# EMBEDDING NOISE SHOULD FOLLOW THE LEARNING RATE IN LLM FINE-TUNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Adding random noise to token embeddings improves supervised fine-tuning (SFT) of large language models (LLMs). The noise helps early in training, when a small instruction set is revisited for several epochs, but it becomes a problem at the end. As the learning rate (LR) decays, the optimizer takes smaller steps, but noise of fixed scale keeps adding a gradient bias that does not shrink. In our case study, keeping the noise through the last quarter of training leaves the model farther from a stationary point of the clean training loss, which is the loss without noise, and the gradient of this loss stays larger than when the noise is removed. We introduce AdaNoiseSFT, which sets the variance of the embedding noise proportional to the current LR, so the noise is strong while the LR is high and fades as the LR decays, with no additional model passes or optimizer state. Our analysis for Adam assumes that the smoothing effect of the noise does not vanish late in training. Under this condition, fixed noise leaves a gradient bias that does not vanish for any LR schedule, whereas coupling the noise to the LR drives this bias to zero when the LR decays to zero. To test that the benefit comes from following the LR schedule itself, we evaluate cosine, linear, step, and warmup stable decay (WSD) LR schedules against independently tuned noise tapers that follow elapsed training time. On Llama-3.1-8B with Alpaca-cleaned and a cosine LR, AdaNoiseSFT reaches 19.5 length-controlled AlpacaEval 2.0 (AE-2) over five seeds, versus 18.1 for the strongest baseline, a time taper selected on a separate selection split. Over ten seeds under linear, step, and WSD LR schedules, it beats the selected time taper by 1.39, 1.14, and 0.91 AE-2 points, respectively. For all four LR schedules, the best noise schedule is the one that follows the shape of the LR.

## 1 INTRODUCTION

Supervised fine-tuning (SFT) adapts pretrained large language models (LLMs) to small instruction datasets (Ouyang et al., 2022), often for several epochs (Taori et al., 2023). Adding random noise to token embeddings during this stage improves instruction following (Jain et al., 2024). Existing methods keep the noise level fixed during training (Jain et al., 2024), or change the sign of the noise (Yadav and Singh, 2023), its scale in each layer (Khadangi et al., 2025), or the weights it perturbs (Zhelnin et al., 2025). In all of them, the noise level does not depend on the learning rate (LR). In the same runs, the LR decays toward zero, for example under a cosine schedule (Loshchilov and Hutter, 2017) or warmup stable decay (WSD) (Hu et al., 2024). Much of the final progress happens in this decay phase (Wen et al., 2025). Near the end of training, the steps become small, but the fixed noise keeps shifting the expected gradient away from the gradient of the clean loss, which is the loss without noise. With this shift, the optimizer settles before it reaches a stationary point (Ajalloeian and Stich, 2020). The noise that helps early in training therefore pulls the final model away from a minimum.

A simple fix is to weaken the regularizer as training goes on. Regularization matters most early in training (Golatkar et al., 2019). Earlier work lowers the dropout rate (Rennie et al., 2014) or the variance of gradient noise (Neelakantan et al., 2015) over time. A recent inference method for diffusion language models also lowers its perturbations over the decoding steps (Wu et al., 2026). We call a noise schedule that follows elapsed time a *time taper*. However, a time taper follows the optimizer only when the LR has the same shape. Under step or WSD LR, the two curves separate. The noise then stays too strong after a sharp LR drop, or it fades too early during a long phase of



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

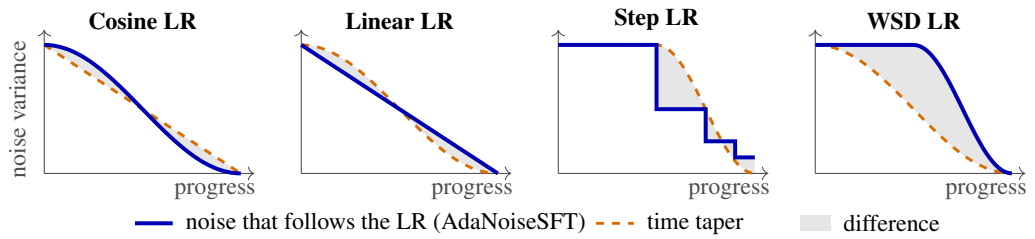



Figure 1: **Schedule shapes.** Each panel shows the noise variance, normalized by its peak, over the training progress after warmup. Blue follows the optimizer LR, as in AdaNoiseSFT. Orange is the time taper that Table 2 selects for that LR, and the shaded area marks where the two differ.

high LR. For each LR schedule, the decay shape must be chosen by hand, and a poor choice costs quality (Section 5.3).

We keep the idea of decaying the noise, but the decay follows the LR. Our contributions are as follows.

- We introduce AdaNoiseSFT, which scales the variance of the embedding noise with the current LR. The noise is strong while the LR is high and fades as the LR decays. AdaNoiseSFT needs no second forward pass and no extra optimizer state.
- We show that fixed noise leaves a gradient shift that does not vanish and keeps the clean loss above its minimum, while noise that decays with the LR removes both effects. The bound for Adam, the optimizer we use, builds on the analysis of Adam by Défossez et al. (2022), and the case of stochastic gradient descent (SGD) follows the bounds for biased gradients of Ajalloeian and Stich (2020).
- We show that the LR shape allows the most noise among schedules whose gradient shift stays within the error that the optimizer already makes at each step, and we test this in practice. Across cosine, linear, step, and WSD LR, AdaNoiseSFT beats the best time taper, tuned separately for each LR, by 0.91 to 1.50 points of length-controlled AlpacaEval 2.0 (AE-2) over ten seeds, and a second model shows the same pattern.
- On three models, AdaNoiseSFT has the highest AE-2 among all compared methods with one forward pass, and the same rule applied to LoRA dropout beats fixed dropout and dropout that decays with elapsed time.

## 2 Related Work

**Noise and consistency regularization for fine-tuning.** NEFTune adds uniform noise of a fixed scale to token embeddings during instruction tuning (Jain et al., 2024), and SymNoise uses sign-flipped noise pairs (Yadav and Singh, 2023). NoiseFiT sets the noise in each layer from a signal-to-noise ratio (Khadangi et al., 2025), GIFT-SW injects noise only into non-salient weights (Zhelnin et al., 2025), and NoisyTune perturbs the pretrained weights once before fine-tuning (Wu et al., 2022). Other methods add a consistency term instead. R3F penalizes the change in output under random embedding noise with a symmetric Kullback–Leibler (KL) divergence (Aghajanyan et al., 2021). FreeLB uses adversarial embedding perturbations (Zhu et al., 2020), SMART adds a smoothness penalty (Jiang et al., 2020), and R-Drop enforces agreement between two dropout passes (Liang et al., 2021).

**When to regularize.** Weight decay and data augmentation matter most during an early critical period (Golatkar et al., 2019). Annealed dropout lowers the dropout rate over training (Rennie et al., 2014), curriculum dropout raises it (Morerio et al., 2017), and early or late dropout confines it to one phase (Liu et al., 2023). Gradient noise with a decaying variance helps train deep networks (Neelakantan et al., 2015), and graduated optimization removes smoothing step by step (Hazan et al., 2016). Stochastic gradient Langevin dynamics (SGLD) ties injected parameter noise to the step size for posterior sampling (Welling and Teh, 2011).

**Learning rate schedules and noise in optimization.** Cosine annealing (Loshchilov and Hutter, 2017) and warmup stable decay (Hu et al., 2024) are standard for LLM training. Under WSD, much

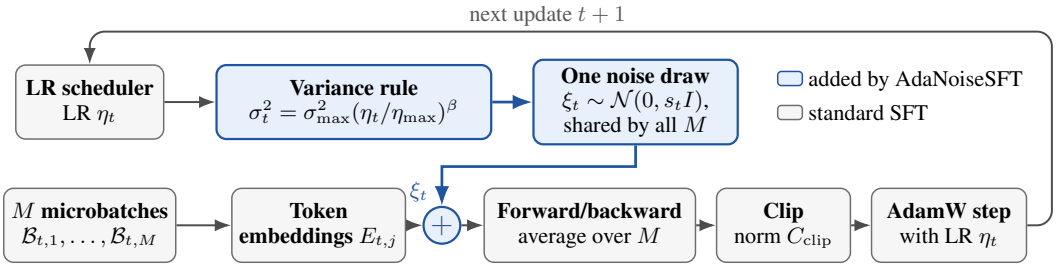



Figure 2: **Training procedure of AdaNoiseSFT for one update.** Blue steps are added by the method, and gray steps are standard SFT. The LR sets the noise variance, one noise draw is added to the token embeddings of all $M$ microbatches, and the rest of the update is standard. Here $s_t$ is the noise variance per embedding coordinate (Section 3.3).

of the loss reduction happens in the final decay (Wen et al., 2025), and the shape of this decay trades bias against variance in the final model (Dremov et al., 2025). Diffusion models of stochastic gradient descent (SGD) relate gradient noise to the LR (Mandt et al., 2017; Smith and Le, 2018; Malladi et al., 2022). Analysis of biased gradients bounds how far a systematic error keeps SGD from a stationary point (Ajalloeian and Stich, 2020), and recent work gives convergence bounds for Adam (Défossez et al., 2022). Appendix A gives an extended review.

# 3 Method

## 3.1 SFT with perturbation regularization

We fine-tune a pretrained language model on a dataset $\mathcal{D}$ of instruction–response pairs $(x, y)$, where $y$ is the response to instruction $x$. The trainable parameters are $\phi$, with initial value $\phi_1$. The main experiments use Low-Rank Adaptation (LoRA) (Hu et al., 2022) with the input embedding map frozen, so the embeddings are inputs to the model and not trainable parameters. The analysis covers this setting. Appendix E also reports fine-tuning of all parameters, which the analysis does not cover. The SFT objective averages the negative log-likelihood of the response tokens,

$$\mathcal{L}(\phi) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\frac{1}{|y|}\sum_k \log p_\phi(y_k \mid x, y_{<k})\Big], \tag{1}$$

where $p_\phi$ is the next-token distribution of the model, $y_k$ is the $k$-th response token, $y_{<k}$ are the tokens before it, and $|y|$ is the response length. We call $\mathcal{L}$ the clean objective because it contains no embedding noise. $\mathcal{L}_{\text{SFT}}(\phi, \mathcal{B}, E)$ denotes the same loss on a minibatch $\mathcal{B}$ when the model receives the token embeddings $E$.

Each optimizer update averages the gradients of $M$ microbatches, a standard setting called gradient accumulation. All microbatches of one update use the same noise draw, and the method does not depend on the value of $M$. We use three versions of the gradient at update $t$. The clean gradient is $g_t := \nabla\mathcal{L}(\phi_t)$. Its estimate from the $M$ microbatches $\mathcal{B}_{t,j}$ of the update, with their embeddings $E_{t,j}$ and without noise, is $\bar{g}_t := \frac{1}{M}\sum_{j=1}^M \nabla_\phi \mathcal{L}_{\text{SFT}}(\phi_t, \mathcal{B}_{t,j}, E_{t,j})$. The same estimate with noise is $\hat{g}_t$, the gradient that the optimizer receives before clipping. We write

$$\begin{aligned} &\hat{g}_t = g_t + b_t + d_t + \varepsilon_t, && \mathbb{E}[\varepsilon_t \mid \mathcal{F}_t] = 0, \\ &d_t := \mathbb{E}[\hat{g}_t \mid \mathcal{F}_t] - \mathbb{E}[\bar{g}_t \mid \mathcal{F}_t], && b_t := \mathbb{E}[\bar{g}_t \mid \mathcal{F}_t] - g_t, \end{aligned} \tag{2}$$

where $\mathcal{F}_t$ is the history before the microbatches and the noise of update $t$ are drawn. Thus $d_t$ is the shift of the expected gradient caused by the noise, $b_t$ is the sampling bias of the data order, and $\varepsilon_t$ is the remaining random part, with zero mean given the history and $\mathbb{E}[\|\varepsilon_t\|^2 \mid \mathcal{F}_t] \le V_t$. The analysis assumes that sampling is unbiased given the history, $b_t = 0$. Appendix C.4 states how $b_t$ enters under reshuffling at every epoch.

### 3.2 VARIANCE COUPLED TO THE LEARNING RATE

AdaNoiseSFT sets the noise variance at update $t$ to

$$\sigma_t^2 = \sigma_{\max}^2 \left( \frac{\eta_t}{\eta_{\max}} \right)^{\beta}, \qquad \beta > 0, \tag{3}$$

where $\eta_t$ is the optimizer LR at update $t$, $\eta_{\max}$ is the peak LR, $\sigma_{\max}^2$ is the peak noise variance, and $\beta$ is the coupling exponent. The LR schedule is fixed before training, so $\eta_{\max}$ is known in advance. Equation (3) is applied during warmup as well as decay, so the variance rises with the LR, reaches $\sigma_{\max}^2$ at the LR peak, and then decays with the LR. $T$ denotes the total number of updates.

**Choice of $\sigma_{\max}$ and $\beta$.** Earlier methods set the noise scale by hand and keep it fixed (Jain et al., 2024; Yadav and Singh, 2023), or lower the noise with elapsed time (Neelakantan et al., 2015). We select $\sigma_{\max}$ on the selection split from a grid around the noise scales of earlier work (Appendix D.2). The exponent $\beta$ sets how closely the variance follows the LR. With $\beta = 1$, the variance is proportional to the LR, as the injected noise is proportional to the step size in stochastic gradient Langevin dynamics (Welling and Teh, 2011). Proposition 3 shows that this shape allows the most noise within a bound on the error per step. We select $\beta$ by clean loss on the selection split from $\{0.5, 1, 1.5, 2\}$, which gives $\beta = 1$. The selection split is the 5% of each corpus that is not used for training (Section 5.1).

### 3.3 GAUSSIAN EMBEDDING NOISE

For the embedding tensor of a padded sequence $E_t \in \mathbb{R}^{n_{\text{tok}} \times d_E}$, where $n_{\text{tok}}$ is the padded sequence length and $d_E$ is the embedding width, we sample

$$\xi_t \sim \mathcal{N}\left(0, \frac{\sigma_t^2}{n_{\text{tok}} d_E} I\right), \qquad \widetilde{E}_t = E_t + \xi_t. \tag{4}$$

Here $I$ is the identity matrix and $\widetilde{E}_t$ is the perturbed embedding tensor. One perturbation is drawn per optimizer update and shared by all its microbatches. We use Gaussian noise because its odd moments are zero, which gives the gradient shift the simple second-order form of Section 4, and because small Gaussian input noise acts as a smoothness penalty on the model (Bishop, 1995). The uniform noise of earlier work (Jain et al., 2024) is also symmetric, and Appendix D.2 relates the two scales. Appendix D.2 describes padding, masking, and a variant with one draw per example.

The full tensor contains $d_\xi := n_{\text{tok}} d_E$ coordinates, so $\mathbb{E}\|\xi_t\|_F^2 = \sigma_t^2$. The noise variance in one embedding coordinate is $s_t := \sigma_t^2 / d_\xi$, with value $s_{\max} := \sigma_{\max}^2 / d_\xi$ at the LR peak. The main setting uses $n_{\text{tok}} = 512$ and $\sigma_{\max} = 10$.

### 3.4 TRAINING PROCEDURE

AdaNoiseSFT changes the perturbation variance before the standard SFT forward and backward pass. Algorithm 1 combines the variance rule of Section 3.2 with the noise draw of Section 3.3 into the full training loop, and Figure 2 shows the same pipeline. In the algorithm, $\phi_t$ and $o_t$ are the parameters and the AdamW moments before update $t$, and Clip rescales the gradient to norm at most $C_{\text{clip}}$. After $T$ updates the trained parameters are $\phi_{T+1}$. Here "Ada" denotes conditioning on the current LR at every update. The name does not refer to a learned controller. Appendix F defines the optional antithetic, consistency, and gradient alignment extensions.

## 4 ANALYSIS

The analysis uses the noise variance per coordinate $s_t = \sigma_t^2 / d_\xi$ of Section 3.3. We write $\ell(\phi, E)$ for the minibatch loss averaged over response tokens, $G_\phi(E) = \nabla_\phi \ell(\phi, E)$ for its gradient, $\mathbb{E}_{\mathcal{B}}$ for the expectation over the next minibatch, and $e_j$ for the $j$th coordinate basis vector in $\mathbb{R}^{d_\xi}$. The quantity $H_t := \frac{1}{2}\mathbb{E}_{\mathcal{B}}\left[\sum_{j=1}^{d_\xi} \nabla_E^2 G_{\phi_t}(E)[e_j, e_j]\right]$ measures how strongly the noise smooths the gradient at the current parameters, that is, how much gradient shift each unit of variance per coordinate produces under the derivative and moment bounds of Assumption 1 in Appendix B.

