\mathbb{E}\,q = 1-2k/8$ (design overlap)	NLL	held-out negative log-likelihood (nats)
$Q,\ R$	permutation; $R = DQ$	τ_b	Kendall's rank correlation
ξ	random redraw with the residual's norm	p_{IUT}	intersection–union test p-value
$U_H,\ \widetilde{U}$	AdamW increment at H; centered update	KM	Kaplan–Meier median (updates)

The model is a pre-layer-normalized decoder with four blocks of width 160, ten heads of width 16, a GELU MLP of width 640, and 13 learned position embeddings (1,241,120 parameters, with an 18-token vocabulary, bias-free attention projections, MLP biases, and a bias-free 8-way readout over the value alphabet at the READ position). The *unclipped* reference recipe runs AdamW for 8,000 updates at $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and decoupled weight decay 0.01, warming the learning rate up linearly over 800 updates to a constant 3×10^{-4}.

The *clipped* recipe clips each separately assigned source to global norm 1.0 on its own (the v source before squaring), so the mean identity of Sec. 2.3 need not survive clipping. Table 5 gives each experiment's departures from the unclipped reference recipe, and the language-model recipes are in Appendix H.

B.2 Outcomes and statistics

Acquisition, the first evaluation with accuracy at least 0.95 (Sec. 2.5), is read off the grid without interpolation, and a run that never crosses is right-censored at its horizon. On the irregular grids above, the AUC differs from the mean of the evaluations.

Accuracy need not stay above 0.95 after acquisition, but dips were brief. Two post-acquisition accuracies in 224 earlier runs fell below 0.90, each followed by at least 0.95 at the next evaluation. Of the 23 acquiring doubled-horizon runs, six dipped below 0.95 for at most two evaluations, and all 23 reached accuracy 1.0 by update 16,000.

The Kaplan–Meier median (Kaplan & Meier, 1958) is the first grid time at which the estimated fraction not yet acquired, $\widehat{S}(t)$, is at most $\frac{1}{2}$. When fewer than half cross, it is reported as a bound such as >8,000. Unless marked distribution-free, its 95% interval inverts Greenwood's log-log pointwise band and is coarse.

Table 5: Departures from the unclipped reference recipe (256 contexts per update, $\beta_1 = 0.9$, 8,000 updates, eight seeds per condition, paired within an experiment). Q/K: query/key. RMS: root-mean-square.

Experiment	Change to the recipe
Sec. 3	
averaged moments, crossed update	per-source clipping
centered update	centered current update
allocation	also run with per-source clipping
learning rate	peak rate 1.5×10^{-4} or 6×10^{-4}
contexts per update	256 to 2,560 contexts
first-moment decay	$\beta_1 \in \{0.3, 0.5, 0.9, 0.95\}$
Sec. 4	
sign masks	16 or 24 seeds
zero long-run gain	$\beta_1 \in \{0.3, 0.5, 0.9\}$
doubled horizon	$\beta_1 = 0.5$; 16,000 updates; rate $\times 544/647$ in one condition
Sec. 7	
long horizon	32,000 updates, evaluated until the first crossing
recipe changes	$\beta_2 = 0.999$ or a cosine schedule
Q/K normalization	Q/K RMS normalization; 3,000 updates; 12 seeds
model shape	width 320 or 640, or eight blocks
alphabet size	$N = 16$

The unit of each test is the seed, or the parent run in continuation studies, and seeds are never pooled across experiments. Language-model sign tests use the held-out-loss endpoint stated with each study. Tests other than paired sign tests are named where they occur. Zero differences and recall pairs with a lost or non-finite run count as non-positive, and the language-model pairing and reflection studies rank a failed run worst. With eight seeds, 8/8 positive pairs give $p = 1/256 \approx 0.0039$ and 7/8 give $9/256 \approx 0.035$, so Holm adjustment over two tests needs 8/8 for the first rejection. Multiple tests within a study are Holm-adjusted or run in a fixed sequence, and we claim no error control across studies. Paired bootstrap intervals (10,000 resamples of whole seeds) and Kaplan–Meier medians are descriptive, and two-sided 95% Clopper–Pearson intervals (Clopper & Pearson, 1934) on acquisition fractions bound acquisition within the horizon, not its timing.

C GRADIENT SOURCES

As in Sec. 2, the identities below hold at a common state H and do not equate realized gradients, covariances, or second-moment inputs once runs diverge.

Among the sources of Table 6, the sign-mixed D has fair signs $d_i = \pm 1$, and the masks $D_{k/8}$ are those of Sec. 2.3. The sign-permuted source uses $R = DQ$, where the permutation Q has $(Qr)_i = r_{\pi(i)}$. One control stores only $\bar{g} + Qr$, with residual covariance $Q\Sigma Q^\top$. In each tensor b of n_b entries, the random redraw sets $\xi_b = \|r_b\| z_b / \|z_b\|$ with $z_b \sim \mathcal{N}(0, I)$. The independent draw uses a query J' drawn independently of J. Sharing one query index across the minibatch keeps the mean but changes the cross-context covariance.

Query subsets. For a context x with query gradients $g_1, \ldots, g_n$ and $\Sigma_x = \frac{1}{n}\sum_j (g_j - \bar{g})(g_j - \bar{g})^\top$, a uniformly drawn subset $\mathcal{S}$ of n' distinct queries has

$$\mathbb{E}[g_{\mathcal{S}} \mid x] = \bar{g}, \qquad \mathrm{Cov}(g_{\mathcal{S}} \mid x) = \frac{n - n'}{n'(n-1)}\,\Sigma_x. \tag{4}$$

Sampled training is the case $n' = 1$, and $n = 511$ with $n' = 64$ gives the sampled gradient of the language model. The partial average of Appendix D.4 keeps the drawn query and adds $n' - 1$ of the other four. Its stored residual s has $\mathbb{E}[s \mid J, H] = qr$ with $q = (5 - n')/(4n')$, because the five residuals sum to zero.

Table 6: Gradient sources and their statistics given H (the pre-step parameters, optimizer state, and minibatch). Every source of u_t and s^m has conditional mean $\bar{g}$, and q is the overlap of the stored residual with r, in expectation for randomized sources.

s^m source	Value	Residual cov.	q
sampled	$\bar{g} + r$	Σ	1
mean-only	$\bar{g}$	0	0
reflected	$\bar{g} - r$	Σ	-1
scaled copy	$\bar{g} + cr$	$c^2\Sigma$	c
sign-mixed	$\bar{g} + Dr$	$D\Sigma D$	0
mask	$\bar{g} + D_{k/8}r$	$D_{k/8}\Sigma D_{k/8}$	$1 - 2k/8$
sign-permuted	$\bar{g} + Rr$	$R\Sigma R^\top$	0
random redraw	$\bar{g} + \xi$	$\mathrm{tr}(\Sigma_{bb})I/n_b$	0
independent draw	$g_{J'}$	Σ	0
partial average	$g_{\mathcal{S}}, \lvert\mathcal{S}\rvert = n'$	$q\Sigma$	$\frac{5-n'}{4n'}$

u_t source	Value	Residual cov.
sampled	$g_J = \bar{g} + r$	Σ
all-query	$\bar{g}$	0
s^v source	Value	$\mathbb{E}[(s_i^v)^2 \mid H]$
sampled	$\bar{g} + r$	$\bar{g}_i^2 + \Sigma_{ii}$
mean-only	$\bar{g}$	$\bar{g}_i^2$
matched	$\sqrt{\bar{g}^2 + r^2}$	$\bar{g}_i^2 + \Sigma_{ii}$
sign-mixed	$\bar{g} + Dr$	$\bar{g}_i^2 + \Sigma_{ii}$
sign-permuted	$\bar{g} + Rr$	$\bar{g}_i^2 + \Sigma_{\pi(i)\pi(i)}$
independent draw	$g_{J'}$	$\bar{g}_i^2 + \Sigma_{ii}$

The matched second moment. Elementwise, $(\bar{g} + r)^2 = (\bar{g}^2 + r^2) + 2\bar{g}r$, and the cross term has conditional mean zero, so the matched square has the sampled square's conditional mean $\bar{g}^2 + \mathbb{E}[r^2 \mid H]$. It never falls below $\bar{g}^2$ and does not depend on the sign of r, so it treats reflected and sampled first moments alike.

Complementary copies. A complementary pair replaces a stored residual Tr by $-Tr$, as with sampled and reflected m, c and $-c$, or $D_{k/8}$ and $D_{(8-k)/8} = -D_{k/8}$ for $k \neq 4$ (Appendix F). Negation keeps every even moment, such as $T\Sigma T^\top$, and a mask's coordinate magnitudes. It flips every odd moment, q_T, and the coupling to a sampled v of the same draw,

$$\mathrm{Cov}\big(\bar{g}_i + d_i r_i,\ (\bar{g}_i + r_i)^2 \mid H\big) = d_i\big(2\bar{g}_i\,\mathbb{E}[r_i^2 \mid H] + \mathbb{E}[r_i^3 \mid H]\big). \tag{5}$$

The matched v removes the first term but not the second, $d_i\,\mathbb{E}[r_i^3 \mid H]$, which vanishes when the conditional third moment is zero (for example, for a symmetric residual). Removing one identified coupling term does not make the m and v sources independent.

The centered update. Let $U_H(g)$ be the whole AdamW increment at state H with current source g, weight decay included. Because U_H is nonlinear, a centered run centers increments, not gradients, and applies $\widetilde{U} = U_H(g_J) - U_H(g_{J'}) + U_H(\bar{g})$ with independent per-context draws J and J', while its stored sources use J. The first two terms are independent with the same law, so $\mathbb{E}[\widetilde{U} \mid H] = U_H(\bar{g})$ and $\mathrm{Cov}(\widetilde{U} \mid H) = 2\,\mathrm{Cov}(U_H(g_J) \mid H)$.

Clipping and a common gradient scale. Per-source clipping (Table 5 and the 150,000-update language-model study) scales each source to global norm at most 1.0 by a draw-dependent factor $w(J) = \min(1, 1/\|g_J\|)$. The clipped sampled source then has mean $\mathbb{E}[w(J)g_J \mid H]$, rather than the clipped mean, so the mean–residual split holds only before clipping. The unclipped language-model experiments (Table 2 and Appendix H) instead multiply every source by one scalar set by the mean,

$$\alpha(\bar{g}) = \begin{cases} 1, & \|\bar{g}\|_2 \le 1, \\ (\|\bar{g}\|_2 + 10^{-6})^{-1}, & \|\bar{g}\|_2 > 1, \end{cases} \qquad \mathbb{E}[\alpha(\bar{g})g_J \mid H] = \alpha(\bar{g})\,\bar{g}. \tag{6}$$

Every identity here holds with $\alpha\bar{g}$ and αr in place of $\bar{g}$ and r, because α depends only on H. Computing $\alpha(\bar{g})$ requires the all-position mean at every update, even in conditions that store no mean-based source (Appendix H.7).

Order of operations. Algorithm 1 restates Eq. (1) and the parameter update of Sec. 2.2 as one update with separate sources.

Algorithm 1. One update with separate sources.

Input: parameters θ_{t-1}, stored moments m_{t-1} and v_{t-1}, a minibatch of contexts (language model: sequences).