**Algorithm 1. AdaNoiseSFT**

**Require:** initial LoRA parameters $\phi_1$ and AdamW moments $o_1$, training set $\mathcal{D}$
**Settings:** LR schedule $\eta_1, \dots, \eta_T$ with peak $\eta_{\max}$, peak noise variance $\sigma^2_{\max}$, exponent $\beta$, microbatches $M$, clipping norm $C_{\text{clip}}$
**Ensure:** trained parameters $\phi_{T+1}$

**for** $t = 1, \dots, T$ **do**
  $\sigma_t^2 \leftarrow \sigma^2_{\max}(\eta_t/\eta_{\max})^\beta$
  draw $\xi_t$ from Eq. (4) and set $\hat{g}_t \leftarrow 0$
  **for** $j = 1, \dots, M$ **do**
    sample microbatch $\mathcal{B}_{t,j}$ from $\mathcal{D}$, with token embeddings $E_{t,j}$
    $\hat{g}_t \leftarrow \hat{g}_t + \frac{1}{M}\nabla_\phi \mathcal{L}_{\text{SFT}}(\phi_t, \mathcal{B}_{t,j}, E_{t,j} + \xi_t)$
  **end for**
  $(\phi_{t+1}, o_{t+1}) \leftarrow \mathsf{AdamWStep}(\phi_t, o_t, \text{Clip}(\hat{g}_t, C_{\text{clip}}), \eta_t)$
**end for**
**return** $\phi_{T+1}$

Lemma 1 in Appendix B.1 expands the gradient shift $d_t$ of Eq. (2) to second order. The lemma gives $d_t = s_t H_t + r_t$ with a remainder $\|r_t\| \le K s_t^2$ and the bound $\|d_t\| \le C_{\text{b}} s_t$, where $K$ and $C_{\text{b}}$ are finite constants. The shift therefore grows in proportion to the noise variance per coordinate.

The next statement needs that this smoothing does not vanish late in training. We call this condition tail nondegeneracy. There are $t_0$ and $h_0 > 0$ such that $\|H_t\| \ge h_0$ for all $t \ge t_0$, and the remainder satisfies $K s_{\max} \le h_0/2$ over the analyzed range of variances.

**Proposition 1** (Persistence under fixed noise)**.** *Under tail nondegeneracy, every fixed variance $s_t \equiv s \in (0, s_{\max}]$ leaves $\|d_t\| \ge h_0 s/2 > 0$ for all $t \ge t_0$, for any LR schedule. If instead $s_t = s_{\max}\eta_t/\eta_{\max}$ and $\eta_t \to 0$, then $\|d_t\| \to 0$.*

The proof is in Appendix B.2. Persistence concerns the size of the gradient shift, and its direction can change along training.

Our experiments train with AdamW without weight decay, so the main bound is stated for Adam. Adam divides each gradient coordinate by the square root of its running second moment plus a stability constant $\epsilon_{\text{A}}$. We write $\bar{P}_t$ for the expected value of this diagonal scaling at update $t$, and $p_{\max} = 1/\epsilon_{\text{A}}$ for its largest possible entry. With momentum rate $b_1$, the step size corrected for the start of the momentum average is $q_t = \eta_t/(1 - b_1^t)$. The quantity $\mathcal{G}_T = \sum_t q_t \mathbb{E}\|g_t\|^2_{\bar{P}_t}$ is a weighted sum of squared clean gradient norms, where $\|v\|_P^2 = v^\top P v$, and $\Delta\mathcal{L} = \mathcal{L}(\phi_1) - \inf_\phi \mathcal{L}(\phi)$ is the initial gap to the minimum.

**Theorem 1** (Perturbation bias under Adam)**.** *Assume that (i) the clean objective $\mathcal{L}$ is bounded below and $L_0$-smooth, (ii) the perturbed gradient is $\hat{g}_t = g_t + d_t + \varepsilon_t$ with $\mathbb{E}[\varepsilon_t \mid \mathcal{F}_t] = 0$, (iii) the gradient shift satisfies $\|d_t\| \le C_{\text{b}} s_t$ for a constant $C_{\text{b}}$, and (iv) $\sup_t \mathbb{E}\|\hat{g}_t\|^2 \le G_2^2$. Then*

$$\frac{\mathcal{G}_T}{\sum_t q_t} \le \frac{2\Delta\mathcal{L}}{\sum_t q_t} + 3p_{\max}C_{\text{b}}^2 \frac{\sum_t q_t s_t^2}{\sum_t q_t} + \Omega_T + \frac{2\mathcal{J}_T^+}{\sum_t q_t}, \tag{5}$$

*where $\Omega_T$ collects step size terms without $s_t$ that vanish when $\eta_t \to 0$ and $\sum_t q_t \to \infty$. The term $\mathcal{J}_T^+$ is the positive part of a remainder from the interaction of momentum with the scaling that Adam applies to each coordinate.*

Assumptions (i), (ii), and (iv) are the standard assumptions of analyses of Adam (Défossez et al., 2022), and (iii) holds by Lemma 1 under the smoothness conditions of Appendix B. The proof is in Appendix C. The left side measures how close training gets to a point where the clean gradient is zero, and the shift caused by the noise enters only through the average of $s_t^2$ weighted by $q_t$. The noise term stays constant under fixed noise and goes to zero with the LR under LR coupling. The bound does not prove that Adam converges, which would also require $\mathcal{J}_T^+$ to grow more slowly than $\sum_t q_t$. For SGD the same argument gives Corollary 1 in Appendix B.4, a bound of the type known for biased gradients (Ajalloeian and Stich, 2020).

**Proposition 2** (Distance to the optimum under fixed noise)**.** *Assume that the clean objective is $L_0$-smooth with minimum value $\mathcal{L}^\star = \inf_\phi \mathcal{L}(\phi)$ and that the variance is fixed at $s \in (0, s_{\max}]$. Consider a point $\bar{\phi}$ where the expected perturbed gradient is zero, $\nabla\mathcal{L}(\bar{\phi}) + d(\bar{\phi}) = 0$, with $d(\phi)$ the gradient shift at parameters $\phi$, and assume that tail nondegeneracy holds at $\bar{\phi}$. Then $\mathcal{L}(\bar{\phi}) - \mathcal{L}^\star \ge$*

$h_0^2 s^2/(8L_0)$. *If $\mathcal{L}$ also satisfies the Polyak–Łojasiewicz inequality $\|\nabla\mathcal{L}(\phi)\|^2 \geq 2\mu(\mathcal{L}(\phi) - \mathcal{L}^\star)$ near $\bar{\phi}$, then $\mathcal{L}(\bar{\phi}) - \mathcal{L}^\star \leq C_{\mathrm{b}}^2 s^2/(2\mu)$.*

The proof is in Appendix B.5. For SGD, such a point is where training with fixed noise stops moving on average, and for Adam this holds up to the covariance term of Appendix C.2. The statement does not show that training reaches the point. There the clean loss stays above its minimum by an amount of order $s^2$ that does not shrink as the LR decays, while under LR coupling the final variance $s_{\max}\eta_T/\eta_{\max}$ drives the bound to zero. The final clean gradient norm is also higher with fixed noise, 0.420 against 0.342 with AdaNoiseSFT over five seeds on Llama-3.1-8B with Alpaca-cleaned (Table 3), and 0.433 against 0.340 for the reserved seed of the case study (Section 5.4).

Theorem 1 and Proposition 2 support decaying the noise, but they do not say which decay to use. Suppose we require the noise to add at most $B_{\mathrm{step}}\eta_t^2$ to each update for a constant $B_{\mathrm{step}} > 0$, the order of the optimizer's own error per step. By Lemma 1, the condition $C_{\mathrm{b}} s_t \leq B_{\mathrm{step}}\eta_t$ guarantees this.

**Proposition 3** (The LR shape is the largest schedule within the step bound)**.** *Every schedule with $C_{\mathrm{b}} s_t \leq B_{\mathrm{step}}\eta_t$ at all updates satisfies $s_t \leq (B_{\mathrm{step}}/C_{\mathrm{b}})\eta_t$. The schedule $s_t = (B_{\mathrm{step}}/C_{\mathrm{b}})\eta_t$ meets the condition with equality at every update, so it is the largest such schedule at every update.*

The proof is in Appendix B.6. The largest schedule is AdaNoiseSFT with $\beta = 1$, and if more noise gives stronger regularization, it regularizes most within the bound. A time taper with another shape uses less noise than allowed where its curve lies below the LR and breaks the bound where it lies above. Under cosine LR a linear taper does both, early and late in training, which matches its lower AE-2 in Section 5.3.

**What is proved and what is not.** Fixed noise keeps training away from the optimum of the clean loss, and noise coupled to the LR removes this effect as the LR decays (Theorem 1, Proposition 2). Proposition 3 explains why the LR shape is preferred, but none of these results predicts benchmark quality, so Section 5.3 tests the shapes directly. Appendix H gives the bias of each schedule in closed form.

# 5 Experiments

## 5.1 Setup

We evaluate Alpaca-cleaned (52k examples) (Taori et al., 2023), UltraChat-50k (a 50k subset of UltraChat-200k) (Ding et al., 2023), and Tulu-3-100k (100k Tulu-3-SFT examples) (Lambert et al., 2024), each split 80/5/5/10 into training, controller (Appendix F), selection, and diagnostic test parts. The main matrix uses Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and Qwen3-8B-Base (Qwen Team, 2025). Unless stated otherwise, runs use LoRA rank 16, three epochs, 100 warmup updates, AdamW, and cosine LR (Loshchilov and Hutter, 2017). Each update accumulates $M = 8$ microbatches of 8 examples, which gives an effective batch size of 64, and the gradient is clipped at $C_{\mathrm{clip}} = 1.0$. The schedule identification study additionally uses linear, step, and WSD LR (Hu et al., 2024) on Llama-3.1-8B and Mistral-7B-v0.3 with Alpaca-cleaned, and repeats step and WSD LR on Qwen3-14B-Base and on UltraChat-50k. LR, perturbation peak, and $\beta$ are selected on the selection split.

Besides vanilla SFT, we compare against nine baselines that cover the main families of SFT regularization. The baselines are embedding noise with a fixed level (NEFTune (Jain et al., 2024), SymNoise (Yadav and Singh, 2023), and fixed Gaussian noise with a default and a tuned scale), time tapers, consistency regularization (R-Drop (Liang et al., 2021)), balancedness-aware regularization for LoRA (BAR) (Li et al., 2024a), and two recent methods from 2025, NoiseFiT (Khadangi et al., 2025) and GIFT-SW (Zhelnin et al., 2025), adapted to one forward pass (Appendix D). R-Drop runs two dropout passes and adds a symmetric KL consistency loss, and BAR penalizes the norm imbalance between the two LoRA factors. We report length-controlled AlpacaEval 2.0 (AE-2) (Dubois et al., 2024), MT-Bench (Zheng et al., 2023), Arena-Hard (Li et al., 2024b), IFEval (Zhou et al., 2023), Inverse-IFEval (Zhang et al., 2026), MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), BIG-Bench Hard (BBH) (Suzgun et al., 2023), and GSM8K (Cobbe et al., 2021). The central schedule comparison uses ten matched seeds and paired Student-$t$ 95% confidence intervals (CIs). Clean gradient norm is computed after averaging the clean gradient over a fixed audit

| Method | Instruction following (target of the method) | | | | | Capability retention | | | |
|---|---|---|---|---|---|---|---|---|---|
| | AE-2↑ | MT-B↑ | Arena-H↑ | IFEval↑ | Inv-IFE↑ | MMLU↑ | TruthQA↑ | BBH↑ | GSM8K↑ |
| *Llama-3.1-8B + Alpaca-cleaned* | | | | | | | | | |
| Vanilla SFT | 12.2 | 6.05 | 14.5 | 44.5 | 65.7 | 61.3 | 37.2 | 50.6 | 34.2 |
| NEFTune | 15.1 | 6.22 | 16.9 | 46.2 | 66.0 | 61.3 | 37.3 | 50.7 | 34.5 |
| SymNoise‡ | 15.3 | 6.36 | 17.4 | 47.0 | 66.3 | 61.1 | 37.4 | 50.7 | 34.4 |
| NoiseFiT-style | 14.2 | 6.18 | 16.4 | 45.7 | 66.0 | 61.1 | 37.7 | 50.7 | 34.5 |
| GIFT-SW-style | 14.1 | 6.24 | 16.1 | 46.1 | 66.0 | 61.3 | 37.3 | 50.9 | 34.4 |
| BAR | 14.0 | 6.31 | 16.2 | 46.0 | 66.0 | 61.2 | 37.4 | 50.5 | 34.5 |
| R-Drop‡ | 15.3 | 6.39 | 16.0 | 46.8 | 66.0 | 61.4 | 37.3 | 51.2 | 35.0 |
| Fixed Gaussian noise ($\sigma = 10$) | 15.0 | 6.33 | 16.4 | 46.3 | 65.8 | 61.3 | 37.4 | 50.9 | 34.5 |
| Tuned fixed Gaussian noise | 15.6 | 6.34 | 16.5 | 46.2 | 65.9 | 61.2 | 37.3 | 50.9 | 34.6 |
| Linear taper | 18.1 | 6.68 | 17.5 | 48.7 | 66.8 | 61.3 | 37.3 | 50.6 | 34.5 |
| **AdaNoiseSFT** | **19.5** | **6.74** | **19.2** | **50.2** | **67.0** | 61.3 | 37.3 | 50.8 | 34.4 |
| Two-pass extension‡ | 20.5 | 6.81 | 20.3 | 50.1 | 67.5 | 61.3 | 37.6 | 51.0 | 34.8 |
| *Mistral-7B-v0.3 + Alpaca-cleaned* | | | | | | | | | |
| Vanilla SFT | 11.0 | 5.87 | 12.5 | 42.3 | 61.9 | 60.1 | 39.4 | 48.9 | 29.5 |
| NEFTune | 13.3 | 6.07 | 14.5 | 43.6 | 62.1 | 60.1 | 39.3 | 49.2 | 30.0 |
| SymNoise‡ | 13.4 | 6.20 | 14.6 | 43.4 | 62.5 | 60.1 | 39.6 | 49.2 | 29.6 |
| NoiseFiT-style | 13.1 | 6.04 | 14.5 | 43.1 | 62.1 | 60.0 | 39.6 | 49.1 | 29.6 |
| GIFT-SW-style | 12.4 | 5.95 | 14.0 | 43.3 | 62.3 | 60.2 | 39.6 | 49.1 | 29.4 |
| BAR | 13.0 | 5.97 | 14.5 | 43.5 | 61.9 | 60.1 | 39.6 | 49.0 | 29.7 |
| R-Drop‡ | 13.2 | 6.04 | 14.6 | 44.5 | 62.5 | 60.1 | 40.0 | 49.3 | 30.1 |
| Fixed Gaussian noise ($\sigma = 10$) | 13.2 | 6.05 | 14.2 | 43.3 | 62.2 | 60.2 | 39.3 | 49.1 | 29.6 |
| Tuned fixed Gaussian noise | 13.2 | 6.05 | 13.7 | 43.6 | 62.0 | 60.1 | 39.3 | 49.2 | 29.5 |
| Linear taper | 15.6 | 6.41 | 15.9 | 45.1 | 62.7 | 60.2 | 39.3 | 49.4 | 29.9 |
| **AdaNoiseSFT** | **17.0** | **6.47** | **17.0** | **46.4** | **62.9** | 60.2 | 39.4 | 49.4 | 30.0 |
| Two-pass extension‡ | 17.8 | 6.53 | 17.9 | 47.1 | 63.3 | 60.1 | 39.8 | 49.5 | 30.2 |
| *Qwen3-8B-Base + Alpaca-cleaned* | | | | | | | | | |
| Vanilla SFT | 16.7 | 6.50 | 20.5 | 51.8 | 69.7 | 72.3 | 47.4 | 55.0 | 47.5 |
| NEFTune | 19.0 | 6.66 | 22.1 | 53.0 | 69.8 | 72.2 | 47.5 | 55.1 | 47.7 |
| SymNoise‡ | 19.1 | 6.81 | 22.8 | 53.0 | 70.1 | 72.1 | 47.3 | 55.1 | 47.4 |
| NoiseFiT-style | 18.7 | 6.65 | 21.9 | 52.3 | 70.4 | 71.9 | 47.2 | 55.2 | 47.8 |
| GIFT-SW-style | 18.2 | 6.59 | 21.9 | 52.5 | 70.3 | 72.2 | 47.3 | 55.2 | 47.3 |
| BAR | 18.5 | 6.69 | 22.0 | 52.9 | 70.0 | 72.5 | 47.4 | 55.2 | 47.8 |
| R-Drop‡ | 18.8 | 6.71 | 21.7 | 53.4 | 70.4 | 72.2 | 47.6 | 55.3 | 48.1 |
| Fixed Gaussian noise ($\sigma = 10$) | 18.8 | 6.65 | 21.9 | 52.5 | 70.3 | 72.2 | 47.4 | 55.2 | 47.5 |
| Tuned fixed Gaussian noise | 18.9 | 6.62 | 22.2 | 52.6 | 70.3 | 72.3 | 47.3 | 55.1 | 47.7 |
| Linear taper | 21.2 | 7.05 | 22.9 | 54.5 | 70.5 | 72.3 | 47.3 | 55.4 | 47.8 |
| **AdaNoiseSFT** | **22.0** | **7.07** | **24.1** | **55.4** | **70.7** | 72.4 | 47.4 | 55.4 | 47.9 |
| Two-pass extension‡ | 22.2 | 7.11 | 24.4 | 55.9 | 71.1 | 72.3 | 47.2 | 55.4 | 47.9 |

Table 1: **Main results on Alpaca-cleaned.** Three model families, five matched seeds, and the default cosine LR. The first five columns measure instruction following, which the method targets, and bold marks the best value among methods with one forward pass. The last four columns measure whether knowledge and reasoning are kept, where no gain is expected and no winner is marked. ‡ marks methods with two model passes per microbatch, including the two-pass extension of AdaNoiseSFT (Appendix F), which is shown for reference and is not compute-matched. NoiseFiT-style and GIFT-SW-style are adaptations with one forward pass (Appendix D), and $\sigma$ is the fixed noise scale. MT-B is MT-Bench, Arena-H is Arena-Hard, Inv-IFE is Inverse-IFEval, and TruthQA is TruthfulQA.

subset of 1,024 training examples. Appendix D gives the tuning grids, evaluation definitions, seed identities, uncertainty protocol, and timing details.