1. At θ_{t-1}, draw a query J per context (language model: 64 of 511 positions per sequence), and any further independent draw that a source uses. Compute g_J and, where a source or the common scale needs it, the all-query (language model: all-position) mean $\bar{g}$ and the residual $r = g_J - \bar{g}$.
2. Form the current source u_t and the stored sources s_t^m and s_t^v from $\bar{g}$ and r (Table 6). The clipped recipe clips each source to global norm 1.0 on its own, the v source before squaring. The unclipped language-model recipe multiplies every source by $\alpha(\bar{g})$ (Eq. (6)).
3. Temporary moments: $m_t^u = \beta_1 m_{t-1} + (1-\beta_1)u_t$ and $v_t^u = \beta_2 v_{t-1} + (1-\beta_2)u_t^{\odot 2}$.
4. Update: $\theta_t = \theta_{t-1} - \eta_t\big(\widehat{m}_t^u/(\sqrt{\widehat{v}_t^u} + \epsilon) + \lambda\theta_{t-1}\big)$, with hats denoting bias correction.
5. Stored moments: $m_t = \beta_1 m_{t-1} + (1-\beta_1)s_t^m$ and $v_t = \beta_2 v_{t-1} + (1-\beta_2)(s_t^v)^{\odot 2}$. The temporary moments are discarded.

With $u_t = s_t^m = s_t^v$, steps 3 and 5 coincide and the update is AdamW.

C.1 Coherent pairs and a bound on the update ratio

A stored pair (s^m, s^v) is *coherent* if $|s_{t,i}^m| \le |s_{t,i}^v|$ for every coordinate i and update t. Coherent pairs include the same sampled, mean-only, sign-mixed, or sign-permuted source in both moments, and mean-only m over the matched v. Mean-only m or v over a sampled partner is not coherent, and neither is sampled or reflected m over the matched v.

Fix a coordinate and start from zero moments. At update t, index the inputs of the temporary moments of Eq. (1) by lag j, with $h_{t,0}^m = h_{t,0}^v = u_t$ and $(h_{t,j}^m, h_{t,j}^v) = (s_{t-j}^m, s_{t-j}^v)$ for $j \ge 1$. Then $m_t^u = (1-\beta_1)\sum_{j<t}\beta_1^j h_{t,j}^m$ and $v_t^u = (1-\beta_2)\sum_{j<t}\beta_2^j (h_{t,j}^v)^2$, and coherence gives $|h_{t,j}^m| \le |h_{t,j}^v|$. Writing $\beta_1^j = (\beta_1^j/\beta_2^{j/2})\,\beta_2^{j/2}$, the Cauchy–Schwarz inequality gives, as in the proof of Lemma 20 in the supplement of Zou et al. (2019),

$$(m_t^u)^2 \le v_t^u \,\frac{(1-\beta_1)^2}{1-\beta_2}\sum_{j=0}^{t-1}\Big(\frac{\beta_1^2}{\beta_2}\Big)^j,$$

and after bias correction, for $\beta_1^2 < \beta_2$,

$$\frac{|\widehat{m}_t^u|}{\sqrt{\widehat{v}_t^u} + \epsilon} \le C_t = \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t}\,\frac{1-\beta_1}{\sqrt{1-\beta_2}}\sqrt{\frac{1-(\beta_1^2/\beta_2)^t}{1-\beta_1^2/\beta_2}}. \tag{7}$$

If $v_t^u = 0$, coherence forces $m_t^u = 0$. Nothing is assumed about the noise, stationarity, or the schedule. Since $C_1 = 1$ and both moments see only u_1 at the first update, every condition reaches the bound there, up to ϵ. The bound C_t rises from 1 toward $\sqrt{19/14} \approx 1.165$ at $(\beta_1, \beta_2) = (0.9, 0.95)$ and toward 2.605 at $(0.5, 0.95)$.

With the matched v, sampled or reflected m obeys a weaker bound, $\sqrt{2}\,C_t$, because $|\bar{g}_i \pm r_i| \le \sqrt{2}\,(\bar{g}_i^2 + r_i^2)^{1/2}$. The language-model comparison of these two at $\beta_1 = 0.5$ is therefore balanced for the bound, but their comparisons with mean-only m, the only coherent pair there, are not. Mean-only m over a sampled v has no bound: if a coordinate's first two sampled gradients are zero while its mean is not, a nonzero numerator meets a zero second moment.

D The long-run gain and its subspace version

D.1 Impulse response under a frozen preconditioner

We assume (A1) a frozen preconditioner $\Lambda = \mathrm{diag}(1/(\sqrt{v} + \epsilon))$ and learning rate η, without bias correction or weight decay, and (A2) no feedback of parameter changes on later gradients. Write $a = 1 - \beta_1$, $u_t = \bar{g}_t + r_t$ for the current source, and $s_t^m = \bar{g}_t + \zeta_t$ for the stored source, with

$\zeta_t = Tr_t$ by default. Unrolling Eq. (1) from $m_0 = 0$, with $s_t^m = 0$ for $t \le 0$, gives the numerator of update t exactly:

$$m_t^u = a\, u_t + a \sum_{\ell \ge 1} \beta_1^\ell\, s_{t-\ell}^m. \tag{8}$$

A residual drawn at update t enters that update with weight aI and update $t+\ell$ with weight $a\beta_1^\ell T$, so it contributes $A_T r$ in total, with $A_T = aI + \beta_1 T$, while the mean receives total weight 1 in every condition. Projecting $A_T r$ on r gives the long-run gain $\kappa = a + \beta_1 q_T$ of Eq. (2). For $T = cI$ it vanishes at $c = -(1-\beta_1)/\beta_1$, which is $-1/9$ at $\beta_1 = 0.9$, -1 (the reflected residual) at 0.5, and $-7/3$ at 0.3.

In this model, flipping the stored copy's sign leaves the stationary per-step covariance of the numerator unchanged. For residuals that are uncorrelated across updates with common covariance Σ, the residual part of a stationary numerator, $z_t = ar_t + a\sum_{\ell \ge 1} \beta_1^\ell T r_{t-\ell}$, has

$$\mathrm{Cov}(z_t) = a^2\Sigma + \frac{a^2\beta_1^2}{1-\beta_1^2}\, T\Sigma T^\top,$$

which is the same for c and $-c$ and for $D_{k/8}$ and $D_{(8-k)/8} = -D_{k/8}$ ($k \ne 4$). A residual and its stored copy never enter the same numerator, so their cross-covariance, the sole term that depends on the sign, appears only at lags $\ell \ge 1$. If residuals are also independent across updates and symmetric under $r \mapsto -r$, the whole law of z_t is equal within each pair, and no single-step account can separate the pairs of Sec. 4. Without symmetry the odd moments differ, so the experiments test accounts that use only the matched second-order quantities, not every single-step description.

Over n_u updates the noise in $\theta_{n_\mathrm{u}} - \theta_0$ has covariance $n_\mathrm{u}\eta^2 \Lambda A_T \Sigma A_T^\top \Lambda$ to leading order (Proposition 1(i)), and we call $P = \mathrm{tr}(A_T \Sigma A_T^\top)/\,\mathrm{tr}\,\Sigma$ the long-run power relative to sampled training. A mask preserves norms, so $P = \kappa^2 + \beta_1^2(1-q^2)$, where $q = \mathrm{tr}(D\Sigma)/\,\mathrm{tr}\,\Sigma$ averages to $1 - 2k/8$ over draws of $D_{k/8}$. At the design overlaps $\pm\frac{1}{2}$, P is 0.91 for $D_{2/8}$ and 0.73 for $D_{6/8}$, of which $\kappa^2 = 0.30$ and 0.12 lie along r, and a fair mask gives 0.82, as much as an independent draw. The sign-mixed residual nevertheless acquired like mean-only m, but the clipped independent draw failed (Sec. 3 and Sec. 4.1). We therefore restrict P to the directions that gradients occupy (Appendix D.2), and Table 7 lists the values.

D.2 Applied noise power in a gradient subspace

For an orthogonal T, the remainder $\beta_1(T - q_T I)r$ is orthogonal to r yet has squared norm $\beta_1^2(1 - q_T^2)\|r\|^2$. Gradients occupy few directions (Gur-Ari et al., 2018), and a remainder outside them may do little. For a subspace S with orthogonal projector Π_S and a stored residual ζ, define the overlap $q_S = \mathbb{E}\langle \Pi_S r, \Pi_S \zeta\rangle / \mathbb{E}\|\Pi_S r\|^2$, the retention $\rho_S = \mathbb{E}\|\Pi_S \zeta\|^2 / \mathbb{E}\|\Pi_S r\|^2$, $\kappa_S = a + \beta_1 q_S$, and the applied noise power in S relative to sampled training,

$$P_S = \mathbb{E}\|\Pi_S(ar + \beta_1 \zeta)\|^2 \big/ \mathbb{E}\|\Pi_S r\|^2, \tag{9}$$

which equals P when S is the whole space.

The proposition below also assumes (A3) residuals that, given the state, have zero mean, are uncorrelated across updates, and have a covariance that changes slowly relative to the memory. Fresh contexts and queries give the zero conditional mean and the lack of correlation, and slow covariance change is an additional approximation. Reading P_S as a predictor of acquisition requires (A4) the same v rule in the compared conditions, true of the stored-m comparisons but not of the batch conditions. It also requires (A5) that applied noise outside S does not affect acquisition. We did not test (A5) for sign-randomized copies, and it fails for fresh noise: in an earlier recipe with 64 contexts per update, global clipping at 1.0, and 32,000 updates, zero-mean Gaussian noise added to the all-query gradient at one eighth, one quarter, or one half of the residual's energy prevented recall when spread over all parameters (0/6 at each dose and at matched per-coordinate intensity, and 0/6 on all but layer norms, embeddings, and readout), slowed it seven- to ninefold on query/key weights alone (6/6 at each dose), and was harmless on all other weights (6/6, also at matched per-coordinate intensity). The noise norms far exceeded the clipping threshold, so these fractions set the noise's share of a clipped update rather than its energy. Such noise raises each coordinate's magnitude and enters the update, which a stored sign mask does not, and it localizes differently: mixing the stored residual's signs on query/key weights alone did not let recall form (Appendix J.3). For SGD, Song et al. (2025) found that updates

Table 7: Overlap, long-run gain, per-step variance, and long-run power of stored first-moment sources under a frozen preconditioner, with sampled v and update. Per-step is the variance of the numerator's residual part per $\operatorname{tr}\Sigma$, and P and P_S are the long-run powers in the full space and in the gradient subspace S (Appendix D.3). Entries marked $\approx$ hold in expectation over random signs or permutations or approximately on frozen states. *At the median overlap and retention over frozen states (Appendix D.3), for the sign-mixed copy on its own run's states.

Stored first-moment source	q	κ	Per-step	P	$\sqrt{P_S}$
$\beta_1 = 0.9$					
sampled, $\bar{g} + r$	1	1	0.0526	1	1
reflected, $\bar{g} - r$	-1	-0.8	0.0526	0.64	0.8
scaled copy $c = \frac{1}{2}$	0.5	0.55	0.0207	0.30	0.55
scaled copy $c = -\frac{1}{2}$	-0.5	-0.35	0.0207	0.12	0.35
mask $D_{2/8}$	0.5	0.55	0.0526	0.91	≈ 0.55
mask $D_{6/8}$	-0.5	-0.35	0.0526	0.73	≈ 0.35
mean-only, $\bar{g}$	0	0.1	0.0100	0.01	0.1
sign-mixed, $\bar{g} + Dr$	≈ 0	≈ 0.1	0.0526	≈ 0.82	0.103^*
sign-permuted, $\bar{g} + Rr$	≈ 0	≈ 0.1	0.0526	≈ 0.82	0.107^*
independent draw, $\bar{g} + r'$	0	0.1	0.0526	0.82	0.906
partial average, $n' = 2$	3/8	0.438	0.0260	0.381	0.617
partial average, $n' = 3$	1/6	0.25	0.0171	0.175	0.418
partial average, $n' = 4$	1/16	0.156	0.0127	0.072	0.268
$\beta_1 = 0.5$					
sampled, $\bar{g} + r$	1	1	1/3	1	1
mean-only, $\bar{g}$	0	0.5	1/4	0.25	0.5
reflected, $\bar{g} - r$	-1	0	1/3	0	0

projected onto the dominant Hessian subspace did not lower the training loss further, a caution for (A5).

Proposition 1. *Assume (A1)–(A3), and let ζ have zero mean, fixed joint second moments with r, and no correlation with the residuals of other updates (for example $\zeta = Tr$, an independent query draw, or a partial average). Then the following hold for every subspace S. (i) The noise part of $\theta_{n_u} - \theta_0$ has covariance $n_u \eta^2 \Lambda \operatorname{Cov}(ar + \beta_1\zeta)\Lambda + O(1)$ as $n_u \to \infty$. (ii) $P_S = \kappa_S^2 + \beta_1^2(\rho_S - q_S^2)$ with $\rho_S \geq q_S^2$, so $\sqrt{P_S} \geq |\kappa_S|$, with equality exactly when $\Pi_S(\zeta - q_S r) = 0$ almost surely. (iii) In units of $\mathbb{E}\|\Pi_S r\|^2$, the stationary numerator residual in S has variance $\gamma(0) = a^2\big(1 + \beta_1^2\rho_S/(1-\beta_1^2)\big)$, which does not depend on q_S, and lag-ℓ covariance $\gamma(\ell) = a^2\beta_1^\ell\big(q_S + \beta_1^2\rho_S/(1-\beta_1^2)\big)$, with $\gamma(0) + 2\sum_{\ell\geq 1}\gamma(\ell) = P_S$. The same holds with Π_S replaced by any fixed linear map.*

Proof. (i) By Eq. (8) the noise part is $-\eta\Lambda\sum_{t=1}^{n_u}\big(ar_t + \beta_1(1-\beta_1^{n_u - t})\zeta_t\big)$, whose uncorrelated terms have covariances within $O(\beta_1^{n_u - t})$ of $\operatorname{Cov}(ar + \beta_1\zeta)$, and these deviations sum to $O(1)$. (ii) Expanding the square gives $a^2 + 2a\beta_1 q_S + \beta_1^2\rho_S$, and completing the square in q_S gives the stated form. With $\chi = \Pi_S(\zeta - q_S r)$, the definition of q_S gives $\mathbb{E}\langle\Pi_S r, \chi\rangle = 0$, so $\rho_S - q_S^2 = \mathbb{E}\|\chi\|^2/\mathbb{E}\|\Pi_S r\|^2 \geq 0$. (iii) With ζ in place of Tr in z_t, the numerators z_t and $z_{t+\ell}$, $\ell \geq 1$, share draws only through r_t paired with ζ_t and through each ζ_s with $s < t$. Summing the geometric series gives $\gamma(\ell)$, and at $\ell = 0$ no cross term appears. The identity $a^2(1 + 2\beta_1/a)\beta_1^2/(1-\beta_1^2) = \beta_1^2$ gives the sum over lags. Only linearity of Π_S is used. □

Values. A scaled copy has $q_S = c$ and $\rho_S = c^2$, so $P_S = \kappa^2$ in every subspace. An independent draw r' has $\rho_S = 1$ and $q_S = 0$ exactly, so $P_S = a^2 + \beta_1^2 = 0.82$. The overlap is zero despite a $1/5$ chance of repeating the drawn query, because a context's centered per-query residuals average to zero over the draw. With an all-query update and a sampled stored copy, $P_S = \beta_1^2 = 0.81$.

For fair masks, $\mathbb{E}_D\|\Pi_S Dr\|^2 = \sum_i r_i^2(\Pi_S)_{ii}$. This averages to $(\dim S/\dim\theta)\|r\|^2$ over random subspaces but equals $\|\Pi_S r\|^2$ for a subspace spanned by coordinate axes, so a small retention must be measured (Appendix D.3). Since $D_{k/8} = (1 - 2k/8)I + \Delta$, where Δ is diagonal with independent

zero-mean entries of variance $k(8-k)/16$, averaging over masks gives $P_S = \kappa^2 + \frac{k(8-k)}{16}\beta_1^2\rho_D$, with ρ_D the fair-mask retention.

Limits of the model. Warmup, bias correction, and per-source clipping violate (A1), and at $\beta_2 = 0.95$ the preconditioner averages over about 20 updates, only twice the first moment's window, so v adapts within the window over which κ is summed. Reflection was not run at $\beta_2 = 0.999$, where (A1) is most defensible, and there the complementary masks both acquired in every run and differed only in speed (Appendix I.2). (A2) removes the feedback through which the compared conditions diverge, and the model, which omits v, cannot explain failures that depend on v (Secs. 5 and 7).

D.3 Retention on frozen states and a post hoc placement

We reused the frozen states of the 48 completed runs of the unclipped scaled-copy experiment (Appendix E.2) at updates 500, 1,000, 1,500, 3,000, and 8,000. At each state we estimated the top-8 residual subspace S from 16 query draws. On 16 held-out draws, we measured the retention of the run's fixed mask D, a signed permutation R, and the seven masks $D_{k/8}$, and compared their q_T with null bands from random masks and permutations. Three descriptive checks, with thresholds set after viewing about seven states (post hoc), were met in all 240 states. The median $|q_T|$ was 0.0063 for D and 0.0016 for R, no mask's median q departed from $1 - 2k/8$ by more than 0.054, and S always held at least 8.5% of the held-out residual energy. Retention was at most 0.053 for D and 0.045 for R, with medians of 0.004 and 0.002. At the median in-subspace overlap and retention, $\sqrt{P_S}$ is 0.103 for D on the sign-mixed run's own states (retention 0.0011, 0.114 over all states) and 0.107 for R, against 0.906 for an independent draw. For the masks, with retention from these states rather than mask runs, $\sqrt{P_S}$ lies within 0.016 of $|\kappa|$ at median and within 0.131 at maximum retention. We did not measure the random redraw, whose retention equals that of R in expectation.

After their outcomes were known, we ordered the 39 conditions of the routing, residual-control, scaled-copy, mask, batch, and β_1 experiments by $\sqrt{P_S}$, with a batch of B contexts at $P_S = 256/B$. A threshold between 0.447 (1,280 contexts, 8/8) and 0.5 (mean-only m at $\beta_1 = 0.5$ in the decay study, 2/8) separates majority-acquiring from majority-failing conditions, both among these and among the 13 later conditions of Appendices D.4 and G. Kendall's τ_b with the fraction acquired is -0.81 over the 39 and -0.77 over all 52. This post hoc placement is an exploratory summary, not a test. Counting the growth of Λ with batch size moves the two larger batches across the threshold. Even $c = -7/3$ at $\beta_1 = 0.3$, which has $\sqrt{P_S} = 0$, acquired in only 5/8 runs. Outside the reference recipe, no fixed threshold separates the outcomes (Appendices E.3, E.4, and I).

D.4 Partial averaging in the stored first moment

Partial averaging stores in m the average of n' of a context's five queries, the drawn one included, with the update and v sampled. It interpolates between sampled training ($n' = 1$) and mean-only m ($n' = 5$), and its stored residual has $q_S = \rho_S = q = (5 - n')/(4n')$ in every subspace (Appendix C). On eight shared seeds in the unclipped scaled-copy recipe, sampled training acquired in 0/8 runs, $n' = 2$ in 1/8, $n' = 3$ and $n' = 4$ in 8/8 each, and mean-only m in 8/8. The AUC was larger at $n' = 4$ than at $n' = 2$ in 8/8 pairs ($p = 0.0039$), and the realized overlaps, 0.376, 0.169, and 0.062, matched the design. Because κ, P_S, and the per-step variance $0.01 + 0.0426q$ all rise with q, the study cannot separate their effects.

E Where the residual acts: recall experiments

These experiments use the unclipped reference recipe of Appendix B ($\beta_1 = 0.9$, sampled update), the sources of Appendix C, and eight seeds per condition, and tests are exact one-sided paired sign tests on the AUC, unless stated otherwise.

E.1 Which moment carries the residual

With independently drawn queries and per-source clipping, mean-only m acquired in 8/8 runs and mean-only v in 1/8 (Table 1(a)). Mean-only m acquired in only 2/8 when the minibatch shared one query index, and averaged moments with a per-tensor scalar v in 5/8.

Table 8: Acquisition by stored source in three experiments with a sampled update, unclipped. Acq.: runs of eight that acquire by update 8,000. Moments not named are sampled, and κ is an expectation for random signs or directions. Bold marks the two single-moment allocations.

(a) Allocation of $\bar{g} + Rr$

Source	κ	Acq.
sampled m, v	1	0/8
sign-permuted v	1	**0/8**
sign-permuted m	0.1	**8/8**
sign-permuted m, v	0.1	8/8
mean-only m, v	0.1	8/8

(b) Residual controls in m

Source	κ	Acq.
sampled	1	0/8
reflected	-0.8	0/8
sign-mixed	0.1	8/8
sign-permuted	0.1	8/8
random redraw	0.1	8/8
mean-only	0.1	8/8

(c) Scaled copies $\bar{g} + cr$ in m

Source	κ	Acq.
$c = 1$ (sampled)	1	0/8
$c = -1$ (reflected)	-0.8	0/8
$c = \frac{1}{2}$	0.55	1/8
$c = -\frac{1}{2}$	-0.35	7/8
$c = 0$ (mean-only)	0.1	8/8
sign-mixed	0.1	8/8

Crossed update. An all-query update over sampled stored moments acquired in 0/8 runs with per-source clipping (Table 1(b)), and a sampled update over mean-only moments in 8/8 (Table 1(a)). The sampled/sampled and all-query/mean-only rows of Table 1 come from another experiment with other seeds. A separate unclipped experiment used the centered update of Appendix C, whose increment has the all-query increment as its conditional mean and twice the sampled increment's covariance. This update acquired in 8/8 runs over mean-only moments and in 0/8 over sampled ones. In both experiments, acquisition tracked the stored moments' source rather than the current update's (descriptive).

Sign-permuted residuals. The allocation experiment stored $\bar{g} + Rr$, where $R = DQ$ applies fixed random signs D and a uniform permutation Q within each tensor (Table 8(a)). Allocated to m, it exceeded sampled training in AUC in 8/8 pairs, as in a clipped replication on new seeds; in v alone, it never acquired. Under a centered update, the signs alone ($\bar{g} + Dr$) and the permutation alone ($\bar{g} + Qr$) each improved on sampled moments in 8/8 pairs (Holm $p = 0.012$), and a two-sided test between them did not reject ($p = 0.727$).