**Schedule controls.** A *time taper* uses the shape of the cosine, linear, step, or WSD schedule over training progress, independently of the optimizer LR. In Table 1, the fixed noise row uses the peak selected for LR coupling. Appendix D.2 gives the step and WSD definitions and the peak grids.

### 5.2 Main comparison

Table 1 tests whether decaying the noise helps. Under cosine LR, LR coupling coincides with the cosine taper, so the table measures the gain of this decaying schedule over fixed noise and the other baselines. On all three models, AdaNoiseSFT has the best score of all methods with one forward pass on each of the five instruction following benchmarks. Its AE-2 exceeds the strongest baseline, the linear taper, by 0.8 to 1.4 points, and tuned fixed Gaussian noise by 3.1 to 3.9 points. Most of the gain over fixed noise therefore comes from decaying the noise at all, and Section 5.3 isolates the part that comes from following the LR. On the four capability benchmarks, the method is not expected to help, and it does not hurt. AdaNoiseSFT never scores below vanilla SFT there and stays within 0.6 points of the best method. The two-pass extension of Appendix F adds a consistency loss with a second model pass. The extension scores higher but costs twice the compute, so the rows

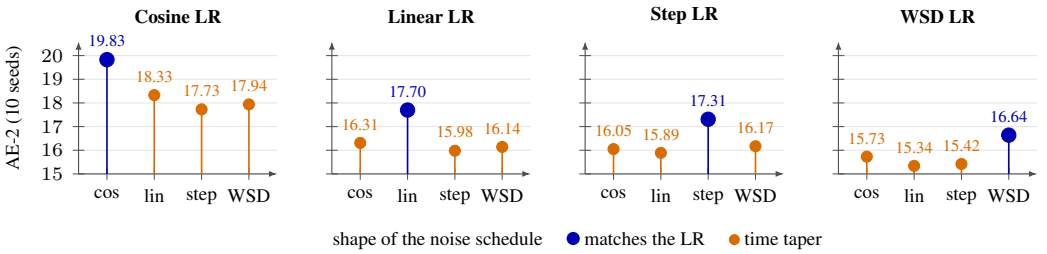



Figure 3: **The best noise schedule follows the LR.** Each panel fixes one LR schedule and shows the test AE-2 of the four noise shapes over ten seeds. Blue marks the shape that matches the LR, and orange marks the time tapers, each tuned separately. Appendix Table 37 lists the same values.

| Optimizer LR | AdaNoiseSFT AE-2 | Selected taper | Taper AE-2 | Δ [95% CI] |
|---|---|---|---|---|
| Cosine | 19.83 | linear | 18.33 | +1.50 [1.27, 1.73] |
| Linear | 17.70 | cosine | 16.31 | +1.39 [1.10, 1.68] |
| Step | 17.31 | WSD | 16.17 | +1.14 [0.84, 1.44] |
| WSD | 16.64 | cosine | 15.73 | +0.91 [0.62, 1.20] |

Table 2: **Schedule shape identification.** Llama-3.1-8B with Alpaca-cleaned. All rows use ten matched seeds, so the cosine row differs from the five seeds operating point in Table 1. Each taper and the displayed comparator are selected on clean selection loss before test AE-2 is evaluated. Δ is AdaNoiseSFT AE-2 minus taper AE-2, with a paired 95% CI.

with one forward pass give the comparison at matched compute (response length and more corpora are in Appendix E).

### 5.3 Schedule shape identification

This section tests whether the decay should follow the LR or the elapsed training time. Under cosine LR, AdaNoiseSFT with $\beta = 1$ and the cosine taper are the same schedule, so cosine LR alone cannot separate the two. We therefore repeat the protocol under linear, step, and WSD LR and select each mismatched comparator on the selection split before test evaluation.

Figure 3 shows the full $4 \times 4$ result. Each panel is one training run setting, defined by its LR schedule, and the four points in the panel are the four noise schedules trained under that LR. If the best noise schedule were the same for every LR, the same point would be highest in every panel. Instead, the highest point is always the noise schedule with the same shape as the LR, marked in blue, and it changes from panel to panel with the LR. The comparator selected by clean loss is also the strongest mismatched shape. Table 2 gives the paired gains, which range from 0.91 to 1.50 AE-2. A second complete matrix on Mistral-7B-v0.3 shows the same pattern, where the matching shape beats each mismatched taper by 0.74 to 1.61 AE-2 (Appendix J.1).

A direct test supports Proposition 3. Under step LR, clipping the WSD taper so that it never exceeds the LR raises AE-2 from 16.17 to 17.05 at the same peak, which recovers about three quarters of the gap to LR coupling at 17.31. Under WSD LR, moving the cosine taper halfway toward the LR raises AE-2 from 15.73 to 16.27, and full coupling reaches 16.64 (Appendix J.3).

### 5.4 Matched tail case study

The case study tests whether the noise can simply be removed at the end of training. The case study uses seed 77, reserved before training and excluded from all result sets. At training progress $\tau = 3/4$, the fraction of updates completed, three branches continue from the same checkpoint, which was trained with fixed noise. The three branches share the optimizer state, future minibatches, and dropout streams, so the noise in the final quarter is the only difference (Figure 4A). Keeping fixed noise raises the clean gradient norm to 0.433 at the end of training, removing the noise lowers it to 0.316, and switching to a tail coupled to the LR ends at 0.340 (Figure 4B). Removal therefore gives the lowest gradient norm, but it does not give the best quality. Over five seeds, removing noise raises AE-2 by +1.14 with CI [0.95, 1.34], and replacing the removed tail with noise coupled to the LR adds another +1.70 with CI [1.54, 1.87] (Figure 4C, Appendix Table 53, and Appendices I and H).

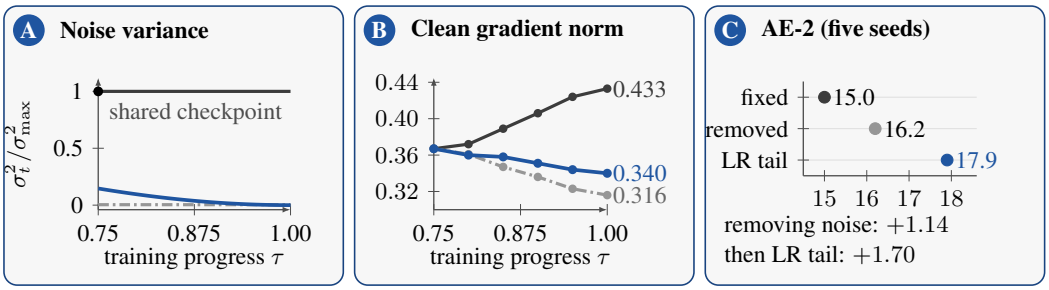



Figure 4: **Matched case study of the final quarter.** From the shared checkpoint at training progress $\tau = 3/4$, training continues with fixed noise (dark gray), without noise (light gray), or with noise coupled to the LR (blue). (A) Noise variance. (B) Clean gradient norm for the reserved seed 77. (C) AE-2 over five seeds and paired gains, with 95% CIs in Appendix Table 53. LR tail is the tail coupled to the LR.

| Noise schedule | AE-2↑ | AdaNoiseSFT − schedule [95% CI] | Clean grad. | Diag. loss | Sharp. |
|---|---|---|---|---|---|
| Fixed Gaussian noise ($\sigma = 10$) | 15.0 | +4.48 [4.04, 4.91] | 0.420 | 1.707 | 0.70 |
| Half variance | 15.6 | +3.94 [3.35, 4.53] | 0.372 | 1.669 | 0.67 |
| Linear taper | 18.1 | +1.42 [1.09, 1.76] | 0.354 | 1.653 | 0.65 |
| $\beta = 0.5$, matched decay variance | 18.2 | +1.27 [0.19, 2.35] | 0.329 | 1.603 | 0.61 |
| **AdaNoiseSFT** | 19.5 | N/A | 0.342 | 1.610 | 0.58 |

Table 3: **Schedule ablation.** Llama-3.1-8B with Alpaca-cleaned and five seeds. Entries are means over seeds 11–55, and paired AE-2 effects are AdaNoiseSFT minus the displayed schedule. The three optimization diagnostics, including clean gradient norm, are not measures of benchmark quality. The AdaNoiseSFT row is the reference for the paired effects, so its effect is marked N/A.

### 5.5 Replication, sensitivity, and controls

Table 3 separates schedule shape from noise level on Llama-3.1-8B with Alpaca-cleaned. Halving the fixed variance recovers little of the gap to LR coupling, the linear taper recovers most of it, and a $\beta = 0.5$ schedule matched in decay phase variance still trails LR coupling by 1.27 AE-2 [0.19, 2.35].

The gain over the selected taper holds in six settings under cosine LR and in three settings each under step and WSD LR, including Qwen3-14B-Base (Appendices D.2 and J.2), and Appendix D.2 reports sensitivity to $\beta$ and controls with SGD.

The same rule also helps a second regularizer. With embedding noise turned off, LoRA dropout coupled to the LR beats fixed dropout by 0.81 to 0.92 AE-2 and the comparator taper by 0.38 to 0.49 AE-2 under cosine, step, and WSD LR (Appendix J.4). The IFEval gains are smaller, and two of the three intervals include zero.

## 6 Conclusion

Embedding noise helps SFT early in training and harms it late, so its variance should decay. Across cosine, linear, step, and WSD LR, the best decay follows the LR rather than elapsed time, and AdaNoiseSFT implements this rule with no extra model pass. The analysis explains why fixed noise keeps training away from the optimum of the clean loss, and LoRA dropout coupled to the LR shows the same advantage, so the rule may apply to other regularizers as well.

## Limitations

The complete $4 \times 4$ studies use Alpaca-cleaned with two models, and the other settings repeat only parts of the comparison (Appendix Table 40). The theory covers LoRA with frozen embeddings and relies on the tail condition and smoothness assumptions of Section 4, and the Adam bound does not prove convergence. Evaluation uses English data and LLM judges, and Appendix G lists further limits.

## Reproducibility Statement

Appendices B and C state the assumptions and proofs, while Appendix D specifies the data partitions, optimizer settings, selection procedure, evaluation definitions, seed matching, and timing protocol. Deterministic schedule budgets follow the definitions stated in the corresponding appendices. Additional implementation details and machine-readable result records are provided in the supplementary material.

## AI Use Statement

We used large language models to aid and polish the writing of this paper. The authors are responsible for all content of the paper, including the method, the experiments, the results, and the citations.

## References


Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*, 2021.

Ahmad Ajalloeian and Sebastian U. Stich. On the convergence of SGD with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 639–668, 2022.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andres Marafioti, Hynek Kydlicek, Agustin Piqueres Lajarin, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clementine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Christopher M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019.

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

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Aleksandr Dremov, Alexander Hägele, Atli Kosson, and Martin Jaggi. Training dynamics of the cooldown stage in warmup-stable-decay learning rate scheduler. *Transactions on Machine Learning Research*, 2025.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Aditya Sharad Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. In *Advances in Neural Information Processing Systems 32*, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Elad Hazan, Kfir Yehuda Levy, and Shai Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1833–1841. PMLR, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. NEFTune: Noisy embeddings improve instruction finetuning. In *International Conference on Learning Representations*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, 2020. doi: 10.18653/v1/2020.acl-main.197.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Afshin Khadangi, Amir Sartipi, Igor Tchappi, and Ramin Bahmani. Noise augmented fine tuning for mitigating hallucinations in large language models. *arXiv preprint arXiv:2504.03302*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5905–5914, 2021.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Bingcong Li, Liang Zhang, and Niao He. Implicit regularization of sharpness-aware minimization for scale-invariant problems. In *Advances in Neural Information Processing Systems 37*, 2024a.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. In *Advances in Neural Information Processing Systems 36*, 2023.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024b.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems 34*, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

Yuhang Liu, Tao Li, Zhehao Huang, Zuopeng Yang, and Xiaolin Huang. Bi-LoRA: Efficient sharpness-aware minimization for fine-tuning large-scale models. In *International Conference on Learning Representations*, 2026.

Zhuang Liu, Zhiqiu Xu, Joseph Jin, Zhiqiang Shen, and Trevor Darrell. Dropout reduces underfitting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22233–22248, 2023.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

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

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems 35*, 2022.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. Memorization in NLP fine-tuning methods. *arXiv preprint arXiv:2205.12506*, 2022.

Pietro Morerio, Jacopo Cavazza, Riccardo Volpi, René Vidal, and Vittorio Murino. Curriculum dropout. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3544–3552, 2017.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Steven J. Rennie, Vaibhava Goel, and Samuel Thomas. Annealed dropout training of deep networks. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 159–164, 2014.

Samuel L. Smith and Quoc V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

Samuel L. Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9058–9067. PMLR, 2020.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. GitHub repository, 2023.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems 26*, 2013.

Fei Wang and Bo Li. Leaner training, lower leakage: Revisiting memorization in LLM fine-tuning with LoRA. *arXiv preprint arXiv:2506.20856*, 2025.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. In *International Conference on Learning Representations*, 2025.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. NoisyTune: A little noise can help you finetune pretrained language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 680–685, 2022. doi: 10.18653/v1/2022.acl-short.76.

Jingxuan Wu, Zhenglin Wan, Xingrui Yu, Yuzhe Yang, Yiqiao Huang, Ivor Tsang, and Yang You. Time-annealed perturbation sampling: Diverse generation for diffusion language models. *arXiv preprint arXiv:2601.22629*, 2026.

Abhay Kumar Yadav and Arjun Singh. SymNoise: Advancing language model fine-tuning with symmetric noise. *arXiv preprint arXiv:2312.01523*, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems 33*, pages 5824–5836, 2020.

Qinyan Zhang, Xinping Lei, Ruijie Miao, Yu Fu, Haojie Fan, Le Chang, Jiafan Hou, Dingling Zhang, Zhongfei Hou, Ziqiang Yang, Changxin Pu, Fei Hu, Jingkai Liu, Xinjie Chen, Jianpeng Jiao, Jiaheng Liu, Tong Yang, Zaiyuan Wang, Ge Zhang, and Wenhao Huang. Inverse IFEval: Can LLMs unlearn stubborn training conventions to follow real instructions? In *International Conference on Learning Representations*, 2026.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems 35*, 2022.

Maxim Zhelnin, Viktor Moskvoretskii, Egor Shvetsov, Maria Krylova, Egor Venediktov, Aleksandr Zuev, and Evgeny Burnaev. GIFT-SW: Gaussian noise injected fine-tuning of salient weights for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6463–6480, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track*, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *International Conference on Learning Representations*, 2025.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. FreeLB: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020.

# A Extended Related Work

Section 2 gives the comparisons needed to position the main contribution. This appendix records additional distinctions that matter for interpretation of the controls and extensions.

**Where perturbations enter.** NEFTune (Jain et al., 2024) and SymNoise (Yadav and Singh, 2023) perturb token embeddings. NoisyTune (Wu et al., 2022) perturbs pretrained weights once before fine-tuning. GIFT-SW (Zhelnin et al., 2025) trains salient columns and injects noise into non-salient columns during fine-tuning. NoiseFiT (Khadangi et al., 2025) chooses layerwise levels from activation statistics. R3F perturbs input embeddings and adds a symmetric KL representation-preserving penalty (Aghajanyan et al., 2021). FreeLB perturbs word embeddings adversarially, while SMART combines smoothness-inducing regularization with proximal optimization (Zhu et al., 2020; Jiang et al., 2020). These methods change perturbation location, objective, distribution, or scale. Our claim is narrower: Gaussian embedding perturbation whose magnitude follows LR.

**Sharpness-aware methods.** Sharpness-aware minimization (SAM) (Foret et al., 2021) adds an inner ascent step with a fixed radius, adaptive sharpness-aware minimization (ASAM) (Kwon et al., 2021) rescales the perturbation by parameter magnitude, and BAR (Li et al., 2024a) uses a balancedness regularizer for LoRA while avoiding SAM's second gradient computation. Bi-LoRA (Liu et al., 2026) studies a two-branch sharpness-aware construction. Work on SAM variants also studies when sharpness-aware updates are most useful during training (Andriushchenko and Flammarion, 2022; Zhou et al., 2025). We use these works as comparison context and do not transfer their scheduling conclusions to embedding noise.

**Timing effects.** Golatkar et al. (2019) show that several regularizers can matter most during an early critical period and can later be removed with little loss of generalization. This observation motivates our shared-checkpoint intervention, which tests whether an embedding-noise effect present before decay fraction $\tau = 3/4$ survives the final decay quarter without perturbation.