Earlier experiments. In three earlier experiments with a centered update, averaging only the stored residual's query-projection part let 1/8 runs acquire. Neither of its tests rejected (4/8 positive pairs in AUC against sampled sources, 0/8 against the sign-permuted complement in a query-transfer readout; Holm $p = 1$ each). Sign-permuting, by contrast, gave a larger AUC than leaving the residual unchanged in 8/8 pairs, for the whole residual and for the part that bypasses the second-block query projection, with the rest averaged. The latter was also ahead, in 8/8 pairs, of an unchanged run whose adaptive step was rescaled at every update to the norm of a run with both moments sign-permuted.

E.2 Residual controls and scaled copies

In the residual-control experiment, 8/8 pairs favored sign-mixed m over sampled training in AUC (Table 8(b); Holm $p = 0.016$ over four tests, with reflection and the learning-rate tests below). Reflected m was favored over sampled training in only 4/8 pairs ($p = 0.637$). Here, as in the scaled-copy and clipped experiments, the batch study, and the decay study at $\beta_1 = 0.9$ and 0.95, mean-only m acquired in 8/8 runs. Reflection keeps the stored residual's covariance, norm, and coordinate magnitudes, but it failed where the sign-mixed, sign-permuted, and redrawn copies acquired. In Table 7, reflection has the same per-step variance as the sign-mixed and sign-permuted copies and a lower full-space power (0.64 against about 0.82), so neither statistic separates them in the observed direction. The long-run gain ($\kappa = -0.8$ against about 0.1) and its subspace version ($\sqrt{P_S} = 0.8$ against 0.103 and 0.107 for the sign-mixed and sign-permuted copies) do separate them in that direction (post hoc).

Independent draws. Two controls under per-source clipping gave m and v query draws of their own on shared contexts, under an all-query or a separately drawn sampled update. Neither acquired (0/8 runs each), so even a stored residual drawn independently of the current one prevented acquisition within the horizon. Because their stored copy re-applies its own residual r' instead of r, it has $\kappa = 0.1$ (Sec. 2.4). The random redraw, also at 0.1, acquired. The two differ in clipping, in the draw that fed v, and in subspace power ($\sqrt{P_S}$). For v, the clipped controls used a query draw of its own, independent

Table 9: Effect of contexts per update (a) and first-moment decay (b) on recall acquisition, unclipped. Acq. as in Table 8. –: not assigned.

(a) Contexts per update

Condition	Contexts	Acq.
sampled	256	0/8
sampled	640	1/8
sampled	1,280	8/8
sampled	2,560	8/8
mean-only m	256	8/8

(b) First-moment decay, $\beta_2 = 0.95$

Condition	β_1	κ	Acq.
mean-only m	0.3	0.7	1/8
all-query	0.3	–	8/8
sign-mixed m	0.3	0.7	0/8
mean-only m	0.5	0.5	2/8
mean-only m	0.9	0.1	8/8
mean-only m	0.95	0.05	8/8
sampled	0.95	1	0/8

of those of the update and m, and the redraw used the sampled square $g_J^{\odot 2}$. The copy r' occupies the same few gradient directions as r, while a redrawn copy (unmeasured) is expected to keep as little energy there as a sign-permuted one (Appendices D.2 and D.3).

Unclipped independent draw. The clipped control above has $\kappa = 0.1$, like mean-only m, but carries the full covariance of r and has $\sqrt{P_S} \approx 0.91$, so the two accounts predict opposite outcomes. We reran the control without clipping, in the recipe of the mask study (Appendix F), on eight new seeds. Each seed covered four conditions: the independent draw with a third draw's square in v, the independent draw with the sampled square in v, sampled training, and mean-only m. An exact one-sided paired sign test on the AUC of the independent draw against sampled training did not reject (4 of 8 positive pairs, $p = 0.64$). Both independent-draw conditions acquired in 0/8 runs, like sampled training (0/8), but mean-only m acquired in 8/8 (KM median 800). The independent draw had a smaller AUC than mean-only m in 8/8 pairs (mean difference 0.72). Of the two accounts, only the subspace power predicted this outcome.

Scaled copies. The half copies of the scaled-copy experiment (Table 8(c)) share the conditional mean, the covariance $\Sigma/4$, and every coordinate magnitude, but have $\kappa = 0.55$ and -0.35. In 7/8 pairs, $c = -\frac{1}{2}$ had a larger AUC than $c = \frac{1}{2}$ (raw $p = 9/256$), but the test did not reject after Holm adjustment over four tests, one of which, the ordering test of Appendix J.2, had $p = 1$ (adjusted $p = 0.070$). Sec. 4.1 repeats it on new seeds.

Step size. Although a mean-only m reduced the median update norm $\eta_t \|d_t\|$ over updates 1–1,000 from 0.039 to 0.016, the two tested learning-rate changes did not reproduce its effect. Sampled training failed at half and at twice the reference learning rate (0/8 runs), and sign-permuted m acquired in 8/8, with the larger AUC in every pair (Holm $p = 0.016$ each). These controls rule out those two changes but leave other dependences on update magnitude open.

E.3 Contexts per update

Because a mean-only m uses all five queries of each context, sampled training at 1,280 contexts per update sees as many (context, query) pairs as mean-only m at 256, spread over the update and both moments. A larger batch, however, also changes the update, v, and the contexts seen, and it shares more contexts with the evaluation set (about 36% at 2,560 contexts, against 4% at 256). The batch study therefore does not isolate sample count, and we assign the larger batches no long-run gain. Under the study's fixed learning rate and warmup, sampled training with the matched pair count acquired in 8/8 runs (KM median 800, as for mean-only m at 256; Table 9(a)). The test for mean-only m at 256 contexts over sampled training at 1,280 did not reject (3/8 pairs, $p = 0.855$).

E.4 First-moment decay

With a mean-only m, the update still carries the current residual at weight $1 - \beta_1$, a long-run gain of 0.1 at $\beta_1 = 0.9$ but 0.7 at 0.3 (Table 9(b)). At $\beta_1 = 0.3$, the informative comparison is between mean-only m and all-query training, the control for the decay itself. The control acquired in 8/8 runs

Table 10: Recall outcomes and overlaps in the sign-mask study (unclipped, $\beta_1 = 0.9$, 8,000 updates). Acquisitions are counted on the 16 shared seeds and on all 24 (–: not run on the extra 8). The Kaplan–Meier (KM) median, in updates, and the mean AUC use every seed of a condition.

	q			Acquired			
Source	Design	Realized	κ	16 seeds	24 seeds	KM median	Mean AUC
$D_{1/8}$	0.75	0.743	0.775	0/16	–	>8,000	0.200
$D_{2/8}$	0.5	0.498	0.55	3/16	5/24	>8,000	0.343
$D_{3/8}$	0.25	0.248	0.325	16/16	–	900	0.911
$D_{4/8}$	0	−0.003	0.1	16/16	–	800	0.922
$D_{5/8}$	−0.25	−0.249	−0.125	16/16	–	900	0.917
$D_{6/8}$	−0.5	−0.498	−0.35	12/16	18/24	1,500	0.714
$D_{7/8}$	−0.75	−0.744	−0.575	0/16	–	>8,000	0.199
$c = \frac{1}{2}$	0.5	0.500	0.55	3/16	7/24	>8,000	0.374
$c = -\frac{1}{2}$	−0.5	−0.500	−0.35	12/16	17/24	1,250	0.692

and mean-only m in 1/8. This result favors the long-run gain over an account in which only the noise stored in m matters.

Like the scaled copies that failed, with $\kappa = 0.55$ (1/8) and 1 (0/8), mean-only m at $\beta_1 = 0.3$ has positive κ, so the study does not test symmetry in the sign of κ. Varying β_1 also changes the lag of the mean gradient, which the all-query control covers only for noise-free updates.

E.5 A LONGER HORIZON

The failure of sampled training is at least a long delay. Over 32,000 updates, sampled training acquired in only 2/8 runs, at updates 15,750 and 24,250, and all-query training in 8/8 by update 750 (KM median 500). The distribution-free 95% limits of the two medians, at least 15,750 and at most 750, put their ratio above 20.

F THE SIGN-MASK STUDY

The sign-mask study of Sec. 4.1 tested, on new seeds, the long-run-gain account that we had drawn post hoc from the completed scaled-copy experiment (Appendix E.2).

Design. Nine conditions stored $\bar{g} + D_{k/8}r$ for $k = 1, \ldots, 7$ or the half copies $\bar{g} \pm \frac{1}{2}r$ in m. Each seed assigned every coordinate to one of eight equally likely bins. For $k \leq 4$, $D_{k/8}$ is negative on the first k bins, and for $k \geq 5$ it equals $-D_{(8-k)/8}$, so $D_{6/8} = -D_{2/8}$ exactly. The design overlap is $\mathbb{E}q = 1 - 2k/8$, giving $\kappa = 0.1 + 0.9\,\mathbb{E}q$, and the realized q is the median over seeds of each run's mean over updates 1–2,000. Each realized q lay within 0.01 of its design value, and measuring it in the preconditioned metric changed it by at most 0.006. Training used the unclipped scaled-copy recipe for 8,000 updates, sampling the current update and v (Sec. 2.1 and Appendix B). Sixteen seeds ran all nine conditions, and eight more ran $D_{2/8}$, $D_{6/8}$, and $c = \pm\frac{1}{2}$ (Table 10).

Tests. Two exact one-sided paired sign tests on the AUC over 24 seeds ran in a fixed sequence, with ties counted as non-positive. On the 16 shared seeds alone, each had 14 positive pairs ($p = 0.0021$). Both rejected, with $D_{6/8}$ ahead of $D_{2/8}$ in 21 of 24 pairs ($p = 1.4 \times 10^{-4}$) and $c = -\frac{1}{2}$ ahead of $c = \frac{1}{2}$ in 20 of 24 ($p = 7.7 \times 10^{-4}$). The second test replicated on fresh seeds the scaled-copy contrast, which had not survived Holm adjustment.

What tracked acquisition (descriptive). Acquisition fell as $|\kappa|$ grew, whatever its sign (Kendall's τ_b between AUC and $|\kappa|$ at the realized q, -0.55 over all runs). Even $D_{7/8}$, which cancels more than $D_{6/8}$, never acquired. A mask stores each coordinate at full size and so has 2.55 times the per-step numerator variance (Appendix D), yet masks and half copies with equal κ had identical counts on the shared seeds. The 95% bootstrap intervals of their paired AUC difference over 24 seeds, mask minus half copy, were -0.090 to 0.000 at $\kappa = 0.55$ and -0.005 to 0.073 at $\kappa = -0.35$, inside an

equivalence margin of ± 0.15. On the eight extra seeds alone, the half copies differed less than the masks, with 5/8 acquisitions for $c = -\frac{1}{2}$ against 4/8 for $c = \frac{1}{2}$, and 6/8 for $D_{6/8}$ against 2/8 for $D_{2/8}$. Complementary masks took early steps of similar size (post hoc, median update norm $\eta_t \|d_t\|$ over updates 1–1,000, 0.045 for $D_{6/8}$ and 0.043 for $D_{2/8}$).

Competing accounts. At the reference shape, the results contradict an account in which full-magnitude residuals with mixed signs suffice. The mixed-sign account predicted a null first test, at least 12/16 acquisitions for both masks, and $D_{2/8}$ above $c = \frac{1}{2}$ by more than 0.15 in AUC. The per-step numerator variance and the marginal statistics $D\Sigma D$ and $|r_i|$ are shared by both arms of each test, so they cannot account for the contrast. Only the $\kappa = -0.35$ versus $+0.55$ contrast separates $|\kappa|$ from $|q|$. Nor do the results separate κ from κ^2, from its subspace version (Appendix D.3), from windowed variants, or from other functions of the signed cross-moment $\operatorname{tr}(D\Sigma)/\operatorname{tr}\Sigma = \pm\frac{1}{2}$.

F.1 Coupling of m and v through one draw

In the mask study, v stored the sampled square $(\bar{g} + r)^{\odot 2}$ of the same draw, and the masks leave free its joint statistics with the stored m source, the covariance of Eq. (5). In both tests this cross-moment flips sign in every coordinate, as q does, and through $1/\sqrt{v}$ it gives the preconditioned stored term a mean bias that also flips. Part of the mask contrast could therefore run through the pairing of m and v rather than through the stored copy's direction.

Masks with matched v. To remove the mean part, $2\bar{g}_i \mathbb{E}[r_i^2]$, a follow-up study stored the matched source $\sqrt{\bar{g}^2 + r^2}$ of Sec. 2.3 in v for every condition, keeping the recipe, masks, and horizon of the mask study. Because $(d_i r_i)^2 = r_i^2$, the v input rule gives complementary masks the same input at a common state, although realized v trajectories still differ once the runs diverge. At a common state the stored m sources then differ in the sign pattern of the residual, which also flips the remaining odd-moment coupling $d_i \mathbb{E}[r_i^3]$, zero when the conditional third moment is zero.

Sixteen new seeds ran $D_{6/8}$ and $D_{2/8}$, and eight of them also ran sampled m with the same v as a descriptive reference. All arms of a seed shared the initialization and data streams. An exact one-sided paired sign test on the AUC rejected, with $D_{6/8}$ ahead in 16 of 16 pairs ($p = 1.5 \times 10^{-5}$; mean paired difference 0.53). $D_{6/8}$ also acquired more often, in 13/16 runs against 2/16 (KM medians 1,500 and beyond 8,000). Sampled m with matched v acquired in 0/8 and did not differ from $D_{2/8}$ (4 of 8 pairs positive).

On the logged coordinates, relative to $\sum_i |\bar{g}_i| r_i^2$, the removed mean part had size 0.98 to 1.02 in every window, and the remaining odd-moment part averaged at most 0.022 in absolute value for either mask. The early update norms of $D_{6/8}$ (updates 1–1,000) were 5% above those of $D_{2/8}$ (median per-seed ratio 1.05). The masks' contrast therefore does not require the mean part of the m–v coupling. The remaining odd-moment part is small in these runs, but its effect was not tested separately.

G Stored copies with zero long-run gain

For a mean-only m, lowering β_1 raises the long-run gain and the per-step noise together, so the decay study (Appendix E.4) cannot separate them. At $\beta_1 = 0.5$ the reflected residual $\bar{g} - r$ has zero long-run gain (Appendix D.1), since its stored copy cancels the current residual in the long-run sum.

Design. Apart from β_1, all runs followed the unclipped reference recipe (Appendix B), with $\beta_2 = 0.95$. The eight conditions of Table 11(a) shared eight seeds. All-query training carries no residual, because it uses $\bar{g}$ in the update and both moments. Every other condition had a sampled current update and a sampled v, so within one decay only the rule for the m source differed. Changing that rule also changes how m pairs with v (below).

The reflected and sampled residuals at $\beta_1 = 0.5$ share their conditional covariance, coordinate magnitudes, and v source. They differ in the sign of the stored copy relative to the current residual, the odd moments if r is skewed, and the pairing of m with v, since reflection stores $\bar{g} - r$ in m but $(\bar{g} + r)^{\odot 2}$ in v. Two sign tests at $\beta_1 = 0.5$ compared the AUC of the reflected residual in a fixed sequence, first against sampled training and then, only if the first rejected, against mean-only m.

Table 11: Zero long-run gain on eight shared seeds, unclipped: (a) 8,000 updates, (b) 16,000 updates at $\beta_1 = 0.5$. Per-step is the stationary covariance of the numerator's residual part under a frozen preconditioner (Sec. 2.4), in units of Σ, and KM (Kaplan–Meier median) exceeds the horizon if fewer than half acquired. Acq.: runs of eight acquiring. Table 9(b) is a separate experiment on other seeds, where mean-only m acquired in 2/8 runs at $\beta_1 = 0.5$ and 1/8 at 0.3.

Stored m	κ	Per-step	Acq.	KM	AUC
(a) 8,000 updates; $\beta_1 = 0.5$					
reflected, $\bar{g} - r$	0	$1/3$	8/8	7,000	0.381
sampled, $\bar{g} + r$	1	$1/3$	0/8	>8,000	0.198
mean-only, $\bar{g}$	0.5	$1/4$	1/8	>8,000	0.203
all-query, $\bar{g}$	–	–	8/8	500	0.952
$\beta_1 = 0.3$					
scaled, $\bar{g} - \frac{7}{3}r$	0	0.754	5/8	8,000	0.254
mean-only, $\bar{g}$	0.7	0.49	0/8	>8,000	0.198
$\beta_1 = 0.9$					
scaled, $\bar{g} - \frac{1}{9}r$	0	0.0105	8/8	900	0.918
reflected, $\bar{g} - r$	-0.8	0.0526	0/8	>8,000	0.199
(b) 16,000 updates; $\beta_1 = 0.5$					
reflected, early-norm-matched	0	$1/3$	8/8	6,750	0.690
reflected, continued	0	$1/3$	8/8	7,000	0.689
mean-only, continued	0.5	$1/4$	7/8	10,250	0.471
sampled, continued	1	$1/3$	0/8	>16,000	0.201

Results. All-query training acquired in 8/8 runs. Against sampled training (0/8 runs) and mean-only m (1/8), the reflected residual acquired in 8/8 runs and had the larger AUC in 8/8 pairs of each comparison ($p = 0.0039$). Seven of its runs first crossed late, between updates 6,750 and 7,500 (KM median 7,000). Over updates 1–1,000, its direction norm $\|d_t\|$ had a median of 647 (the median over seeds of each run's median), 19% above the 544 of sampled training. At $\beta_1 = 0.9$, the zero-gain copy $c = -1/9$ acquired in 8/8 runs and reflected m in 0/8. With $\beta_1 = 0.3$, the zero-gain copy $c = -7/3$ acquired in 5/8 runs, all at update 7,250 or later, and mean-only m in 0/8.

G.1 Continued runs and a matched step size

A follow-up on the same seeds, designed after these results, asked whether the failures at $\beta_1 = 0.5$ were only delays and whether the larger early steps explained the reflected residual's lead. It continued the reflected, mean-only, and sampled runs to update 16,000. From update 0 it trained an early-norm-matched reflected residual, whose learning rate was scaled by $544/647 \approx 0.84$, warmup and decoupled weight decay included, so the per-step decay $\eta_t\lambda$ shrank by the same factor. Three Holm-adjusted sign tests on the AUC over updates 0–16,000 compared the continued reflected residual with mean-only m and the early-norm-matched one with sampled training and with mean-only m. The early step sizes matched. In units of the reference learning-rate schedule, the matched residual's median update norm over updates 1–1,000 was 1.003 times sampled training's (per-seed ratios 0.999 to 1.008).

All three tests rejected, each with 8/8 positive pairs (Holm $p = 0.012$; Table 11(b)). The matched residual acquired in 8/8 runs, against none for sampled training. Mean-only m acquired in 7/8 runs (KM median 10,250); its earlier failure was thus a delay. It still trailed the reflected residual. Matching the early step size changed little (Table 11(b)). Against sampled training this match held only early. Over updates 1,001–6,000 the matched residual's steps were 0.92 to 0.94 times those of sampled training, which makes its test against sampled training conservative if larger steps help. Post hoc, mean-only m's median update norm over updates 1–1,000 was 546 in units of the reference schedule, against 544 for sampled training and 546 for the matched residual. Over updates 1,001–3,000 and 3,001–6,000 the matched residual's norms were 1.00 to 1.03 times mean-only m's in the seven seeds that had not yet acquired (the eighth acquired at update 1,250), so against mean-only m the step sizes stayed matched through update 6,000. The full update sequence was not matched.

Per-step variance at zero long-run gain. Zero long-run gain did not determine when runs acquired. The three copies with $\kappa = 0$ acquired at KM medians from 900 to 8,000 updates, later with more per-step variance (post hoc, Table 11(a)). Three confounds leave this a hypothesis. First, the decays differ, and so does how m smooths and lags the mean signal. Second, the copy $c = -7/3$ stores 2.33 times the sampled residual over a sampled v, and its stored pair left the range that a common source allows (Appendix C.1) in essentially every update. Third, early update norms $\eta_t \|d_t\|$ (learning rate times Adam direction) over updates 1–1,000 grew with per-step variance (0.018, 0.121, and 0.179).

G.2 Reflection with a matched second moment

In the studies above, v stored the sampled square, so the reflected and sampled first moments also differed in how m pairs with v (Appendix F.1). A later study gave every condition (reflected, sampled, and mean-only m) the matched source $\sqrt{\bar{g}^2 + r^2}$ in v, whose input is then identical across conditions at a common state. The runs used $\beta_1 = 0.5$ and the recipe of this appendix for 12,000 updates on eight new seeds, with each seed's conditions sharing the initialization and data streams. The reflected condition again used $0.84\times$ the learning rate, which brought its early update norms to sampled training's (median per-seed ratio 0.997 over updates 1–1,000, within 0.7% in every seed).

Exact one-sided paired sign tests on the AUC compared, in a fixed order, the reflected with the sampled first moment and then, only if that test rejected, with mean-only m. The reflected residual had the larger AUC in 8/8 pairs ($p = 0.0039$, mean difference 0.52) and acquired in 8/8 runs, all by update 8,000 (KM median 4,500). None of the sampled-m runs acquired. The advantage of reflection at $\beta_1 = 0.5$ therefore does not depend on the sign-flipping part of the m–v pairing that the sampled square carries.

Against mean-only m with the same matched v, the reflected AUC was also larger in 8/8 pairs ($p = 0.0039$, mean difference 0.34), in the order of both κ and $\sqrt{P_S}$ (0, 0.5, and 1 for reflected, mean-only, and sampled m, Table 7). Both conditions mostly acquired, the reflected residual in 8/8 runs and mean-only m in 6/8 by update 12,000 (2/8 by 8,000, KM median 9,250), so this is a difference in speed, not an established difference in how many runs acquire. Against mean-only m, the reflected update norms were 0.99 to 1.00 times over updates 1–1,000 and 1.00 to 1.02 times in later windows before acquisition (post hoc), and against sampled training only the early step sizes were matched. Once the reflected runs had acquired, their update norms fell far below sampled training's. The matched v leaves the odd-moment part of the pairing in place.

H Language-model studies

H.1 Recipe

The pre-layer-normalized decoder has four blocks of width 256, eight heads, an MLP of width 1,024, and a 512-token context (29,046,272 parameters, of which 25,755,648 are in the untied embedding and output matrices and 3,159,040 in the four blocks). It trains on 227,963 non-overlapping 512-token sequences of WikiText-103 (Merity et al., 2016), tokenized with the Pythia tokenizer (the GPT-NeoX-20B byte-pair encoding, 50,277 tokens padded to a vocabulary of 50,304). Batches hold 64 sequences, and AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$ unless stated, $\epsilon = 10^{-8}$, weight decay 0.01, and an 800-update warmup to a constant learning rate of 3×10^{-4}. Loss is measured on 64 held-out 512-token windows (the diagnostic set) and on 1,024 (the primary set).