**Dropout schedules.** Dropout admits both regularization and noise interpretations (Wager et al., 2013). Annealed dropout lowers the rate (Rennie et al., 2014), curriculum dropout raises it (Morerio et al., 2017), and later analysis separates early and late roles (Liu et al., 2023). We do not test a dropout schedule, so these studies provide context for phase-dependent regularization.

**Annealed stochasticity and continuation.** Stochastic gradient Langevin dynamics (SGLD) ties parameter-space noise to step size for posterior sampling (Welling and Teh, 2011). The scaling is different here. Under first-order propagation, embedding-noise variance $s_t$ contributes $O(\eta_t^2 s_t)$ variance to a parameter update, so $s_t \propto \eta_t$ gives $O(\eta_t^3)$ update variance. SGLD instead injects parameter-space variance proportional to $\eta_t$. Explicit gradient-noise schedules can decay with training time (Neelakantan et al., 2015), and continuation methods progressively remove smoothing (Hazan et al., 2016).

**Adaptive-optimizer context.** SGD noise and LR are linked in diffusion approximations (Mandt et al., 2017; Smith et al., 2020). Adaptive-optimizer stochastic differential equation (SDE) analyses

retain moment dynamics explicitly (Malladi et al., 2022). Appendix C does not use these works to derive the schedule. Its formal result keeps the current-gradient Adam denominator and first-moment momentum explicit.

**Diagnostics.** Training-data extraction and memorization motivate the exact-continuation probe (Carlini et al., 2021; 2023; Mireshghallah et al., 2022; Wang and Li, 2025). The sharpness diagnostic follows the local perturbation perspective of SAM and large-batch generalization work (Foret et al., 2021; Keskar et al., 2017). R-Drop (Liang et al., 2021) motivates the bidirectional consistency form used only in the optional two-pass extension. Tulu-3 (Lambert et al., 2024) supplies the source collection for our Tulu-3-100k transfer corpus. Gradient-alignment signals appear in multi-task optimization (Yu et al., 2020). Our optional controller uses such a signal to modulate the envelope of the variance coupled to the LR.

# B PROOFS

Throughout this appendix, $\|\cdot\|$ is the Euclidean norm, $\|\cdot\|_F$ is the Frobenius norm, $\mathcal{F}_t$ is the history immediately before the fresh minibatch and perturbation at update $t$, and $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. We reserve $E$ for embedding activations and use the blackboard-bold operator $\mathbb{E}$ exclusively for expectation. For Gaussian embedding noise, $s$ denotes the variance per coordinate in $\xi \sim \mathcal{N}(0, sI_{d_\xi})$, where $d_\xi = n_{\text{tok}} d_E$ and $\sigma^2 = s\, d_\xi$. Constants below absorb dimension-dependent derivative and Gaussian-moment factors. The perturbation analysis treats the provided embedding activations $E$ as external to the trainable coordinates $\phi$, which applies directly to the main LoRA setting with a frozen input-embedding map. The full parameter SmolLM2 check (Ben Allal et al., 2025) is empirical and is outside the scope of Lemma 1 and Corollary 1.

**Assumption 1** (Gaussian-line regularity)**.** *Let $\ell(\phi, E)$ denote the response-token-normalized minibatch loss and $G_\phi(E) := \nabla_\phi \ell(\phi, E)$. The map $G_\phi$ is four times continuously differentiable in $E$. For $Z \sim \mathcal{N}(0, I_{d_\xi})$, every analyzed $0 \le s \le s_{\max}$, and every interpolation coefficient $\vartheta \in [0,1]$, the first four directional derivatives of $G_\phi(E + \vartheta\sqrt{s}Z)$ admit minibatch-uniform random envelopes with the Gaussian moments used below. In particular,*

$$\mathbb{E}\left[\|Z\|^4 \sup_{\vartheta\in[0,1]} \|D_E^4 G_\phi(E + \vartheta\sqrt{s}Z)\|_{\text{op}}\right] < \infty,$$

*and corresponding square-integrable envelopes hold for the lower-order terms used in Proposition 5. The mixed Jacobian*

$$J_{\phi E}(\phi, E) := \nabla_E G_\phi(E) = \nabla_\phi \nabla_E^\top \ell(\phi, E)$$

*satisfies $\|J_{\phi E}\|_F \le C_J$. Finally, $\nabla_\phi \mathcal{L}$ and each minibatch gradient are $L_0$-Lipschitz in the trainable parameters. These bounds hold uniformly over the union of optimization trajectories generated by the schedules compared in the analysis and over $s \in [0, s_{\max}]$. The derivative envelopes, $C_J$, $L_0$, and the resulting constants can therefore be chosen common to those schedules. This remains a trajectory-local bounded-region assumption.*

## B.1 BIAS LEMMA AND ITS PROOF

**Lemma 1** (Perturbation-induced bias)**.** *There are finite constants $K$ and $C_{\text{b}}$, the same for all updates of the compared runs, such that*

$$d_t = s_t H_t + r_t, \qquad \|r_t\| \le K s_t^2, \qquad \|d_t\| \le C_{\text{b}} s_t. \tag{6}$$

Here $r_t$ is the higher order remainder of the expansion.

Write $\hat{g}_t = \hat{g}(\mathcal{B}_t, \xi_t)$ for the perturbed minibatch gradient. For a fixed minibatch $\mathcal{B}$, define

$$d^{\mathcal{B}} := \mathbb{E}_\xi[\hat{g}(\mathcal{B}, \xi)] - \nabla \ell_{\mathcal{B}}(\phi_t).$$

By Eq. (2), $d_t = \mathbb{E}_{\mathcal{B}}[d^{\mathcal{B}}]$, where $\mathbb{E}_{\mathcal{B}}$ averages over the conditional distribution of the next minibatch under the implemented sampler. This identity does not require an unbiased sampler. Appendix C.4 handles the separate sampling-bias term $b_t$.

Let $\xi = \sqrt{s}Z$ with $Z \sim \mathcal{N}(0, I_{d_\xi})$ and write $G(E) = \nabla_\phi \ell_{\mathcal{B}}(\phi, E)$. Taylor expansion through third order gives

$$G(E+\xi) = G(E) + J_{\phi E}\xi + \frac{1}{2}\mathcal{Q}[\xi,\xi] + \frac{1}{6}\mathcal{T}[\xi,\xi,\xi] + \mathcal{R}_4(\xi),$$

where $\mathcal{Q} = \nabla_E^2 G(E)$, $\mathcal{T} = \nabla_E^3 G(E)$, and the integral remainder satisfies

$$\|\mathcal{R}_4(\xi)\| \le \frac{\|\xi\|^4}{6}\int_0^1 (1-\vartheta)^3 \|D_E^4 G(E+\vartheta\xi)\|_{\mathrm{op}}\, d\vartheta.$$

Assumption 1 gives $\mathbb{E}\|\mathcal{R}_4(\xi)\| = O(s^2)$ uniformly over the analyzed minibatches. Gaussian symmetry makes both the linear and cubic expectations zero, hence

$$d^{\mathcal{B}} = \frac{s}{2}\sum_{j=1}^{d_\xi} \mathcal{Q}[e_j, e_j] + O(s^2).$$

Averaging over minibatches gives $d_t = s_t H_t + r_t$ with $\|r_t\| \le K s_t^2$ for a finite constant $K$ shared across the compared trajectories. Uniform boundedness of $H_t$ on the analyzed region then gives $\|d_t\| \le C_{\mathrm{b}} s_t$. Since $s_t = \sigma_t^2/d_\xi$, the equivalent bound is $\|d_t\| \le \bar{C}_{\mathrm{b}}\sigma_t^2$ with $\bar{C}_{\mathrm{b}} = C_{\mathrm{b}}/d_\xi$. The constants $K$ and $C_{\mathrm{b}}$ include Gaussian moments that grow with $d_\xi$, so the lemma describes how the shift scales with the variance and does not certify that the remainder is small at a given noise level. The condition $K s_{\max} \le h_0/2$ used in Section 4 is an assumption about this remainder. □

### B.2 Proof of Proposition 1

Lemma 1 gives $d_t = s_t H_t + r_t$ with $\|r_t\| \le K s_t^2$. Under tail nondegeneracy, every fixed $s \in (0, s_{\max}]$ satisfies $Ks \le K s_{\max} \le h_0/2$, and therefore, for $t \ge t_0$,

$$\|d_t\| \ge s\|H_t\| - \|r_t\| \ge s(h_0 - Ks) \ge \frac{h_0 s}{2} > 0.$$

The fixed-noise lower bound does not involve the LR sequence. For $s_t = s_{\max}\eta_t/\eta_{\max}$, Lemma 1 gives $\|d_t\| \le C_{\mathrm{b}} s_{\max}\eta_t/\eta_{\max} \to 0$ whenever $\eta_t \to 0$. □

### B.3 SGD as a special case

For SGD, consider

$$\phi_{t+1} = \phi_t - \eta_t \hat{g}_t, \tag{7}$$

with $\hat{g}_t$ decomposed as in Eq. (2). Assume $L_0$-smoothness, a lower-bounded clean objective, conditionally centered stochastic variation with second moment at most $V_t$, $b_t = 0$, the bound of Lemma 1, and $0 < \eta_t \le 1/(4L_0)$. Let $S_k(T) = \sum_{t=1}^T \eta_t^k$ and $\Delta\mathcal{L} = \mathcal{L}(\phi_1) - \inf_\phi \mathcal{L}(\phi)$.

**Corollary 1** (Stationarity bound for SGD)**.** *Under these assumptions,*

$$\min_{1\le t\le T} \mathbb{E}\|\nabla\mathcal{L}(\phi_t)\|^2 \le \frac{4\Delta\mathcal{L}}{S_1(T)} + \frac{2L_0\sum_t \eta_t^2 V_t}{S_1(T)} + 3C_{\mathrm{b}}^2\frac{\sum_t \eta_t s_t^2}{S_1(T)}. \tag{8}$$

The last term is the contribution of the perturbation bias. It is the SGD form of the noise term in Theorem 1, with $q_t = \eta_t$ and no adaptive scaling.

### B.4 Proof of Corollary 1

Write $\nabla_t := \nabla\mathcal{L}(\phi_t)$. Smoothness and the update in Eq. (7) imply

$$\mathcal{L}(\phi_{t+1}) \le \mathcal{L}(\phi_t) - \eta_t \nabla_t^\top(\nabla_t + d_t + \varepsilon_t) + \frac{L_0\eta_t^2}{2}\|\nabla_t + d_t + \varepsilon_t\|^2.$$

Conditioning on $\mathcal{F}_t$ removes the linear $\varepsilon_t$ term, and

$$\mathbb{E}_t\|\nabla_t + d_t + \varepsilon_t\|^2 \le 2\|\nabla_t\|^2 + 2\|d_t\|^2 + V_t.$$

Young's inequality gives $-\eta_t \nabla_t^\top d_t \le \frac{\eta_t}{2}\|\nabla_t\|^2 + \frac{\eta_t}{2}\|d_t\|^2$. Therefore

$$\mathbb{E}_t \mathcal{L}(\phi_{t+1}) \le \mathcal{L}(\phi_t) - \Big(\frac{\eta_t}{2} - L_0\eta_t^2\Big)\|\nabla_t\|^2 + \Big(\frac{\eta_t}{2} + L_0\eta_t^2\Big)\|d_t\|^2 + \frac{L_0\eta_t^2}{2}V_t.$$

The step-size condition gives $L_0\eta_t^2 \le \eta_t/4$. Using Lemma 1 and summing gives

$$\frac{1}{4}\sum_{t=1}^{T}\eta_t \mathbb{E}\|\nabla_t\|^2 \le \Delta\mathcal{L} + \frac{3}{4}C_{\mathrm{b}}^2\sum_{t=1}^{T}\eta_t s_t^2 + \frac{L_0}{2}\sum_{t=1}^{T}\eta_t^2 V_t.$$

Divide by $S_1(T)/4$ and use the weighted average to obtain Eq. (8). □

**Schedule-dependent fluctuation envelope.** Assumption 1 also gives the leading samplewise expansion

$$\hat{g}_t^{(s)} - \hat{g}_t^{(0)} = J_{\phi E}\xi_t + O_{L^2}(s_t).$$

Here $\hat{g}_t^{(s)}$ and $\hat{g}_t^{(0)}$ are the update-$t$ gradients with and without a perturbation of variance per coordinate $s_t$, and $O_{L^2}(s_t)$ denotes a term whose root-mean-square norm is $O(s_t)$. Thus perturbation contributes $O(s_t)$ to the conditional second moment on the bounded variance range. If the remaining minibatch fluctuation is bounded by $V_0$, the envelope $V_t \le V_0 + C_{\mathrm{v}}s_t$ yields

$$2L_0V_0\frac{S_2(T)}{S_1(T)} + 2L_0C_{\mathrm{v}}\frac{\sum_t \eta_t^2 s_t}{S_1(T)}$$

for the fluctuation term. No independence between minibatch and perturbation fluctuations is required.

### B.5 Proof of Proposition 2

At $\bar{\phi}$ the expected perturbed gradient is zero, so $\nabla\mathcal{L}(\bar{\phi}) = -d(\bar{\phi})$. The conditions of Proposition 1 at $\bar{\phi}$ give $\|d(\bar{\phi})\| \ge s\|H(\bar{\phi})\| - Ks^2 \ge s(h_0 - Ks_{\max}) \ge h_0 s/2$, where $H(\phi)$ is the smoothing term $H_t$ of Section 4 evaluated at parameters $\phi$, hence $\|\nabla\mathcal{L}(\bar{\phi})\| \ge h_0 s/2$. For an $L_0$-smooth function that is bounded below, one gradient step of size $1/L_0$ from $\bar{\phi}$ gives $\mathcal{L}^\star \le \mathcal{L}(\bar{\phi}) - \|\nabla\mathcal{L}(\bar{\phi})\|^2/(2L_0)$, so

$$\mathcal{L}(\bar{\phi}) - \mathcal{L}^\star \ge \frac{\|\nabla\mathcal{L}(\bar{\phi})\|^2}{2L_0} \ge \frac{h_0^2 s^2}{8L_0}.$$

Under the Polyak–Łojasiewicz inequality near $\bar{\phi}$, $\mathcal{L}(\bar{\phi}) - \mathcal{L}^\star \le \|\nabla\mathcal{L}(\bar{\phi})\|^2/(2\mu) = \|d(\bar{\phi})\|^2/(2\mu) \le C_{\mathrm{b}}^2 s^2/(2\mu)$ by Lemma 1. □

### B.6 Proof of Proposition 3

Dividing $C_{\mathrm{b}}s_t \le B_{\mathrm{step}}\eta_t$ by $C_{\mathrm{b}} > 0$ gives $s_t \le (B_{\mathrm{step}}/C_{\mathrm{b}})\eta_t$ at every update. The schedule $s_t = (B_{\mathrm{step}}/C_{\mathrm{b}})\eta_t$ attains this bound with equality at every update, so no schedule that meets the condition is larger at any update. By Lemma 1, the condition implies $\|\eta_t d_t\| \le \eta_t C_{\mathrm{b}} s_t \le B_{\mathrm{step}}\eta_t^2$. Setting $s_{\max} = (B_{\mathrm{step}}/C_{\mathrm{b}})\eta_{\max}$ recovers Eq. (3) with $\beta = 1$. □

### B.7 Cosine schedule ratios

With $s_t \equiv s_{\max}$,

$$\sum_t \eta_t s_t^2 = s_{\max}^2 S_1(T), \qquad \text{so the last term of Eq. (8) equals } 3C_{\mathrm{b}}^2 s_{\max}^2.$$

With $s_t = s_{\max}(\eta_t/\eta_{\max})^\beta$,

$$\sum_t \eta_t s_t^2 = s_{\max}^2 \eta_{\max}^{-2\beta} S_{1+2\beta}(T).$$

If $\eta_t \to 0$ and $S_1(T) \to \infty$, then $S_{1+2\beta}(T)/S_1(T) \to 0$. For any $\epsilon > 0$, choose $t_0$ such that $\eta_t^{2\beta} \le \epsilon$ for $t > t_0$. Then

$$S_{1+2\beta}(T) \le S_{1+2\beta}(t_0) + \epsilon S_1(T),$$

which proves the limit.

For the finite cosine family, let

$$h(\tau) = \frac{1 + \cos(\pi\tau)}{2}, \qquad \tau \in [0, 1].$$

The identity

$$\int_0^1 h(\tau)^\kappa \, d\tau = \frac{\Gamma(\kappa + \frac{1}{2})}{\sqrt{\pi}\,\Gamma(\kappa + 1)}$$

follows from the change of variables $x = \pi\tau/2$ and the standard cosine-power integral. Define the decay phase quantities

$$E^{\mathrm{dec}}_{\mathrm{var}}(\beta) = \frac{\Gamma(\beta + \frac{1}{2})}{\sqrt{\pi}\,\Gamma(\beta + 1)}, \qquad E^{\mathrm{dec}}_{\mathrm{bias}}(\beta) = \frac{2\Gamma(2\beta + \frac{3}{2})}{\sqrt{\pi}\,\Gamma(2\beta + 2)}. \tag{9}$$

Here $E^{\mathrm{dec}}_{\mathrm{var}}(\beta)$ is the average normalized noise variance over the cosine decay phase, and $E^{\mathrm{dec}}_{\mathrm{bias}}(\beta)$ is the LR-weighted average of its square, which is the continuous limit of the bias budget. In particular, $E^{\mathrm{dec}}_{\mathrm{var}}(1) = 1/2$, $E^{\mathrm{dec}}_{\mathrm{bias}}(1/2) = 3/4$, and $E^{\mathrm{dec}}_{\mathrm{bias}}(1) = 5/8$.