The sampled gradient g_J averages the next-token losses at 64 of each sequence's 511 supervised positions, drawn without replacement, and the all-position gradient $\bar{g}$ averages all 511, so $r = g_J - \bar{g}$ has conditional mean zero. The common gradient scale $\alpha(\bar{g})$ of Eq. (6), written α, keeps this exactly, while per-source clipping to global norm 1.0 keeps it only before clipping. Within a study, conditions share eight seeds, and unless a fixed testing sequence is stated (Appendix H.7), each study's two tests are exact one-sided paired sign tests with Holm adjustment, under which 8/8 pairs give $p = 0.0078$.

H.2 The clipped factorial and attention onset

The first study ran 150,000 updates (about 42 passes) with per-source clipping (Appendix C). It crossed the source of the current update, all-position or sampled, with that of both stored moments.

Table 12: Held-out negative log-likelihood (NLL, nats) on the primary set at update 10,000, with current update αg_J (α: the common gradient scale of Eq. (6)). NLL is the median of eight seeds, and Δ is the median paired difference from the panel's first row, of the same sign in 8/8 seeds. Bold marks the condition that each study's two tests compared with others, which in (a) is not the lowest-loss row.

Condition	Stored m	Stored v	NLL	Δ
(a) $\beta_1 = 0.9$				
sampled	αg_J	αg_J	5.166	–
both averaged	$\alpha\bar{g}$	$\alpha\bar{g}$	4.605	-0.563
mean-only m	$\alpha\bar{g}$	αg_J	5.461	$+0.299$
mean-only v	αg_J	$\alpha\bar{g}$	4.783	-0.384
matched v	$\alpha\bar{g}$	$\alpha\sqrt{\bar{g}^2 + r^2}$	**4.889**	-0.277
shared signs	$\alpha(\bar{g} + Dr)$	$\alpha(\bar{g} + Dr)$	5.142	-0.024
(b) $\beta_1 = 0.5$, *new seeds*				
sampled m	αg_J	matched	5.096	–
mean-only m	$\alpha\bar{g}$	matched	4.981	-0.117
reflected	$\alpha(\bar{g} - r)$	matched	**4.910**	-0.185

Further conditions with a sampled update made only one moment all-position. This study's loss comparisons are descriptive. For sampled gradients, clipping was almost always active (median fraction of clipped updates per evaluation interval 0.98 at update 10,000 and 1.000 at 150,000, against 0.40 and 0.64 for all-position gradients).

At update 150,000, sampling the stored moments raised the held-out loss on the diagnostic set by a median paired 0.469 nats with an all-position update and 0.519 with a sampled one. Sampling the current update raised it less (0.023 and 0.087). Every paired difference was positive. An all-position m over a sampled v, with a sampled update, ended below sampled moments in 7/8 seeds (post hoc). Yet it spiked 146 times in its eight runs (consecutive evaluations more than 0.5 nats above the running minimum), and no other condition spiked.

The prefix-matching score (Olsson et al., 2022) is the largest mean attention of any head from the second copy of a repeated random sequence to the token after the earlier match. Onset is the first of four consecutive evaluations, 500 updates apart, at which this score is at least 0.10. In the factorial, changing only m did not reliably delay onset. Sampling m under a sampled update, over an all-position or a sampled v, delayed onset in only 4/8 and 6/8 pairs (median ratios 0.98 and 1.17). Sampling both stored moments did delay onset, in 8/8 pairs under an all-position update and 7/8 under a sampled one (median paired ratios 2.03 and 1.93, non-crossers counted at 150,000), while the update's source left the Kaplan–Meier onset in place (45,500 against 45,500 updates with all-position moments, 101,500 against 100,000 with sampled ones). An eight-block run of the state contrast under a sampled update, with the same clipping, gave 8/8 pairs (median ratio 1.77).

Two replications of 100,000 updates, at the common scale on WikiText-103 and clipped on an equally large FineWeb-Edu sample, showed that sampled moments delayed onset in 8/8 pairs ($p = 0.0078$ each), with median capped ratios 1.74 and 2.10. These are lower bounds, since three sampled-moment runs on WikiText-103 and one on FineWeb-Edu never reached onset and were assigned 98,500. At equal attained loss, the attention lead of all-position moments reversed. They had the higher mean score in 8/8 common-scale WikiText-103 pairs over updates 5,000–100,000 but in 1/8 over a shared range of attained loss (descriptive, as the states have different histories).

H.3 Pairing of the first and second moments

This study trained the six stored pairs $(m; v)$ of Table 12(a) at the common scale for 10,000 updates (about 2.8 passes), before any onset seen in this setup. Its two tests asked whether the matched v ended below mean-only m, which differs only in its v source, and below sampled moments on the primary set at update 10,000. Mean-only v was added post hoc, and the other comparisons are descriptive.

The matched v ended below both in 8/8 seeds ($p = 0.0078$ each), by a median of 0.575 nats against mean-only m (range 0.416–0.716) and 0.277 against sampled moments (0.267–0.292). Only mean-

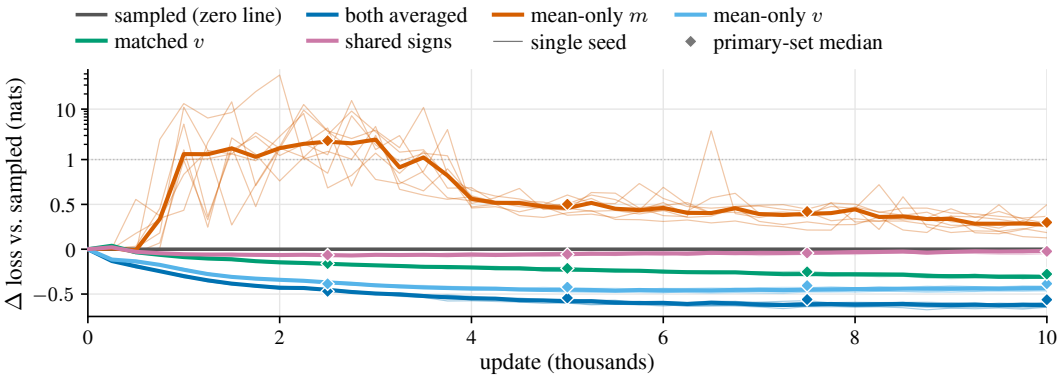


Fig. 4: Loss of each $\beta_1 = 0.9$ condition of Table 12(a) minus that of sampled moments in the same seed, on the diagnostic set: single seeds (thin) and medians (thick). Diamonds mark medians on the primary set. The vertical axis is linear up to 1 nat and logarithmic above it (dotted line) up to 60 nats. One seed of mean-only m peaked at $+47$ nats at update 2,000.

only m was unstable (Fig. 4). Its diagnostic loss rose 4.6–46.6 nats above its running minimum in every seed, against at most 0.013 for the others. The median endpoint of mean-only v lay 68% of the way from sampled to both-averaged moments, against 49% for matched v (descriptive). A sign pattern D shared by both moments lowered the loss by 0.024 nats (8/8).

H.4 Reflected first moment at $\beta_1 = 0.5$

The pairing study was repeated on new seeds with $\beta_1 = 0.5$ and the matched v throughout, storing in m the sampled αg_J, mean-only $\alpha\bar{g}$, or reflected $\alpha(\bar{g} - r)$, with long-run gains κ (Eq. (2)) of 1, 0.5, and 0. Under a frozen preconditioner, sampled and reflected m give the same stationary per-step numerator covariance, larger than that of mean-only m (Sec. 2.4). Reflected m was tested for a lower final loss than sampled m and, separately, than mean-only m.

Both tests rejected ($p = 0.0078$ each; Table 12(b)). In 8/8 seeds, reflected m ended more than 0.05 nats below sampled m (median 0.185 nats, range 0.176–0.198) and below mean-only m (median 0.070, range 0.061–0.078). The endpoints were ordered by long-run gain in every seed. From update 2,500 on, both differences had the same sign at each evaluation on the primary set. At the shared learning rate, reflected m's median direction norm over the 10,000 updates was 1.19 times mean-only m's and 1.10 times sampled m's in every seed (post hoc). Sampled m took 1.08 times mean-only m's steps yet ended 0.117 nats above it, so larger steps alone did not order these endpoints. A follow-up on the same seeds scaled reflection's learning rate by 0.843 to match mean-only m's median update norm, giving ratios of 0.998–0.999 to mean-only m and 0.92–0.93 to sampled m (post hoc). Reflected m then ended below sampled m in 8/8 seeds (median 0.077 nats, Holm $p = 0.0078$) and above mean-only m in 8/8 (median 0.033 nats), so step size explains its lead over mean-only m. These runs used RTX A6000 GPUs with the same software, and a rerun of one original reflected run matched its loss to within 2×10^{-6} over 250 updates.

Reflection is a probe of the stored residual, not a training method. No all-position condition was run at $\beta_1 = 0.5$, and reflected m ended above both-averaged moments and matched v at $\beta_1 = 0.9$, a descriptive comparison across seeds and decays.

H.5 Per-tensor localization and sign mixing

The pairing and reflection studies logged the per-tensor Adam direction d_t at every update, which we analyzed post hoc. The direction energy $\|d\|^2$ under mean-only m over a sampled v was 47 times that under sampled moments, and 93.5% of it lay in the output matrix, while the improvements lay elsewhere. First-order progress along the all-position mean ($\sum_t \langle d_t, \alpha\bar{g}_t \rangle$) measures alignment rather than loss. On this measure, 96.6% of matched v's improvement over sampled moments lay outside the embedding and output matrices, as did 96.0% and 95.9% of the improvements of reflected m over

Table 13: Final held-out NLL (nats) on the primary set after 5,000-update continuations of the eight sampled-moment parents, with current update αg_J, and the number of parents ending below the sampled condition. The follow-up's tests used the peak NLL, given in the text.

			NLL		Below
Condition	Stored m	Stored v	median	range	sampled
First study					
sampled	αg_J	αg_J	4.30	4.30–4.44	
sign-permuted m	$\alpha(\bar{g} + Rr)$	αg_J	393.57	13.01–2,434.47	0/8
sign-mixed m	$\alpha(\bar{g} + Dr)$	αg_J	6.64	4.59–50.19	0/8
mean-only m	$\alpha\bar{g}$	αg_J	7.58	4.32–13.22	0/8
both averaged	$\alpha\bar{g}$	$\alpha\bar{g}$	3.95	3.95–4.09	8/8
Follow-up: coherent pairs					
matched v	$\alpha\bar{g}$	$\alpha\sqrt{\bar{g}^2 + r^2}$	3.933		8/8
shared signs	$\alpha(\bar{g} + Dr)$	$\alpha(\bar{g} + Dr)$	4.196		8/8
shared signed permutation	$\alpha(\bar{g} + Rr)$	$\alpha(\bar{g} + Rr)$	4.027		8/8
averaged in vocabulary tensors	$\alpha\bar{g}$ there, αg_J elsewhere		4.009		8/8
averaged in other tensors	$\alpha\bar{g}$ there, αg_J elsewhere		4.244		8/8

sampled and mean-only m. One untested hypothesis is that mean-only m over a sampled v disrupts the output rows of rarely predicted tokens, whose second moments are poorly estimated (Chowdhery et al., 2023; Kunstner et al., 2024).

Sign mixing in m acted like mean-only m in recall (Appendix E.2), but here a sign pattern shared by both moments recovered about a tenth of matched v's improvement, $(5.166 - 5.142)/(5.166 - 4.889) \approx 0.09$. In a single-seed, single-step measurement after preconditioning, a sign-scrambled residual moved the output logits 10–46% as much as the unscrambled one (at most 5% in recall), mostly through the output weights of rarely predicted tokens. The measurement does not explain why the shared sign pattern recovered so little.

H.6 Continuation from sampled parents

The parents are the eight common-scale WikiText-103 models with sampled moments, saved at update 100,000 with both moments and the optimizer clock; five had passed onset. Each continuation ran 5,000 updates, and a parent's conditions shared data and position masks.

The first study stored, over a sampled v, the sign-permuted m, $\alpha(\bar{g} + Rr)$ with a fixed signed permutation $R = DQ$ within each tensor, or the sign-mixed m, $\alpha(\bar{g} + Dr)$, both of which acquired in recall. Neither lowered the loss in any parent (0/8, $p = 1$ for each test), and both raised it sharply (Table 13), although every update stayed finite. Averaging both moments lowered it in every parent, by 0.338–0.355 nats.

A follow-up added five coherent pairs to ask whether the increase came from building m and v from different histories. Because mean-only m had ended within 0.03 nats of the sampled condition in two parents, the follow-up's tests used the peak NLL on the primary set over updates 1,000, 2,500, and 5,000, an endpoint chosen post hoc. In 8/8 parents, the matched v peaked lower than mean-only m, and the shared sign pattern lower than sign-mixed m ($p = 0.0078$ each; median peak reductions 16.0 and 33.0 nats). On the final NLL, both contrasts were also positive in 8/8 parents. The matched v ended below both-averaged moments (median 3.933 against 3.951), recovering a median of 106% of the improvement from averaging both, against about half from scratch. Confined to the vocabulary tensors (the embedding and output matrices), averaging both moments recovered 83% of this improvement, and 18% when confined to the other tensors. Each of these percentages is the median over parents of that parent's fraction, not a ratio of the medians in Table 13. The mismatch of m and v need not be the only cause of the loss increase.

H.7 Position order without a mean-based source

Here the current update and both moments used the sampled gradient, as in plain AdamW, and no condition stored or applied a mean-based source. Each update still computed the all-position gradient $\bar{g}$, because the common scale $\alpha(\bar{g})$ of Eq. (6) multiplies the sampled gradient, so the study is not free of the exact mean. Only the order of supervised positions departed from the sampled condition of Appendix H.3 ($\beta_1 = 0.9$, 64 of 511 positions per sequence, 10,000 updates). In three conditions, each batch of 64 sequences was held for a block of 8 updates. Cycling drew, per held sequence and block, a random permutation of the 511 positions and used its consecutive 64-position segments in updates 1–7. Update 8 used the remaining 63 positions plus one drawn uniformly from the other 448, so every update averaged exactly 64 distinct positions per sequence and each position was used once per block except one, used twice. Drawing with replacement took a fresh uniform 64-subset at every update (335.8 distinct positions per sequence and block on average), and echo reused the block's first subset for all 8 updates, as data echoing reuses batches (Choi et al., 2019). In all three, each update's subset is marginally a uniform 64-subset. A fourth condition, the sampled training of Appendix H.3, drew a fresh batch at every update, and eight seeds ran all four conditions with a shared initialization and sequence stream.

Two exact one-sided paired sign tests on the primary set at update 10,000, in a fixed sequence, both rejected. Cycling ended below drawing with replacement in 8/8 seeds (median 0.134 nats, range 0.120–0.143, $p = 0.0039$), and drawing with replacement below echo in 8/8 (median 0.777 nats, range 0.739–0.856, $p = 0.0039$). Both differences exceeded 0.05 nats in every seed. The endpoint is a block boundary, and at the mid-block evaluations of updates 2,500 and 7,500 cycling was also below drawing with replacement in 8/8 seeds (medians 0.070 and 0.118 nats). Holding a batch cost loss, since fresh batches ended 0.076 nats below cycling (8/8). How much it cost depended on the order within the held batch.

At fixed parameters, a cycled block's residuals nearly cancel, summing to $(d - \bar{g})/64$, where d is the gradient at the repeated positions. Per sequence and coordinate, with σ^2 the variance of the per-position gradients, the variance of the block's residual sum is $0.000244\sigma^2$ for cycling, against $0.109559\sigma^2$ with replacement (about 449 times larger) and $0.876471\sigma^2$ for echo. Each update's residual variance is $0.013695\sigma^2$ in all three. Within a block the parameters move, and a cycled or echoed subset is not independent of the current state, so the residual has zero mean only marginally, and this cancellation holds only at fixed parameters. The ordering agrees with the long-run gain but also with position coverage (511, 335.8, and 64), the number of distinct supervised targets, the residual variance stored in m, and random reshuffling (Mishchenko et al., 2020). It is therefore not evidence for the long-run gain over these accounts. The study concerns early, pre-onset loss of one 29M-parameter model on one corpus, not the recall mechanism.

I Boundaries and replications

Runs use the unclipped recipe of Appendix B unless stated, and tests are exact one-sided paired sign tests on the AUC.

Recipe changes. Sampled training acquired within the horizon (Table 14) under query/key (Q/K) normalization (Henry et al., 2020), fourfold width (Appendix I.2), 1,280 contexts per update (Appendix E.3), and Lion (Chen et al., 2023), and later under Muon (Jordan et al., 2024). None of these changes shows where in AdamW's memory the residual acts, as sampled training did not fail under the first four and Muon replaced AdamW's moments on the block weight matrices.

Q/K normalization applied RMSNorm ($\epsilon = 10^{-6}$) to the query and key projections of every block. Unclipped, sampled training acquired so fast (11/12 runs, KM median 500) that speed tests of mean-only m above sampled training and of $c = -\frac{1}{2}$ above $c = \frac{1}{2}$ would be uninformative. Lion, which has no second moment, used $(\beta_1, \beta_2) = (0.9, 0.99)$, decoupled weight decay 0.3, clipping to global norm 1.0, and learning rate 10^{-5}, the smallest of 10^{-4}, 3×10^{-5}, and 10^{-5} at which two all-query calibration runs both acquired within 3,000 updates. Muon used Nesterov momentum 0.95, five Newton–Schulz iterations, learning rate 0.02, decoupled weight decay 0.01, and no clipping; its runs were evaluated every 250 updates on 1,024 freshly drawn contexts.

I.1 Second-moment decay 0.999 and a cosine schedule

Do the failure and the repair by mean-only m need a short memory of v or a constant learning rate? One setting used $\beta_2 = 0.999$, so that v averages over about 1,000 updates instead of 20, and the other decayed the learning rate to zero at update 8,000 on a cosine schedule, scaling the weight decay with it. In each setting, sampled training, mean-only m, and sign-mixed m ran on eight seeds.

Eight of the 48 runs stopped early after CUDA out-of-memory errors caused by another process on the same GPU. In both settings, sampled training acquired in none of its completed runs, and mean-only m and sign-mixed m in all of theirs (Table 14). Every complete pair favored mean-only m, but with each missing pair counted as non-positive, the Holm-adjusted tests of mean-only m above sampled training did not reject (7/8 and 6/8 positive pairs, Holm $p = 0.070$ and 0.14). When the eight runs were rerun post hoc, both settings ended at 0/8 for sampled training and 8/8 for mean-only m.

I.2 Model shape and alphabet size

Three replications changed only the model shape: width 320 (MLP width 1,280) or 640 (MLP width 2,560) in four blocks, or eight blocks of width 160, with the reference learning rate and warmup. Each ran six conditions on eight seeds (Table 15). At every shape, mean-only m exceeded sampled training in AUC in 8/8 pairs (Holm $p = 0.0078$ each), though at width 640, where sampled training always acquired, only in speed.

In these eight-seed studies, the test of $D_{6/8}$ above $D_{2/8}$ did not reject at any of the three shapes. A replication at eight blocks on 16 new seeds, with the same recipe and masks, did: $D_{6/8}$ had the larger AUC in 13 of 16 pairs ($p = 0.011$, mean difference 0.42) and acquired in 11/16 runs against 1/16 (KM median 1,500). At the reference shape with $\beta_2 = 0.999$, where sampled training fails (Appendix I), 16 further seeds were run with the same masks and sampled v: both masks acquired in every run, and $D_{6/8}$ had the larger AUC in 16/16 pairs ($p = 1.5 \times 10^{-5}$) with KM medians of 1,000 and 1,250 updates, but also 1.09 times $D_{2/8}$'s early update norm, so the contrast reduced to a small speed difference. Neither counts nor crossing times separated the masks at widths 320 and 640. There, every run with either mask acquired, and in 15 of 16 pairs both runs first crossed at the same evaluation (update 500 at width 320, 300 at width 640). The paired AUC differences were small. Mask $D_{6/8}$ had the larger AUC in 6/8 pairs at width 320 (mean difference, $D_{6/8}$ minus $D_{2/8}$, 0.0035, Holm $p = 0.14$), and $D_{2/8}$ in 6/8 at width 640 (one tie, mean difference -0.00014, 95% bootstrap interval -0.00038 to 0.00023, Holm $p = 0.996$). Of the eight-block runs, 5/8 acquired with $D_{6/8}$ and 1/8 with $D_{2/8}$, but only 5/8 AUC pairs were positive (Holm $p = 0.36$).

Alphabet size 16. Sign-permuted m acquired in 8/8 runs at $N = 8$ but in 2/8 in each of two experiments at $N = 16$, and in the first its AUC exceeded that of sampled training in 4/8 pairs (Holm $p = 1$; Appendix J.1). In the second, which had no test, mean-only, sign-mixed, and sign-permuted m acquired in 8/8, 6/8, and 2/8 runs. All three have κ near 0.1 (the signed ones in expectation), so

Table 14: Recipe changes. Runs acquiring recall (accuracy 0.95) within the stated updates, unclipped unless marked. Crossing is the Kaplan–Meier (KM) median or the range of first crossings, in updates. Q/K: query/key. –: not run.

		Sampled training		Comparison	
Change	Updates	Acq.	Crossing	Condition	Acq.
reference	8,000	0/8	>8,000	mean-only m	8/8
Q/K norm.	3,000	11/12	500	five stored sources	12/12
width 640	8,000	8/8	400	five stored sources	8/8
1,280 contexts	8,000	8/8	800	mean-only m, 256 contexts	8/8
$\beta_2 = 0.999$	8,000	0/7	>8,000	mean-only m	7/7
cosine schedule	8,000	0/6	>8,000	mean-only m	6/6
Q/K norm., clipped	8,000	8/8	500–1,000	all-query update	8/8
Lion, clipped	8,000	4/4	4,000–5,500	all-query	4/4
Muon	32,000	8/8	10,500–22,250	–	–
Muon, 1,280 contexts	32,000	8/8	750–2,000	–	–

the long-run gain does not explain their order. Averaging v as well did not add to mean-only m (both-averaged moments acquired in 6/8 runs in the first experiment and 1/8 in the second).