For a linear variance taper $s(\tau) = s_{\max}(1 - \tau)$,

$$\frac{\int_0^1 h(\tau)(1-\tau)^2 \, d\tau}{\int_0^1 h(\tau)\, d\tau} = \frac{1}{3} + \frac{2}{\pi^2},$$

while the step-weighted exposure is

$$\int_0^1 h(\tau)(1-\tau)\, d\tau = \frac{1}{4} + \frac{1}{\pi^2}, \qquad \int_0^1 h(\tau)^2 \, d\tau = \frac{3}{8}.$$

The total variance of the linear taper and $\beta = 1$ LR coupling is $1/2$ of fixed over the decay phase. Under linear LR with normalized LR $1 - \tau$, cosine taper variance gives $\int_0^1 (1-\tau)h(\tau)^2 \, d\tau / \int_0^1 (1-\tau)\, d\tau = 3/8 + 2/\pi^2$, while coupled variance gives $\int_0^1 (1-\tau)^3 \, d\tau / \int_0^1 (1-\tau)\, d\tau = 1/2$.

For the last quarter, direct integration gives

$$\frac{\int_{3/4}^1 h(\tau)^3 \, d\tau}{\int_{3/4}^1 h(\tau)\, d\tau} \approx 0.009322, \qquad \frac{\int_{3/4}^1 h(\tau)(1-\tau)^2 \, d\tau}{\int_{3/4}^1 h(\tau)\, d\tau} \approx 0.037278.$$

Also

$$\frac{\int_{3/4}^1 h(\tau)\, d\tau}{\int_0^1 h(\tau)\, d\tau} = \frac{\frac{1}{8} - \frac{\sqrt{2}}{4\pi}}{1/2} \approx 0.024921.$$

If Eq. (8) is applied only to the final-quarter sub-horizon, its step-size denominator contains just $0.024921$ of the full decay phase step-size mass. The two tail ratios above are therefore local schedule diagnostics. They do not by themselves provide a stronger whole-run stationarity guarantee.

### B.8 Exact quadratic least-squares example

**Proposition 4** (Exact residual for deterministic GD under Gaussian input noise)**.** *Let $X \in \mathbb{R}^{m \times d_w}$ be a design matrix, $y \in \mathbb{R}^m$ its target vector, and $\mathcal{L}(w) = \frac{1}{2m}\|Xw - y\|^2$. Write $\hat{\Sigma} = X^\top X/m$, assume $\lambda_{\min}(\hat{\Sigma}) > 0$, define $c_{\mathrm{LS}} = X^\top y/m$, and set $w^\star = \hat{\Sigma}^{-1} c_{\mathrm{LS}}$. Let $\lambda_{\max}$ be the largest eigenvalue of $\hat{\Sigma}$. Gaussian input noise $z \mapsto z + \xi$ with $\xi \sim \mathcal{N}(0, sI)$ gives $\mathcal{L}_s(w) = \mathcal{L}(w) + \frac{s}{2}\|w\|^2$. Consider deterministic gradient descent on the smoothed objective, $w_{t+1} = w_t - \eta_t \nabla \mathcal{L}_{s_t}(w_t)$. If $0 < \eta_t \le 1/(\lambda_{\max} + s)$ and $\sum_t \eta_t = \infty$, fixed $s_t \equiv s > 0$ converges to $w_s = (\hat{\Sigma} + sI)^{-1} c_{\mathrm{LS}}$ and $\|\nabla \mathcal{L}(w_t)\|^2 \to s^2 \|w_s\|^2$. If $0 \le s_t \le s_{\max}$, $\eta_t \le 1/(\lambda_{\max} + s_{\max})$, $s_t \to 0$, and $\sum_t \eta_t = \infty$, then $w_t \to w^\star$ and $\|\nabla \mathcal{L}(w_t)\| \to 0$.*

**Proof.** For one row $z \in \mathbb{R}^{d_w}$, target $y$, and loss $\ell(w, z, y) = \frac{1}{2}(w^\top z - y)^2$, Gaussian input noise $\xi \sim \mathcal{N}(0, sI)$ gives (Bishop, 1995)

$$\mathbb{E}_\xi \ell(w, z + \xi, y) = \ell(w, z, y) + \frac{s}{2}\|w\|^2.$$

Averaging over samples gives $\mathcal{L}_s(w) = \mathcal{L}(w) + \frac{s}{2}\|w\|^2$ and $\nabla\mathcal{L}_s(w) = \hat{\Sigma}w - c_{\mathrm{LS}} + sw$.

For fixed $s$, the unique minimizer is $w_s = (\hat{\Sigma} + sI)^{-1}c_{\mathrm{LS}}$. With $e_t^w = w_t - w_s$,

$$e_{t+1}^w = \big(I - \eta_t(\hat{\Sigma} + sI)\big)e_t^w.$$

If $\eta_t \le 1/(\lambda_{\max} + s)$, then

$$\|e_T^w\| \le \|e_1^w\| \exp\{-\lambda_{\min} S_1(T-1)\} \to 0$$

whenever $\sum_t \eta_t = \infty$. Since $\nabla\mathcal{L}(w_t) = \hat{\Sigma}e_t^w - sw_s$, the squared clean gradient norm converges to $s^2\|w_s\|^2$.

For vanishing $s_t$, let $e_t^w = w_t - w^\star$ and $\delta_t^w = \|e_t^w\|$. Then

$$\delta_{t+1}^w \le (1 - \eta_t\lambda_{\min})\delta_t^w + \eta_t s_t\|w^\star\|.$$

Fix $\epsilon > 0$ and choose $t_0$ so that $s_t\|w^\star\| \le \lambda_{\min}\epsilon$ for $t \ge t_0$. Then

$$(\delta_t^w - \epsilon)_+ \le (\delta_{t_0} - \epsilon)_+ \exp\{-\lambda_{\min}(S_1(t-1) - S_1(t_0-1))\} \to 0.$$

Hence $w_t \to w^\star$ and $\|\nabla\mathcal{L}(w_t)\| \to 0$. No assumption $\eta_t \to 0$ is needed for either case beyond the displayed stability and divergent-total-step conditions. □

### B.9 Antithetic variance reduction

**Proposition 5** (Antithetic variance reduction for the parameter gradient)**.** *Under Assumption 1, let*

$$g_\pm := \nabla_\phi \ell(\phi, E \pm \xi), \qquad \xi \sim \mathcal{N}(0, sI),$$

*and let* $J_{\phi E} = \nabla_E \nabla_\phi \ell(\phi, E)$*. Then*

$$\mathbb{E}\|g_+ - \mathbb{E}g_+\|^2 = s\|J_{\phi E}\|_F^2 + O(s^2), \qquad \mathbb{E}\left\|\frac{g_+ + g_-}{2} - \mathbb{E}\frac{g_+ + g_-}{2}\right\|^2 = O(s^2).$$

*The two estimators have the same expectation by Gaussian symmetry, and Lemma 1 bounds their common expected-gradient shift by* $C_{\mathrm{b}}s$*.*

*Proof.* Taylor expansion gives

$$G(E \pm \xi) = G(E) \pm J_{\phi E}\xi + \frac{1}{2}\mathcal{Q}[\xi, \xi] + \mathcal{R}_\pm(\xi),$$

with $\mathbb{E}\|\mathcal{R}_\pm(\xi)\|^2 = O(s^3)$. A single draw therefore has leading centered variance $s\|J_{\phi E}\|_F^2$, while the antithetic average cancels the linear term exactly and leaves $O(s^2)$ variance. Gaussian symmetry gives equality of the expectations. □ □

## C Adaptive-Optimizer Theory

This appendix accounts for how the perturbation-induced bias enters practical Adam when the denominator is formed from the current stochastic gradient. The result is a finite-horizon decomposition that preserves Adam momentum and adaptive normalization. Full optimizer convergence additionally depends on the boundedness, smoothness, and parameter conditions developed in existing Adam analyses (Défossez et al., 2022; Zhang et al., 2022; Li et al., 2023). Section C.1 first treats a history-measurable diagonal preconditioner. Section C.2 then retains the current-gradient second moment, both bias corrections, and first-moment momentum.

Throughout, $\mathcal{F}_t$ is the history available immediately before drawing the fresh minibatch and perturbation at step $t$, $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$, $g_t := \nabla\mathcal{L}(\phi_t)$, and

$$\hat{g}_t = g_t + d_t + \varepsilon_t, \qquad \mathbb{E}_t\varepsilon_t = 0, \qquad \|d_t\| \le C_{\mathrm{b}}s_t,$$

which is the case $b_t = 0$ of Eq. (2). A data-ordering scheme whose next-minibatch conditional mean differs from $g_t$ contributes an additional sampling-bias term. Section C.4 states this explicitly.

### C.1 PREDICTABLE DIAGONAL PRECONDITIONING

For a positive diagonal matrix $P$, write $\|x\|_P^2 = x^\top P x$. Consider

$$\phi_{t+1} = \phi_t - \eta_t P_t \hat{g}_t, \tag{10}$$

where $P_t$ is $\mathcal{F}_t$-measurable and $0 < p_{t,i} \le p_{\max}$.

**Proposition 6** (Predictable adaptive preconditioning)**.** *Assume that $\mathcal{L}$ is lower bounded and $L_0$-smooth, $\|d_t\| \le C_{\mathrm{b}} s_t$, and $\mathbb{E}_t\|\hat{g}_t\|^2 \le G_2^2$. Then*

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}\|g_t\|_{P_t}^2}{S_1(T)} \le \frac{2\Delta\mathcal{L}}{S_1(T)} + \frac{\sum_{t=1}^T \eta_t \mathbb{E}\|d_t\|_{P_t}^2}{S_1(T)} + L_0 p_{\max}^2 G_2^2 \frac{S_2(T)}{S_1(T)}. \tag{11}$$

*If $P_t \succeq p_{\min} I$ uniformly, then*

$$\min_{t \le T} \mathbb{E}\|g_t\|^2 \le \frac{1}{p_{\min}} \left[ \frac{2\Delta\mathcal{L}}{S_1(T)} + p_{\max} C_{\mathrm{b}}^2 \frac{\sum_t \eta_t s_t^2}{S_1(T)} + L_0 p_{\max}^2 G_2^2 \frac{S_2(T)}{S_1(T)} \right]. \tag{12}$$

*Proof.* Smoothness gives

$$\mathcal{L}(\phi_{t+1}) \le \mathcal{L}(\phi_t) - \eta_t \langle g_t, P_t \hat{g}_t \rangle + \frac{L_0 \eta_t^2}{2} \|P_t \hat{g}_t\|^2.$$

Conditioning on $\mathcal{F}_t$ is valid because $P_t$ is predictable. Hence

$$\mathbb{E}_t \langle g_t, P_t \hat{g}_t \rangle = \langle g_t, P_t (g_t + d_t) \rangle \ge \tfrac{1}{2}\|g_t\|_{P_t}^2 - \tfrac{1}{2}\|d_t\|_{P_t}^2,$$

and $\mathbb{E}_t\|P_t \hat{g}_t\|^2 \le p_{\max}^2 G_2^2$. Summation and telescoping prove Eq. (11). The Euclidean form follows from $p_{\min} I \preceq P_t \preceq p_{\max} I$. □

**Corollary 2** (Lagged second-moment normalization)**.** *Let $b_2 \in [0, 1)$ and $\epsilon_{\mathrm{A}} > 0$. Set $P_1 = \epsilon_{\mathrm{A}}^{-1} I$. Initialize $\psi_1 = 0$. For $t \ge 2$, define*

$$\psi_t = b_2 \psi_{t-1} + (1 - b_2) \hat{g}_{t-1}^{\odot 2}, \qquad \hat{\psi}_t = \frac{\psi_t}{1 - b_2^{t-1}}, \qquad P_t = \mathrm{diag}\left( \frac{1}{\sqrt{\hat{\psi}_{t,i}} + \epsilon_{\mathrm{A}}} \right).$$

*If $\|\hat{g}_t\|_\infty \le G_\infty$ almost surely, then $P_t$ is $\mathcal{F}_t$-measurable and*

$$(G_\infty + \epsilon_{\mathrm{A}})^{-1} I \preceq P_t \preceq \epsilon_{\mathrm{A}}^{-1} I.$$

*Here $x^{\odot 2}$ denotes the elementwise square of $x$. Thus Proposition 6 applies. This corollary covers a lagged adaptive denominator but omits first-moment momentum.*

**Lemma 2** (Perturbation component carried by momentum)**.** *Let $m_t = b_1 m_{t-1} + (1 - b_1)\hat{g}_t$ with $m_0 = 0$, and define*

$$b_t^{(d)} = (1 - b_1) \sum_{k=1}^t b_1^{t-k} d_k.$$

*Then*

$$\|b_t^{(d)}\| \le C_{\mathrm{b}} \bar{s}_t, \qquad \bar{s}_t = (1 - b_1) \sum_{k=1}^t b_1^{t-k} s_k. \tag{13}$$

*If $s_t \equiv s_{\max}$, then $\bar{s}_t = s_{\max}(1 - b_1^t)$. If $s_t \to 0$, then $\bar{s}_t \to 0$.*

*Proof.* Apply the triangle inequality and $\|d_k\| \le C_{\mathrm{b}} s_k$. The fixed-variance expression is the geometric sum. For $s_t \to 0$, split the convolution into a finite prefix, whose weight vanishes geometrically, and a tail whose entries are uniformly small. □

### C.2 Practical Adam with a directional adaptation remainder

Practical Adam (Kingma and Ba, 2015) forms its second moment from the same fresh stochastic gradient used in the numerator, so $P_t$ is not $\mathcal{F}_t$-measurable. In particular, $\mathbb{E}_t[P_t\varepsilon_t]$ cannot be replaced by $P_t\mathbb{E}_t\varepsilon_t$. We therefore retain this dependence explicitly.

For compact optimizer notation, this subsection writes $x_t := \phi_t$ and $f := \mathcal{L}$. Consider zero-decay Adam

$$\begin{aligned} m_t &= b_1 m_{t-1} + (1-b_1)\hat{g}_t, \\ \psi_t &= b_2\psi_{t-1} + (1-b_2)\hat{g}_t^{\odot 2}, \qquad \hat{\psi}_t = \frac{\psi_t}{1-b_2^t}, \\ P_t &= \operatorname{diag}\left(\frac{1}{\sqrt{\hat{\psi}_{t,i}} + \epsilon_{\mathrm{A}}}\right), \\ x_{t+1} &= x_t - \eta_t P_t \frac{m_t}{1-b_1^t}. \end{aligned} \tag{14}$$

Here $b_1$ and $b_2$ are the decay rates of the first and second moment estimates $m_t$ and $\psi_t$, $\hat{g}_t^{\odot 2}$ is the elementwise square of the gradient, $\epsilon_{\mathrm{A}} > 0$ is the stability constant, and $P_t$ is the resulting diagonal preconditioner. The main experiments use AdamW with zero decoupled weight decay, so Eq. (14) matches their adaptive-gradient algebra.

Set

$$p_{\max} := \epsilon_{\mathrm{A}}^{-1}, \qquad q_t := \frac{\eta_t}{1-b_1^t}, \qquad A_t := q_t P_t, \qquad \lambda_m := \frac{b_1}{1-b_1},$$

with $A_0 = 0$ and $q_0 = 0$. Define

$$z_t := x_t - \lambda_m A_{t-1} m_{t-1}, \qquad g_t^z := \nabla f(z_t), \qquad e_t^z := g_t^z - g_t,$$

where $g_t = \nabla f(x_t)$. The current-denominator quantities are

$$\begin{aligned} \bar{P}_t &:= \mathbb{E}_t P_t, & c_t &:= \mathbb{E}_t[P_t\varepsilon_t], \\ \upsilon_t &:= \lambda_m(A_t - A_{t-1})m_{t-1}, & U_t &:= A_t\hat{g}_t + \upsilon_t, \\ \chi_t &:= q_t c_t + \mathbb{E}_t\upsilon_t. \end{aligned} \tag{15}$$

Finally, let

$$\mathcal{G}_T := \sum_{t=1}^{T} q_t \mathbb{E}\|g_t\|_{\bar{P}_t}^2, \qquad S_T^q := \sum_{t=1}^{T} q_t, \qquad \mathcal{J}_T := -\sum_{t=1}^{T} \mathbb{E}\langle g_t^z, \chi_t\rangle. \tag{16}$$

Here $\mathcal{G}_T$ is the weighted sum of squared gradient norms measured in the preconditioner metric, and $S_T^q$ is the sum of the weights $q_t$. The signed remainder $\mathcal{J}_T$ contains current-gradient normalization covariance and the interaction between momentum and changes in the adaptive step.