I.3 A MULTI-QUERY RECALL TASK

The second task is a compact multi-query associative recall (Arora et al., 2024). A sequence lists K key–value pairs from disjoint halves of a 256-token vocabulary and then repeats the keys in random order, and the model outputs each key's value. All-query training supervises all K queries ($\bar{g}$), and sampled training one uniformly drawn query per sequence (g_J); a condition is written as its sources (u, s^m, s^v) of Eq. (1). We trained a two-block decoder of width 128 (four heads, MLP width 512) without clipping for 6,000 updates of 256 sequences, with AdamW at $(\beta_1, \beta_2) = (0.9, 0.95)$, learning rate 10^{-3} after 500 warmup updates, and weight decay 0.01. A run acquires when its accuracy on a fixed evaluation set of 1,024 sequences, evaluated every 100 updates, first reaches 0.9.

First experiment. Eight conditions ran on twelve seeds at each of $K = 6$, 8, 10, and 16. Two sequential AUC tests at $K = 6$, of $(g_J, \bar{g}, g_J)$ against (g_J, g_J, g_J) and against $(g_J, g_J, \bar{g})$, each rejected with 12/12 positive pairs ($p = 2.4 \times 10^{-4}$). Averaging m alone, $(g_J, \bar{g}, g_J)$, acquired in 12/12 runs at $K = 6$ and 11/12 at $K = 16$ but in 0/12 at $K = 10$, and averaging both moments, $(g_J, \bar{g}, \bar{g})$, acquired in only 2/12 at $K = 6$. At $K = 6$, both half copies failed, while $(g_J, \bar{g} + Dr, g_J)$, which stores a full-norm residual, acquired in 12/12. Runs at $K = 8$ are uninformative and not shown, since $(\bar{g}, \bar{g}, \bar{g})$ never acquired there. A β_1 experiment at $K = 10$ on twelve further seeds tested $(g_J, \bar{g}, g_J)$ at $\beta_1 = 0.95$ above 0.9 (7/12 positive pairs, $p = 0.39$). Because all-query training at $\beta_1 = 0.95$ acquired in only 2/12 runs, this comparison is uninformative.

Second experiment. On fresh seeds, a second experiment crossed all-query and sampled sources for the update, m, and v at $K = 6$, 10, and 16. With the all-query update, the stored first moment decided the outcome at $K = 6$ and 16. Under the sampled update, as in the first experiment, $(g_J, \bar{g}, \bar{g})$ acquired in only 3/12 runs at $K = 6$ and $(g_J, \bar{g}, g_J)$ in 12/12 at $K = 16$. The main test compared $\kappa = 0$ with $\kappa = 0.2$ at equal conditional mean and covariance (the $\pm r/9$ rows) at $K = 10$, where neither acquired and all-query training acquired in only 6/12 runs. That comparison is uninformative. A further test favored $(g_J, \bar{g}, \bar{g})$ over $(\bar{g}, g_J, g_J)$ at $K = 6$ (12/12 pairs), a difference in AUC alone. Three Holm-adjusted tests of $(\bar{g}, g_J, \bar{g})$ above $(\bar{g}, g_J, g_J)$ did not reject at $K = 6$ and reversed at $K = 10$ and 16, where neither condition acquired.

We have no account of why $(g_J, \bar{g}, \bar{g})$ fails where $(g_J, \bar{g}, g_J)$ acquires, and at $K = 10$ both failed. With the half copies and the runs at $N = 16$, these results show that the long-run gain, which ignores v and the preconditioner's dynamics, is incomplete.

I.4 QUERY ORDER WITHOUT THE MEAN

We varied query order to test whether the long-run gain predicts anything when no condition uses the exact mean. In the unclipped reference recipe, each batch of 256 contexts was kept for five consecutive updates, 1,600 batches in 8,000 updates, with one query per context and update. Three

Table 15: Model shape and alphabet size. Runs of eight acquiring recall (accuracy 0.95) by update 8,000, unclipped. –: not run.

	Shape, $N = 8$			$N = 16$	
Condition	width 320	width 640	8 blocks	first	second
sampled	1/8	8/8	0/8	0/8	0/8
mean-only m	8/8	8/8	8/8	–	8/8
sign-mixed m	8/8	8/8	8/8	–	6/8
sign-permuted m	–	–	–	2/8	2/8
reflected m	–	–	–	0/8	–
both-averaged m and v	8/8	8/8	8/8	6/8	1/8
$D_{2/8}$ ($\kappa = 0.55$)	8/8	8/8	1/8	–	–
$D_{6/8}$ ($\kappa = -0.35$)	8/8	8/8	5/8	–	–

Table 16: Multi-query task. Runs of 12 acquiring (accuracy 0.9) by update 6,000, unclipped. –: not run.

Source (u, s^m, s^v)			First experiment			Second experiment		
Update	Stored m	Stored v	$K = 6$	10	16	$K = 6$	10	16
g_J	g_J	g_J	0	0	0	–	0	0
$\bar{g}$	$\bar{g}$	$\bar{g}$	12	8	12	12	6	12
g_J	$\bar{g}$	g_J	12	0	11	–	0	12
g_J	g_J	$\bar{g}$	0	0	0	–	0	–
g_J	$\bar{g}$	$\bar{g}$	2	0	0	3	0	0
g_J	$\bar{g} + Dr$	g_J	12	2	8	–	–	–
g_J	$\bar{g} + \frac{1}{2}r$	g_J	0	0	0	–	–	–
g_J	$\bar{g} - \frac{1}{2}r$	g_J	1	0	0	–	–	–
$\bar{g}$	$\bar{g}$	g_J	–	–	–	12	5	12
$\bar{g}$	g_J	g_J	–	–	–	0	0	0
$\bar{g}$	g_J	$\bar{g}$	–	–	–	0	0	0
g_J	$\bar{g} - r/9$	$\bar{g} - r/9$	–	–	–	–	0	–
g_J	$\bar{g} + r/9$	$\bar{g} + r/9$	–	–	–	–	0	–

conditions differed only in the queries of updates 2–5 of each block. Cycling drew each context's five queries without replacement. The residuals of a block therefore summed to zero at fixed parameters (measured ratio of the squared block sum to the summed squares 2.7×10^{-13}, against 0.94 with replacement and 5.0 for a repeated draw). Drawing with replacement visited 3.36 distinct queries per context and block on average, and the echo condition repeated one query five times. The three conditions of a seed shared the initialization, the contexts, and the first query of every block, and 12 seeds ran all three.

An exact one-sided paired sign test on the AUC of cycling against drawing with replacement did not reject (8 of 12 positive pairs, $p = 0.19$), so the second test in the fixed sequence, with replacement against echo, was not run. Cycling acquired in 2/12 runs, drawing with replacement in 0/12, and echo in 0/12 (mean AUC 0.26, 0.20, and 0.20), which left both tested conditions near the floor.

The design does not isolate the long-run gain. Coverage (5, 3.36, and 1 distinct queries per context and block) and the residual variance stored in m (about 0.19, 1, and 4.3 times that of drawing with replacement) are ordered in the same way. Every condition also reuses each batch five times, unlike the fresh batches of the other studies, and every scheduled evaluation falls on a block boundary. A null result on 12 seeds does not show that query order has no effect.

The interventions with a mean-based source need the exact mean. With a mean estimated from two draws, the reflected source equals the second draw's gradient, which is the independent draw that failed (Appendix E.2).

J Other studies

These studies did not identify the recall computation through which stored residuals affect acquisition. In the residual-control experiment (Appendix E.2), the first block's READ-to-query attention fell from about 1/13 at initialization in every condition, to 0.063–0.064 at update 600 in the sign-mixed, sign-permuted, redrawn, and mean-only conditions and to 0.044–0.046 in sampled and reflected runs (seed means). The first four conditions all acquired by update 900 (KM median 800), and at update 1,000 their attention averaged 0.270–0.279, against 0.034 in sampled and reflected runs (ranges 0.030–0.038), so the later gap accompanies acquisition. These descriptive analyses did not show how the residual prevents acquisition, and one used an update chosen post hoc. Recall runs follow Sec. 2.1 with a sampled current update and eight seeds per condition, unclipped in the first three studies. With $M_t = \beta_1 M_{t-1} + (1 - \beta_1)\bar{g}_t$, the accumulated residual is $E_t = m_t - M_t$.

J.1 ACCUMULATED STATE

In runs forked at update 500, we rotated E_t at fixed norm to a target cosine with M_t during updates 501–1,000, each target against a norm-matched same-angle sham. Zero alignment did not rescue sampled runs (above the sham in 3/8 pairs, Holm $p = 1$), but an absolute cosine of 0.5 lowered the AUC of sign-permuted runs in every pair (Holm $p = 0.012$). This Holm family's third test, sign-permuted m above sampled training at $N = 16$, did not reject (4/8 positive pairs, Holm $p = 1$).

J.2 FORCED SIGN

Prompted post hoc by that study, we set the cosine of E_t with M_t in sign-permuted runs to $+0.5$ or -0.5. Both targets lowered the AUC over updates 500–3,000 below their shams in 8/8 seeds. A third test asked whether -0.5 lowered it more than $+0.5$, and the reverse held in every seed ($p = 1$). These three tests and the $c = \pm\frac{1}{2}$ contrast of Appendix E.2 (raw $p = 9/256$) formed one Holm family, with adjusted $p = 0.016$ for each sham test and 0.070 for the $c = \pm\frac{1}{2}$ contrast.

J.3 LOCALIZATION

Storing the sign-mixed source $\bar{g} + Dr$ in m only for the query and key weights gave 0/8 acquisitions, and storing it in their complement or everywhere gave 8/8. Two Holm families of two exact one-sided Wilcoxon signed-rank tests on the AUC compared each partial mask with sampled training and whole-model mixing with each partial mask. The complement exceeded sampled training in AUC, and mixing everywhere exceeded query/key only (Holm $p = 0.0078$ each) and the complement (7/8 seeds, paired mean 0.002, Holm $p = 0.027$). Query/key only against sampled training was null at the floor. As query and key held a median of only 17% of the preconditioned stored-residual energy, this outcome does not place the effect outside them. The partial masks also changed different amounts of the stored residual.

J.4 ATTENTION GUIDANCE AND CLAMP

A first-block attention-logit bias of 4 from READ to the query, removed by update 3,000, was tested for an AUC increase of at least 0.10 over both no bias and a fixed-position bias over updates 3,000–8,000. It exceeded them by that margin in 8/8 and 6/8 seeds, and the intersection–union test did not reject ($p_{\mathrm{IUT}} = 0.145$). Clamping first-block READ attention to the query in two heads of sampled runs gave 5/8 acquisitions with the clamp in place, in an earlier recipe with 64 contexts per update, but 0/8 when the same training was evaluated without the clamp.

J.5 HEAD ABLATIONS

At update 100,000, we ablated the two top prefix-matching heads in each of the sixteen common-scale language models (Appendix H). Relative to matched random head pairs, the ablation raised the NLL on repeated text, and more so than on non-repeated text, in 8/8 models with all-position moments and 7/8 with sampled ones (intersection–union test, $p_{\mathrm{IUT}} = 0.0352$). The all-position lead on these readouts vanished or reversed at equal attained loss. These ablations support a functional role for the heads, not an attention-specific explanation of the difference between sampled and all-position moments.