**Proposition 7** (Finite-horizon bias accounting for practical Adam)**.** *Assume that $f$ is lower bounded and $L_0$-smooth, $\hat{g}_t = g_t + d_t + \varepsilon_t$ with $\mathbb{E}_t\varepsilon_t = 0$ and $\|d_t\| \le C_{\mathrm{b}} s_t$, and $\mathbb{E}\|\hat{g}_t\|^2 < \infty$ for every finite $t$. Then*

$$\frac{1}{2}\mathcal{G}_T \le \Delta + \frac{3}{2}\sum_t q_t\mathbb{E}\|d_t\|_{\bar{P}_t}^2 + \frac{3}{2}\sum_t q_t\mathbb{E}\|e_t^z\|_{\bar{P}_t}^2 + \frac{L_0}{2}\sum_t \mathbb{E}\|U_t\|^2 + \mathcal{J}_T, \tag{17}$$

*where $f_\star := \inf_x f(x) > -\infty$ and $\Delta = f(x_1) - f_\star$. No predictability assumption is made on $P_t$, and Eq. (17) has no step-size ceiling.*

*Proof.* Since $1 + \lambda_m = (1-b_1)^{-1}$,

$$z_{t+1} - z_t = -A_t\hat{g}_t - \lambda_m(A_t - A_{t-1})m_{t-1} = -U_t. \tag{18}$$

This identity keeps the fresh gradient multiplied by the current $A_t$. Conditioning on $\mathcal{F}_t$ gives

$$\mathbb{E}_t U_t = q_t\bar{P}_t(g_t + d_t) + \chi_t, \tag{19}$$

where the covariance $c_t = \mathbb{E}_t[P_t \varepsilon_t]$ is kept explicit and is not set to zero.

Smoothness at $z_t$ yields

$$\mathbb{E}_t f(z_{t+1}) \le f(z_t) - q_t \langle g_t^z, \bar{P}_t(g_t + d_t)\rangle - \langle g_t^z, \chi_t\rangle + \frac{L_0}{2}\mathbb{E}_t\|U_t\|^2.$$

For any positive-semidefinite $M$,

$$-\langle g + e, M(g+d)\rangle \le -\tfrac{1}{2}\|g\|_M^2 + \tfrac{3}{2}\|d\|_M^2 + \tfrac{3}{2}\|e\|_M^2,$$

which follows by applying Cauchy–Schwarz in the $M$ seminorm and Young's inequality separately to the three cross terms. Substitute this inequality, take total expectations, sum over $t$, and use $z_1 = x_1$ and $f(z_{T+1}) \ge f_\star$ to obtain Eq. (17). □

A simpler form replaces the finite-step second moments by one uniform moment bound.

**Corollary 3** (Second-moment practical-Adam bound)**.** *Under Proposition 7, additionally assume*

$$\sup_t \mathbb{E}\|\hat{g}_t\|^2 \le G_2^2 < \infty. \tag{20}$$

*Define*

$$C_{2,T} = \sum_t q_t^2, \qquad C_{3,T} = \sum_t q_t q_{t-1}^2, \qquad C_{\Delta,T} = \sum_t (q_t + q_{t-1})^2,$$

*and* $\mathcal{J}_T^+ = \max\{\mathcal{J}_T, 0\}$. *Then*

$$\begin{aligned}\frac{\mathcal{G}_T}{S_T^q} \le \frac{2\Delta}{S_T^q} + 3p_{\max}C_{\mathrm{b}}^2\frac{\sum_t q_t s_t^2}{S_T^q} + 3L_0^2\lambda_m^2 p_{\max}^3 G_2^2\frac{C_{3,T}}{S_T^q} \\ + 2L_0 p_{\max}^2 G_2^2\frac{C_{2,T} + \lambda_m^2 C_{\Delta,T}}{S_T^q} + \frac{2\mathcal{J}_T^+}{S_T^q}.\end{aligned} \tag{21}$$

*Proof.* The uncorrected first moment is a sub-convex combination of past optimizer-input gradients, so weighted Jensen gives $\mathbb{E}\|m_t\|^2 \le G_2^2$. Since $0 \prec P_t \preceq p_{\max} I$,

$$\mathbb{E}\|e_t^z\|_{\bar{P}_t}^2 \le L_0^2\lambda_m^2 p_{\max}^3 q_{t-1}^2 G_2^2,$$

$$\mathbb{E}\|\upsilon_t\|^2 \le \lambda_m^2 p_{\max}^2 (q_t + q_{t-1})^2 G_2^2,$$

and

$$\mathbb{E}\|U_t\|^2 \le 2p_{\max}^2 G_2^2\big[q_t^2 + \lambda_m^2(q_t + q_{t-1})^2\big].$$

Also $\mathbb{E}\|d_t\|_{\bar{P}_t}^2 \le p_{\max} C_{\mathrm{b}}^2 s_t^2$. Substitute these inequalities into Eq. (17), upper-bound $\mathcal{J}_T$ by $\mathcal{J}_T^+$, and divide by $S_T^q/2$. □

*Remark* (Interpretation of the constants). Corollary 3 is a mechanism-level bound. Its constants should not be interpreted as a numerical certificate. With the standard small Adam constant $\epsilon_{\mathrm{A}}$, $p_{\max} = 1/\epsilon_{\mathrm{A}}$ and its powers can be very large. Proposition 7, which retains the realized update terms directly, is the more faithful finite-horizon accounting statement.

**Fixed perturbations and perturbations coupled to the LR.** The perturbation-bias ratio in Eq. (21) is

$$R_T := \frac{\sum_t q_t s_t^2}{S_T^q}. \tag{22}$$

For fixed $s_t = s_{\max}$, $R_T = s_{\max}^2$. For $s_t = s_{\max}(\eta_t/\eta_{\max})^\beta$,

$$R_T = \frac{s_{\max}^2}{\eta_{\max}^{2\beta}}\frac{\sum_t \eta_t^{1+2\beta}/(1-b_1^t)}{\sum_t \eta_t/(1-b_1^t)}. \tag{23}$$

Thus $R_T \to 0$ for an infinite schedule with $\eta_t \to 0$, $\sum_t \eta_t = \infty$, and $\beta > 0$. This statement concerns the perturbation-bias term only. Perturbations can also change the gradient-noise distribution, the denominator, and $\mathcal{J}_T$.

If additionally $\eta_t \to 0$, $\sum_t \eta_t = \infty$, $s_t \to 0$, and

$$\mathcal{J}_T^+ = o(S_T^q), \tag{24}$$

then the weighted adaptive-metric stationarity measure $\mathcal{G}_T/S_T^q$ converges to zero. Under additional boundedness, total operator-norm variation implies the directional condition. The next subsection makes the relation explicit and shows that the converse can fail.

**Finite cosine schedules.** For $\eta_{t,T} = \eta_{\max} h(t/T)$ with $h(u) = (1+\cos(\pi u))/2$, the bias-correction prefix is negligible in the normalized Riemann-sum limit and

$$\frac{R_T}{s_{\max}^2} \longrightarrow \frac{\int_0^1 h(u)^{1+2\beta}\, du}{\int_0^1 h(u)\, du} = \frac{2\Gamma(2\beta+\frac{3}{2})}{\sqrt{\pi}\,\Gamma(2\beta+2)}. \tag{25}$$

The ratio is $3/4$ at $\beta = 1/2$ and $5/8$ at $\beta = 1$. A finite fixed-peak cosine family therefore gives whole-run attenuation, and the normalized perturbation-bias ratio remains nonzero.

### C.3 Directional versus total adaptive-step variation

The directional remainder has the equivalent form

$$\mathcal{J}_T = -\sum_{t=1}^{T} \mathbb{E}\left\langle g_t^z, (A_t - A_{t-1})(\varepsilon_t + \lambda_m m_{t-1})\right\rangle, \tag{26}$$

because $A_{t-1}$ is $\mathcal{F}_t$-measurable and $\mathbb{E}_t \varepsilon_t = 0$. Define the total operator variation

$$\mathcal{O}_T = \sum_t \mathbb{E}\|A_t - A_{t-1}\|_{\mathrm{op}}.$$

If $\|g_t^z\| \le G_{\max}$ and $\|\varepsilon_t + \lambda_m m_{t-1}\| \le B_\varepsilon$ almost surely, then

$$\mathcal{J}_T^+ \le G_{\max} B_\varepsilon\, \mathcal{O}_T. \tag{27}$$

Thus $\mathcal{O}_T = o(S_T^q)$ is one sufficient route to the directional condition under boundedness. The converse can fail: operator variation may persist in directions that do not interact with the descent geometry, as the following example shows.

**Proposition 8** (Persistent denominator variation)**.** *Let $f \equiv 0$ and let the scalar optimizer-input gradients be independent and equiprobable on $\{-g_{\mathrm{hi}}, -g_{\mathrm{lo}}, g_{\mathrm{lo}}, g_{\mathrm{hi}}\}$, where $0 < g_{\mathrm{lo}} < g_{\mathrm{hi}}$. Then $d_t = 0$ and the gradients are bounded and unbiased. For any fixed $b_2 < 1$, the Adam second-moment recursion has*

$$\liminf_t \mathbb{E}|P_t - P_{t-1}| > 0. \tag{28}$$

*For a regular diminishing sequence such as $\eta_t = t^{-\theta}$ with $0 < \theta \le 1$, this implies $\mathcal{O}_T = \Theta(S_T^q)$ up to the finite bias-correction transient. Yet $g_t = g_t^z = 0$, so $\mathcal{J}_T = 0$.*

*Proof.* Write $Y_t = \hat{g}_t^2 \in \{g_{\mathrm{lo}}^2, g_{\mathrm{hi}}^2\}$. The bias-corrected second moment obeys, for $t \ge 2$,

$$\hat{\psi}_t = (1 - W_t)\hat{\psi}_{t-1} + W_t Y_t, \qquad W_t = \frac{1-b_2}{1-b_2^t}.$$

Because $\hat{\psi}_{t-1} \in [g_{\mathrm{lo}}^2, g_{\mathrm{hi}}^2]$ after the first update,

$$\mathbb{E}_t|\hat{\psi}_t - \hat{\psi}_{t-1}| = \frac{W_t}{2}\left(|g_{\mathrm{lo}}^2 - \hat{\psi}_{t-1}| + |g_{\mathrm{hi}}^2 - \hat{\psi}_{t-1}|\right) \ge \frac{W_t(g_{\mathrm{hi}}^2 - g_{\mathrm{lo}}^2)}{2}.$$

For $u, v \in [g_{\mathrm{lo}}^2, g_{\mathrm{hi}}^2]$,

$$\left|\frac{1}{\sqrt{u}+\epsilon_{\mathrm{A}}} - \frac{1}{\sqrt{v}+\epsilon_{\mathrm{A}}}\right| = \frac{|u-v|}{(\sqrt{u}+\sqrt{v})(\sqrt{u}+\epsilon_{\mathrm{A}})(\sqrt{v}+\epsilon_{\mathrm{A}})} \ge \frac{|u-v|}{2g_{\mathrm{hi}}(g_{\mathrm{hi}}+\epsilon_{\mathrm{A}})^2}.$$

Since $W_t \to 1 - b_2 > 0$, Eq. (28) follows. For monotone $q_t$,

$$|q_t P_t - q_{t-1}P_{t-1}| \ge q_t|P_t - P_{t-1}| - p_{\max}|q_t - q_{t-1}|.$$

The first term sums on the order of $S_T^q$, whereas the second has bounded total variation. The reverse inequality gives the matching $O(S_T^q)$ upper bound. Since the objective is flat, $g_t = g_t^z = 0$ and the directional remainder vanishes identically. □

This proposition explains why $b_2 \approx 1$ is not by itself an asymptotic argument: a small but persistent per-step denominator change can still accumulate linearly. In the flat-objective construction, $\mathcal{O}_T/S_T^q$ stays nonzero while $\mathcal{J}_T = 0$, so the converse of Eq. (27) fails. Related Adam analyses likewise identify effective-step behavior as central to convergence and provide counterexamples when exponential second-moment averaging is uncontrolled (Reddi et al., 2018; Chen et al., 2019).

**Directional-remainder diagnostic.** Estimating $\mathcal{J}_T^+/S_T^q$ and $\mathcal{G}_T/S_T^q$ empirically would require a common normalization and frozen-state replication for the conditional quantities. A single observed trajectory supplies only a pathwise proxy, so we do not use these terms as empirical diagnostics in the main results.

### C.4 Weight decay, sampling, and clipping

**Decoupled weight decay.** For nonzero AdamW decay, the virtual update becomes $z_{t+1} - z_t = -U_t - \eta_t \lambda_t^{\mathrm{wd}} x_t$. Here $\lambda_t^{\mathrm{wd}}$ denotes the decoupled weight-decay coefficient. The resulting inner-product and curvature contributions should be bounded separately from $d_t$. The main runs use zero decay, so this term is absent from the main optimizer comparison.

**Sampling.** Independent uniform minibatches support the conditional-unbiasedness model for the empirical objective. Under epoch-wise random reshuffling, $\mathbb{E}_t \hat{g}_t = \nabla f(x_t) + d_t + b_t$, where $b_t$ is the sampling-bias term of Eq. (2) and need not scale with $s_t$.

**Gradient clipping.** Let $\mathrm{Clip}_{G_{\mathrm{clip}}}$ denote global-norm clipping at threshold $G_{\mathrm{clip}}$. If clipping is applied before the Adam moments, define

$$b_t^{\mathrm{clip}} = \mathbb{E}_t \, \mathrm{Clip}_{G_{\mathrm{clip}}}(\hat{g}_t^{\mathrm{raw}}) - \mathbb{E}_t \hat{g}_t^{\mathrm{raw}}.$$

Then $\|b_t^{\mathrm{clip}}\| \le \mathbb{E}_t(\|\hat{g}_t^{\mathrm{raw}}\| - G_{\mathrm{clip}})_+$. Clipping bounds the optimizer input while also changing its conditional mean, so this term sits outside the unclipped conditional-gradient model used in the main theorem.

## D Experimental Details

### D.1 Core training procedure

Algorithm 1 in Section 3 gives the core training loop.

**Architecture and optimizer.** LoRA (Hu et al., 2022) uses rank $r_{\mathrm{L}} = 16$ and scale 32 on attention projections $q, k, v, o$ and multilayer-perceptron (MLP) projections gate, up, and down. Unless a table states otherwise, the optimizer implementation is AdamW (Loshchilov and Hutter, 2019) with peak learning rate selected from $\{10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}\}$ on the selection split, zero weight decay, gradient clipping at 1.0, effective batch size 64, and three epochs. With weight decay set to zero, the decoupled shrinkage term is inactive in the reported experiments. Appendix Table 41 reports the selection check for the primary baseline families. Each selects $2 \times 10^{-4}$. Training uses bfloat16 (BF16) arithmetic and fixed-length sequence tensors. The main perturbation is applied to token embeddings after lookup and before the first transformer block.

**Training constants.** Two quantities in Algorithm 1 are fixed constants of the training setup and are the same for every method. Each optimizer update accumulates gradients over $M = 8$ microbatches of 8 examples, which gives the effective batch size of 64, and the accumulated gradient is clipped at global norm $C_{\mathrm{clip}} = 1.0$. One Gaussian embedding perturbation is drawn per optimizer update and shared by all $M$ microbatches of that update.

**Data partitions.** For each source corpus, 80% of examples are used for supervised updates, 5% form the optional controller split, 5% are reserved for model selection, and 10% form a held-out diagnostic split. The controller split is unused by the one-pass AdaNoiseSFT method. Hyperparameters are selected only on the selection split, and external benchmarks are excluded from selection. Member extraction probes are sampled from examples that participate in parameter updates. Validation loss and non-member extraction use held-out source examples. The clean-gradient audit uses a fixed 1,024-example subset from the training partition so that it tracks the clean training objective in Section 4. Before model selection, we sort stable training-example identifiers by the SHA-256

hash of the public salt `adanoise-audit-v1` concatenated with the identifier and retain the first 1,024 examples. The subset is never used for hyperparameter selection. It is partitioned into 16 audit blocks of 64 examples, and each block is evaluated with the same training microbatch size and gradient accumulation used for SFT. We average the unperturbed parameter gradients across all examples before taking the Euclidean norm.

**Learning-rate and perturbation schedules.** Warmup ends at update $T_{\text{warm}} = 100$. For $t > T_{\text{warm}}$, define normalized post-warmup progress

$$\tau_t = \frac{t - T_{\text{warm}}}{T - T_{\text{warm}}}.$$

The cosine LR uses $a(\tau) = (1 + \cos(\pi\tau))/2$, and the linear LR uses $a(\tau) = 1 - \tau$. The step LR uses multipliers $1$, $1/2$, $1/4$, and $1/8$, with drops at $\tau = 0.50$, $0.75$, and $0.90$, so its terminal plateau is $a(1) = 1/8$. Cosine, linear, and WSD reach a zero endpoint. WSD keeps $a(\tau) = 1$ through the first half of post-warmup training and then applies cosine decay over the second half. For the 1950-update reference schedule, WSD therefore uses 100 warmup, 925 stable, and 925 decay updates.

At $\beta = 1$, AdaNoiseSFT sets the normalized variance $v_t = \sigma_t^2/\sigma_{\max}^2$ equal to the normalized LR $a_t = \eta_t/\eta_{\max}$. The comparator family contains cosine taper, linear taper, step taper, and WSD taper time variance profiles defined only by $\tau$. Within each optimizer LR family, the time profile identical to $a_t$ is labeled AdaNoiseSFT because it duplicates the method's variance sequence exactly. Every distinct taper independently selects $\sigma_{\max} \in \{5, 10, 15, 20\}$ on the clean selection split before benchmark evaluation. For cosine, linear, and WSD, the final LR endpoint is zero and therefore makes no parameter update. For step LR, the final multiplier is $1/8$. At $\beta = 1$, LR coupling therefore leaves a terminal variance plateau $v_t = 1/8$ (equivalently $\sigma_t = \sigma_{\max}/\sqrt{8}$) rather than vanishing at the endpoint.

The shared-checkpoint tail branch begins at the first discrete update at or after $\tau = 3/4$. Tail-swap controls evaluate the matching constant at the first post-switch update and rescale the new branch to equal the pre-switch variance at that same discrete update, removing a variance jump. The WSD budget-matched linear diagnostic remains separate from the central four-shape comparison because matching total variance changes its peak amplitude.

**Perturbation sampling.** For the one-pass method, a single Gaussian embedding perturbation is sampled per optimizer update and reused across the eight accumulated microbatches. For paired schedule-shape comparisons, the same pair seed determines minibatch order, dropout randomness, benchmark decoding seeds, and a base Gaussian perturbation stream. Each schedule rescales that base stream according to its own variance rule. This common-random-number design targets the paired schedule contrast and can make paired intervals substantially narrower than marginal across-seed variation. For the fixed-length embedding tensor $E \in \mathbb{R}^{n_{\text{tok}} \times d_E}$, where $n_{\text{tok}}$ is the padded sequence length, the covariance is $\sigma_t^2/(n_{\text{tok}} d_E) I$. Attention and loss masks exclude padded positions from the response-token objective. The per-example control in Table 16 preserves the same marginal coordinate variance while changing across-example covariance. The primary NEFTune row follows independent per-example uniform draws while using the same fixed-length tensor dimension for direct marginal-variance matching. Thus the NEFTune comparison matches coordinate variance but not across-example covariance. Appendix Table 21 repeats NEFTune with example-specific non-padding token counts, and Appendix Table 16 provides the estimator-matched per-example Gaussian control. Paired antithetic variants additionally reuse the same non-embedding randomness across the positive and negative perturbation views.

**Baseline onset convention.** Fixed-amplitude embedding-noise baselines, including NEFTune and SymNoise, use the same LR-linked warmup ramp as the schedule controls and then hold their selected post-warmup amplitude constant. Under this shared onset convention, $\alpha = \sqrt{3}\,\sigma$ matches post-warmup variance per coordinate between NEFTune and Gaussian noise. The selected values are $\alpha = 10$ for NEFTune and $\sigma = 5$ for tuned fixed Gaussian, which are close after variance matching.

**Model selection.** Peak scales are selected by clean loss on the 5% selection split, with ties resolved toward the smaller scale. Fixed Gaussian noise, the linear taper, and AdaNoiseSFT are selected independently in Appendix Table 30. Appendix Table 51 evaluates $\sigma \in \{1, 2, 3, 5\}$ and selects $\sigma = 5$. The coupling exponent is selected from $\{0.5, 1, 1.5, 2\}$ by the same clean-loss criterion before external evaluation. In the four-shape identification experiment, the peak optimizer LR and every

distinct time-taper peak are re-selected on the same disjoint selection split for each optimizer schedule. Among the distinct non-matching taper families, comparator identity is then fixed by the lowest clean loss on that same split before any held-out benchmark evaluation. The reported AdaNoiseSFT rows select $\eta_{\max} = 2 \times 10^{-4}$ and $\sigma_{\max} = 10$. For UltraChat-50k and Tulu-3-100k, $\sigma_{\max}$ is re-selected within each corpus. Appendix Table 31 reports the local selection record. This prevents padding-rate differences from silently changing the chosen effective operating point. The primary same-peak fixed control nevertheless uses the AdaNoiseSFT peak $\sigma = 10$ because its purpose is to isolate schedule shape. The supplementary SGD comparisons use zero momentum and zero decay, identical minibatch orders, and a peak learning rate selected from $\{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$ on the same selection criterion. After warmup, the inverse-square-root schedule is $\eta_t = \eta_{\max}\sqrt{T_{\text{warm}}/t}$.

| Method | Selection grid |
|---|---|
| Fixed Gaussian peak scale | $\sigma \in \{1, 2, 3, 5, 10, 15, 20\}$ |
| Linear taper peak scale | $\sigma_{\max} \in \{5, 10, 15, 20\}$ |
| Cosine taper peak scale | $\sigma_{\max} \in \{5, 10, 15, 20\}$ |
| Step taper peak scale | $\sigma_{\max} \in \{5, 10, 15, 20\}$ |
| WSD taper peak scale | $\sigma_{\max} \in \{5, 10, 15, 20\}$ |
| AdaNoiseSFT peak scale | $\sigma_{\max} \in \{5, 10, 15, 20\}$ |
| AdaNoiseSFT exponent | $\beta \in \{0.5, 1, 1.5, 2\}$ |
| NEFTune uniform noise | $\alpha \in \{5, 10, 15, 20\}$ |
| SymNoise | $\alpha \in \{5, 10, 15, 20\}$ |
| NoiseFiT-style | $k \in \{2, 4, 8\}$, scale multiplier $c \in \{0.5, 1, 2\}$ |
| GIFT-SW-style | salient fraction $q_{\text{sal}} \in \{0.5, 1, 2\}\%$, $c \in \{0.5, 1, 2\}$ |
| R-Drop | $\lambda_{\text{RD}} \in \{0.5, 1.0, 2.0\}$ |
| BAR | $\rho_{\text{BAR}} \in \{0.01, 0.02, 0.05, 0.10\}$ |

Table 4: **Explicit selection grids used in this study.** Here $\lambda_{\text{RD}}$ is the R-Drop consistency weight and $\rho_{\text{BAR}}$ is the BAR regularization coefficient. Adapted baselines are tuned within the displayed search spaces under the same selection protocol.

| Method | Tuned parameter(s) | Selected value | Selection criterion |
|---|---|---|---|
| NEFTune | $\alpha$ | 10 | clean selection loss |
| SymNoise | $\alpha$ | 10 | clean selection loss |
| NoiseFiT-style | $k, c$ | 4, 1 | clean selection loss |
| GIFT-SW-style | $q_{\text{sal}}, c$ | 1%, 1 | clean selection loss |
| R-Drop | $\lambda_{\text{RD}}$ | 1.0 | clean selection loss |
| BAR | $\rho_{\text{BAR}}$ | 0.05 | clean selection loss |
| Fixed Gaussian | $\sigma$ | 5 | clean selection loss |
| Linear taper | $\sigma_{\max}$ | 10 | clean selection loss |
| AdaNoiseSFT | $\sigma_{\max}, \beta$ | 10, 1 | clean selection loss |

Table 5: **Selected hyperparameters for the primary Llama-3.1-8B + Alpaca-cleaned comparison.** Every value is chosen from the grid in Table 4 using the disjoint selection split.

**Generation and evaluation.** MMLU is evaluated zero-shot by choice likelihood. IFEval reports prompt-level strict accuracy. Inverse-IFEval uses the benchmark's LLM-judge evaluation on its English subset with the question set, judge, and rubric held fixed across methods. BBH uses a 3-shot direct-answer prompt without chain of thought, and GSM8K uses a zero-shot direct-answer prompt scored on the final numeric answer. These settings are compared only within this study because they differ from common chain-of-thought protocols. TruthfulQA reports multiple-choice 1 (MC1) accuracy. AlpacaEval-2 uses the official length-controlled AlpacaEval 2.0 configuration with the `weighted_alpaca_eval_gpt4_turbo` annotator (Dubois et al., 2024). MT-Bench uses FastChat single-answer grading, and Arena-Hard uses its 500-prompt two-position protocol (Zheng et al., 2023; Li et al., 2024b). We retain and release evaluator snapshots, prompt versions, language subsets, and raw judgments with the supplementary evaluation records. Absolute judge scores are interpreted within this evaluation stack because cross-paper levels can shift with templates, annotator snapshots, and decoding settings. As a calibration anchor, Meta-Llama-3-8B-Instruct scores 22.74 AE-2 and 20.28 Arena-H under this evaluation stack, against public values of 22.92 and 20.6, respectively. AE-2, MT-Bench, and Arena-Hard use temperature 0.7, top-$p$ 0.9, and at most 1024 new

tokens. Deterministic instruction-following, direct-answer, and extraction diagnostics use greedy decoding. Likelihood-based tasks do not sample answers.

**Failure-mode probes.** Carl@50 uses member probes selected before training from the 80% update partition and restricted to targets of at least 100 tokens (Carlini et al., 2023). Each probe supplies the instruction and first 50 target tokens, followed by greedy decoding of 50 tokens. Extraction is counted only when the decoded sequence exactly matches target positions 51–100. A matched held-out source sample provides the non-member control. The target sample size is 2,500 eligible examples per membership group. If fewer are available, all eligible examples are used and the actual denominator is reported. Eligibility is recomputed for each tokenizer.

The sharpness diagnostic is a one-step sharpness-aware minimization (SAM) proxy at $\rho_{\mathrm{SAM}} = 0.05$ (Foret et al., 2021). With $g = \nabla_\phi \mathcal{L}(\phi)$, we set $\hat{\epsilon} = \rho_{\mathrm{SAM}} g / \|g\|$ for $g \neq 0$ and report $\mathcal{L}(\phi + \hat{\epsilon}) - \mathcal{L}(\phi)$. For LoRA models, the value depends on the stored factor parameterization and is compared only between matched parameterizations. Training and diagnostic losses are clean response-token means, averaged within each example and then across examples. Gap is diagnostic loss minus training loss. The clean-gradient diagnostic is the Euclidean norm of the unperturbed SFT gradient averaged across the fixed 1,024-example training-audit subset. For the seed-77 endpoint comparison, uncertainty from audit composition is summarized with delete-one-audit-block jackknife standard errors.

**Checkpoint bias-to-step audit.** The schedule-shape mechanism audit is evaluated at training progress $\tau \in \{0.25, 0.45, 0.60, 0.70, 0.80, 0.86\}$. At each saved checkpoint, the model and optimizer are frozen and the same pre-registered 1,024-example training-audit subset is reused. Let $g_t^{\mathrm{clean}}$ be the clean gradient averaged across the audit subset. Within each 64-example audit block, we evaluate 16 antithetic Gaussian perturbation pairs at the schedule's current variance using the training microbatch size, accumulate each paired gradient over the whole block, and average the paired gradients before aggregation across blocks. Let $g_t^{\mathrm{pert}}$ denote the resulting perturbed gradient averaged across the audit subset. We estimate $d_t$ by $\hat{d}_t = g_t^{\mathrm{pert}} - g_t^{\mathrm{clean}}$ and report $u_t^{\mathrm{audit}} = \|\hat{d}_t\| / v_t$, normalized by its first-checkpoint value within each schedule, and summarize variance scaling by the coefficient of variation of $u_t^{\mathrm{audit}}$. The deterministic ratio $v_t/(\eta_t/\eta_{\max})$ is reported only as schedule geometry, not as empirical evidence. This audit protocol is fixed before checkpoint inspection and is not used for hyperparameter selection.

**Uncertainty.** Primary configurations use five matched seeds: 11, 22, 33, 44, and 55. Schedule-identification comparisons use ten matched seeds by adding 66, 88, 99, 111, and 122. The same ten identities are used for cosine, linear, step, and WSD LR. Seed 77 is reserved for the case study. Means, sample standard deviations, and Student-$t$ confidence intervals (CIs) are computed from unrounded seed-level values, with rounding applied only for display, and paired comparisons subtract observations with the same seed identity. Unless stated otherwise, reported seed CIs are pointwise effect-estimation intervals. The first five observations in each ten seeds comparison coincide with the corresponding primary seed identities whenever that configuration appears in the main matrix. The cross-setting cosine replications use the pre-specified linear taper comparator, and the cross-setting linear replications use the pre-specified cosine taper comparator. The step/WSD cross-backbone and cross-data replications reuse the same ten seed identities and repeat selection independently within each setting, selecting the displayed non-matching comparator by clean selection loss before test AE-2 evaluation. The prompt-sensitivity table resamples matched prompt-level evaluator records 2,000 times within each evaluation set and recomputes the paired metric difference for each bootstrap sample. For AE-2, each resample reruns the same length-control aggregation used for the reported score. Seed CIs and prompt-bootstrap intervals quantify different sources of uncertainty. The seed CIs condition on one paired decoding seed per training seed. Appendix Table 8 reports a separate generation-only sensitivity on one representative checkpoint pair, and those decoding replicates are not folded into the training-seed CIs.

**Software and timing.** The reference software stack is Transformers 4.52.4, PEFT 0.15.2, and TRL 0.18.1. Training jobs use one GPU from an eight-GPU NVIDIA H200 SXM node with 141 GB of memory per GPU. Timing uses BF16 arithmetic, synchronized start and stop boundaries, and an otherwise idle assigned device. It includes training passes and controller audits, excludes evaluation, and normalizes each seed to its matching vanilla run. Eight microbatches are accumulated per optimizer update. AdaNoiseSFT evaluates one perturbed forward/backward pass for each microbatch, while paired variants evaluate two sign-flipped passes. The controller adds two clean

audit-microbatch passes every 50 updates after update 200. The Qwen3-32B adaptation check is reported separately from the base-model scaling comparison. No additional memory-saving or timing claim is made here without a corresponding configuration record.

**Model-specific scope.** All LoRA experiments use the same target-module set $q, k, v, o, \text{gate}, \text{up}, \text{down}$ unless a table explicitly states otherwise. Cross-family differences are therefore not attributed to target-module scope. The 14B experiment is a larger-model comparison on the same Alpaca-cleaned setting. The 32B row uses a post-trained Qwen3-32B checkpoint and is treated as an adaptation check, reported separately from the base-model scale comparison. Data-regime controls use nested Tulu subsets while keeping selection and diagnostic examples disjoint from the training union. The data-regime sweep in Appendix Table 23 uses Llama-3.1-8B and a separately prepared 125k-example Tulu-3-SFT source pool. Under the same 80/5/5/10 split, its full training partition contains 100k examples. Table 11 instead starts from a 100k source pool and therefore contains 80k training examples. In the sweep, Fixed denotes fixed Gaussian noise at $\sigma = 10$.

**Adapted baseline scope.** The one-pass NoiseFiT-style row uses a signal-to-noise-ratio (SNR)-selected noise component with plain SFT loss, omitting the original hybrid loss and repeated noisy passes (Khadangi et al., 2025). The comparison grid selects the top $k$ layers by SNR from a frozen training calibration subset and multiplies the reference noise scale by $c$. The GIFT-SW-style row uses a salient-column mask and Gaussian noise in the complementary weight columns within a LoRA comparison (Zhelnin et al., 2025). The original GIFT-SW trains salient columns directly. These adapted rows are implementation-specific comparators and should not be interpreted as faithful reproductions of either published method. Calibration cost is included in the method-specific accounting reported with these adapted comparators.

### D.2 Noise normalization, schedule controls, and further replications

The normalization uses the full fixed-length tensor, including padded positions. Padding remains excluded by attention and loss masks, so those coordinates do not contribute to the response-token objective. One $n_{\text{tok}} \times d_E$ perturbation tensor is sampled per optimizer update, broadcast across the microbatch example axis, and reused across gradient accumulation microbatches in that update. This update-shared draw is the pre-specified one-pass implementation and makes each optimizer update correspond to one perturbation realization. Because the minibatch loss averages over examples, drawing independent noise for each example leaves the expected shift $d_t$ of the analysis unchanged and changes only the variance term $V_t$. Appendix Table 16 evaluates that per-example variant directly.

The main setting uses $n_{\text{tok}} = 512$ and $\sigma_{\max} = 10$. Under the same fixed-length Frobenius normalization, NEFTune uniform noise with scale $\alpha$ has variance per coordinate $\alpha^2/(3d_\xi)$, so variance matching gives $\alpha = \sqrt{3}\,\sigma$. Tuned NEFTune is selected independently on the selection split. Because the original NEFTune normalization uses the active token count, Appendix Table 21 repeats the comparison with example-specific non-padding normalization.

**Schedule controls.** Fixed-amplitude embedding-noise baselines use the common warmup ramp to their selected post-warmup amplitude. The fixed Gaussian noise control then holds variance at $\sigma^2_{\max}$. Appendix D gives baseline-specific onset and tuning details. For the LR-shape study, we define four variance profiles indexed by training progress over normalized post-warmup progress $\tau$: cosine taper, linear taper, step taper, and WSD taper. They use the same functional shapes as the corresponding optimizer schedules but depend only on $\tau$ and ignore the realized optimizer LR. In the empirical comparisons, the term time taper refers only to a profile that is distinct from AdaNoiseSFT under the optimizer LR being tested. Each distinct profile selects $\sigma_{\max}$ independently on the selection split. A profile numerically identical to Eq. (3) is labeled AdaNoiseSFT itself and is not counted as a separate comparator. The step LR uses post-warmup multipliers $1$, $1/2$, $1/4$, and $1/8$, with drops at normalized progress $0.50$, $0.75$, and $0.90$. WSD uses a stable half followed by cosine decay to zero. The step taper and WSD tapers copy these progress-dependent shapes while remaining independent of the optimizer state. These controls alter only the variance schedule. Peak scale is selected separately for AdaNoiseSFT, the linear taper, and fixed Gaussian noise families on the clean selection loss. The main row for fixed Gaussian noise uses $\sigma_{\max} = 10$, the peak selected by LR coupling, without retuning its amplitude. This control with the same peak changes schedule

shape while holding the peak fixed. Appendix Table 4 lists the peak grids, Tables 30 and 51 report peak-selection losses, and Appendix Table 18 reports exponent selection.

The cosine LR contrast is positive on AE-2 and IFEval in six model/corpus settings (Appendix Table 36). Step/WSD ordering also persists under one backbone change and one corpus change (Appendix Table 39). Values of $\beta$ far from 1 reduce quality, while the effect of values close to 1 is less certain. In the equal-area control, the variance profile keeps the same peak, endpoint, and average level as LR coupling while its shape moves away from the LR. The WSD control with the same total variance gives the same total variance with a higher peak, and schedules that swap only the final-quarter noise change both the early noise level and the amount of decay at the end (Appendix H). Per-example Gaussian draws add +0.15 AE-2 with 95% CI [0.06, 0.23]. Update-shared draws remain the pre-specified primary estimator.

Fixed-to-coupled gains remain positive under inverse-square-root SGD and cosine SGD and are largest with AdamW (Appendix Table 17). A schedule that follows the norm of the previous parameter update, instead of the scheduled LR, is nearly tied with LR coupling under the two SGD settings (14.6 vs. 14.5 and 15.2 vs. 15.1 AE-2). Under AdamW with cosine LR, it falls well below LR coupling (16.7 vs. 19.5), so the size of the realized updates does not explain the AdamW gain.

# E Additional Tables and Diagnostics

## E.1 Seed reporting

The matched seeds are 11, 22, 33, 44, and 55. Sample standard deviations (SD) use $n-1$ in the denominator. Mean and paired intervals use $t_{0.975,4}$ and are computed before display rounding.

| Configuration | 11 | 22 | 33 | 44 | 55 | Mean | SD |
|---|---|---|---|---|---|---|---|
| Llama / NEFTune | 15.23 | 14.88 | 14.87 | 15.70 | 14.79 | 15.09 | 0.38 |
| Llama / Fixed Gaussian noise ($\sigma = 10$) | 15.62 | 14.72 | 14.43 | 15.21 | 15.14 | 15.02 | 0.46 |
| Llama / Linear taper | 18.28 | 17.63 | 17.74 | 18.43 | 18.30 | 18.08 | 0.36 |
| Llama / AdaNoiseSFT | 19.60 | 19.30 | 18.81 | 20.15 | 19.63 | 19.50 | 0.49 |
| Mistral / NEFTune | 13.67 | 13.01 | 13.09 | 13.25 | 13.56 | 13.32 | 0.29 |
| Mistral / Fixed Gaussian noise ($\sigma = 10$) | 13.79 | 12.66 | 12.85 | 12.49 | 14.41 | 13.24 | 0.82 |
| Mistral / Linear taper | 15.91 | 14.42 | 15.99 | 15.74 | 16.07 | 15.63 | 0.69 |
| Mistral / AdaNoiseSFT | 17.42 | 16.58 | 16.83 | 16.48 | 17.76 | 17.02 | 0.55 |
| Qwen3-8B-Base / NEFTune | 18.82 | 19.61 | 19.50 | 19.01 | 18.05 | 19.00 | 0.62 |
| Qwen3-8B-Base / Fixed Gaussian noise ($\sigma = 10$) | 19.10 | 19.46 | 19.13 | 18.68 | 17.84 | 18.84 | 0.63 |
| Qwen3-8B-Base / Linear taper | 20.49 | 21.86 | 21.57 | 21.12 | 20.80 | 21.17 | 0.56 |
| Qwen3-8B-Base / AdaNoiseSFT | 22.00 | 22.56 | 22.39 | 21.97 | 21.18 | 22.02 | 0.53 |

Table 6: **AE-2 by matched seed: primary backbone comparisons.**

| Configuration | 11 | 22 | 33 | 44 | 55 | Mean | SD |
|---|---|---|---|---|---|---|---|
| Qwen3-14B-Base / NEFTune | 19.98 | 20.68 | 20.55 | 20.59 | 21.88 | 20.74 | 0.69 |
| Qwen3-14B-Base / Fixed Gaussian noise ($\sigma = 10$) | 20.38 | 21.04 | 20.14 | 20.31 | 20.93 | 20.56 | 0.40 |
| Qwen3-14B-Base / Linear taper | 22.13 | 22.11 | 22.03 | 22.39 | 23.89 | 22.51 | 0.78 |
| Qwen3-14B-Base / AdaNoiseSFT | 22.62 | 22.82 | 22.44 | 22.90 | 24.30 | 23.02 | 0.74 |
| Llama / Fixed $\to$ 0 at $\tau = 3/4$ (shared checkpoint) | 16.64 | 16.08 | 15.55 | 16.43 | 16.12 | 16.17 | 0.41 |
| Llama / Fixed $\to$ AdaNoiseSFT at $\tau = 3/4$ | 18.29 | 17.64 | 17.26 | 18.12 | 18.03 | 17.87 | 0.41 |
| Llama / Half-level variance | 16.38 | 14.96 | 15.03 | 15.78 | 15.66 | 15.56 | 0.59 |
| Llama / Linear taper | 18.28 | 17.63 | 17.74 | 18.43 | 18.30 | 18.08 | 0.36 |
| Llama / AdaNoiseSFT | 19.60 | 19.30 | 18.81 | 20.15 | 19.63 | 19.50 | 0.49 |
| Llama / WSD fixed | 13.82 | 14.25 | 13.46 | 14.94 | 15.24 | 14.34 | 0.75 |
| Llama / WSD linear taper | 14.40 | 15.03 | 14.38 | 15.62 | 16.22 | 15.13 | 0.80 |
| Llama / WSD budget-matched linear | 14.83 | 15.14 | 14.79 | 15.87 | 16.18 | 15.36 | 0.63 |
| Llama / WSD, AdaNoiseSFT | 15.92 | 16.29 | 15.83 | 17.23 | 17.50 | 16.55 | 0.76 |

Table 7: **AE-2 by matched seed: larger backbone and schedule interventions.**

| Decoding seed | Linear taper AE-2 | AdaNoiseSFT AE-2 | Paired Δ |
|---|---|---|---|
| 101 | 17.58 | 18.76 | +1.18 |
| 202 | 17.91 | 18.84 | +0.93 |
| 303 | 17.72 | 18.91 | +1.19 |
| 404 | 17.83 | 18.77 | +0.94 |
| 505 | 17.64 | 18.86 | +1.22 |
| Mean | 17.74 | 18.83 | +1.09 |

Table 8: **Generation-only repeated-decoding sensitivity on the seed-33 Llama/Alpaca checkpoint pair.** No retraining is performed. The paired-effect sample SD is 0.14. These decoding replicates are reported separately from the training-seed CIs.

| Comparison | Mean | Mean 95% CI | Paired Δ | Paired 95% CI |
|---|---|---|---|---|
| Llama: Fixed Gaussian noise ($\sigma = 10$) / NEFTune | 15.02 | [14.45, 15.60] | -0.07 | [-0.60, 0.46] |
| Llama: Linear taper / Fixed Gaussian noise ($\sigma = 10$) | 18.08 | [17.62, 18.53] | +3.05 | [2.72, 3.38] |
| Llama: AdaNoiseSFT / Linear taper | 19.50 | [18.89, 20.11] | +1.42 | [1.09, 1.76] |
| Mistral: Fixed Gaussian noise ($\sigma = 10$) / NEFTune | 13.24 | [12.22, 14.26] | -0.08 | [-0.83, 0.67] |
| Mistral: Linear taper / Fixed Gaussian noise ($\sigma = 10$) | 15.63 | [14.77, 16.48] | +2.39 | [1.44, 3.33] |
| Mistral: AdaNoiseSFT / Linear taper | 17.02 | [16.33, 17.70] | +1.39 | [0.65, 2.13] |
| Qwen3-8B-Base: Fixed Gaussian noise ($\sigma = 10$) / NEFTune | 18.84 | [18.06, 19.62] | -0.16 | [-0.48, 0.17] |
| Qwen3-8B-Base: Linear taper / Fixed Gaussian noise ($\sigma = 10$) | 21.17 | [20.48, 21.86] | +2.33 | [1.61, 3.04] |
| Qwen3-8B-Base: AdaNoiseSFT / Linear taper | 22.02 | [21.36, 22.68] | +0.85 | [0.34, 1.37] |
| Qwen3-14B-Base: Fixed Gaussian noise ($\sigma = 10$) / NEFTune | 20.56 | [20.06, 21.06] | -0.17 | [-0.87, 0.53] |
| Qwen3-14B-Base: Linear taper / Fixed Gaussian noise ($\sigma = 10$) | 22.51 | [21.54, 23.48] | +1.95 | [1.10, 2.79] |
| Qwen3-14B-Base: AdaNoiseSFT / Linear taper | 23.02 | [22.10, 23.93] | +0.51 | [0.36, 0.66] |
| Llama: Fixed $\to 0$ at $\tau = 3/4$ (shared checkpoint) / Fixed Gaussian noise ($\sigma = 10$) | 16.17 | [15.65, 16.68] | +1.14 | [0.95, 1.34] |
| Llama: Fixed $\to$ AdaNoiseSFT at $\tau = 3/4$ / Fixed $\to 0$ at $\tau = 3/4$ (shared checkpoint) | 17.87 | [17.35, 18.38] | +1.70 | [1.54, 1.87] |
| Llama: Half-level variance / Fixed Gaussian noise ($\sigma = 10$) | 15.56 | [14.84, 16.29] | +0.54 | [0.30, 0.77] |
| Llama: Linear taper / Fixed Gaussian noise ($\sigma = 10$) | 18.08 | [17.62, 18.53] | +3.05 | [2.72, 3.38] |
| Llama: AdaNoiseSFT / Linear taper | 19.50 | [18.89, 20.11] | +1.42 | [1.09, 1.76] |
| Llama: WSD linear taper / WSD fixed | 15.13 | [14.14, 16.12] | +0.79 | [0.59, 0.99] |
| Llama: WSD budget-matched linear / WSD fixed | 15.36 | [14.58, 16.15] | +1.02 | [0.80, 1.24] |
| Llama: WSD, AdaNoiseSFT / WSD, budget-matched linear | 16.55 | [15.60, 17.50] | +1.19 | [1.02, 1.36] |

Table 9: **AE-2 mean and paired-difference intervals.** The subtracted reference method follows the slash.

### E.2 Cross-corpus benchmark rows

| Method | AE-2↑ | MT-B↑ | Arena-H↑ | IFEval↑ | Carl@50 (%)↓ | Sharp.↓ |
|---|---|---|---|---|---|---|
| Vanilla SFT | 14.6 | 6.17 | 16.7 | 48.2 | 13.2 | 0.77 |
| NEFTune | 15.8 | 6.37 | 18.3 | 49.2 | 11.7 | 0.70 |
| Fixed Gaussian noise ($\sigma = 10$) | 16.5 | 6.32 | 18.0 | 48.9 | 11.4 | 0.71 |
| Linear taper | 17.3 | 6.59 | 19.8 | 50.3 | 10.5 | 0.65 |
| AdaNoiseSFT | 19.2 | 6.65 | 20.2 | 51.8 | 9.9 | 0.61 |
| Two-pass extension[‡] | 20.2 | 6.71 | 20.7 | 52.2 | 9.4 | 0.60 |

Table 10: **UltraChat-50k on Llama-3.1-8B.** The source split leaves 80% for supervised training.

| Method | AE-2↑ | MT-B↑ | Arena-H↑ | IFEval↑ | Carl@50 (%)↓ | Sharp.↓ |
|---|---|---|---|---|---|---|
| Vanilla SFT | 11.7 | 6.27 | 17.4 | 50.1 | 12.6 | 0.75 |
| NEFTune | 13.4 | 6.45 | 19.1 | 51.0 | 10.9 | 0.69 |
| Fixed Gaussian noise ($\sigma = 10$) | 13.9 | 6.45 | 19.3 | 50.6 | 10.6 | 0.68 |
| Linear taper | 15.9 | 6.66 | 20.5 | 52.1 | 9.8 | 0.61 |
| AdaNoiseSFT | 17.1 | 6.74 | 21.2 | 53.6 | 9.0 | 0.59 |
| Two-pass extension[‡] | 17.7 | 6.81 | 21.6 | 53.8 | 8.7 | 0.57 |

Table 11: **Tulu-3-100k on Llama-3.1-8B.** The source split leaves 80% for supervised training.

### E.3 Response length and compute structure

| Method | AE-2 length (tokens) | SFT passes | Relative cost |
|---|---|---|---|
| Vanilla SFT | 146 | 1 | 1.000 |
| NEFTune | 172 | 1 | 1.011 |
| SymNoise‡ | 162 | 2 | 1.945 |
| NoiseFiT-style | 158 | 1 | 1.029 |
| GIFT-SW-style | 161 | 1 | 1.024 |
| BAR | 157 | 1 | 1.054 |
| R-Drop‡ | 156 | 2 | 1.932 |
| Fixed Gaussian noise ($\sigma = 10$) | 168 | 1 | 1.015 |
| Tuned fixed Gaussian noise | 164 | 1 | 1.015 |
| Linear taper | 162 | 1 | 1.018 |
| AdaNoiseSFT | 160 | 1 | 1.018 |
| Two-pass extension‡ | 165 | 2 | 1.955 |

Table 12: **Mean response length, number of SFT model passes, and relative training cost.**

| Variant | Carl@50 (%)↓ | Sharp. | Rel. cost |
|---|---|---|---|
| *Llama-3.1-8B + Alpaca-cleaned* | | | |
| Vanilla SFT | 13.9 | 0.79 | 1.000 |
| NEFTune | 11.7 | 0.72 | 1.011 |
| Fixed Gaussian noise ($\sigma = 10$) | 11.2 | 0.70 | 1.015 |
| Tuned fixed Gaussian noise | 11.3 | 0.68 | 1.015 |
| Linear taper | 10.5 | 0.65 | 1.018 |
| **AdaNoiseSFT** | 9.1 | 0.58 | 1.018 |
| Two-pass extension‡ | 8.7 | 0.55 | 1.955 |
| *Qwen3-14B-Base + Alpaca-cleaned* | | | |
| Vanilla SFT | 10.8 | 0.67 | 1.000 |
| NEFTune | 9.6 | 0.62 | 1.010 |
| Fixed Gaussian noise ($\sigma = 10$) | 9.2 | 0.61 | 1.015 |
| Tuned fixed Gaussian noise | 9.0 | 0.61 | 1.013 |
| Linear taper | 8.7 | 0.58 | 1.019 |
| **AdaNoiseSFT** | 7.9 | 0.54 | 1.021 |
| Two-pass extension‡ | 7.6 | 0.53 | 1.954 |

Table 13: **Memorization, sharpness, and training-cost diagnostics for the primary Llama setting and the Qwen3-14B-Base transfer setting.** Relative cost is normalized to vanilla within each model. Quality metrics are reported in the main comparison and the corresponding seed-level and transfer tables.

### E.4 Additional scale checks

| Method | AE-2↑ | IFEval↑ | Carl@50 (%)↓ | Sharp.↓ |
|---|---|---|---|---|
| Vanilla SFT | 21.2 | 61.5 | 9.6 | 0.61 |
| NEFTune | 22.2 | 62.1 | 8.6 | 0.58 |
| Fixed Gaussian noise ($\sigma = 10$) | 22.7 | 61.5 | 8.5 | 0.57 |
| AdaNoiseSFT | 25.0 | 63.4 | 7.4 | 0.51 |

Table 14: **Qwen3-32B.** LoRA uses the post-trained Qwen3-32B checkpoint, so it is reported as an adaptation check and kept separate from the base-model scaling comparison.

| Method | AE-2↑ | IFEval↑ | Carl@50 (%)↓ | Sharp.↓ |
|---|---|---|---|---|
| Vanilla SFT | 5.6 | 30.5 | 17.4 | 1.04 |
| NEFTune | 9.1 | 32.1 | 15.1 | 0.96 |
| Fixed Gaussian noise ($\sigma = 10$) | 8.5 | 32.3 | 14.8 | 0.95 |
| AdaNoiseSFT | 12.8 | 35.9 | 12.7 | 0.82 |

Table 15: **SmolLM2-1.7B (Ben Allal et al., 2025).** Full-parameter adaptation with length 1024. Sharpness is interpreted within this parameterization.

### E.5 Full ablation matrix and optimizer checks

| Method | AE-2↑ | Sharp.↓ | Cost↓ |
|---|---|---|---|
| AdaNoiseSFT | 19.5 | 0.58 | 1.018 |
| Per-example Gaussian draws | 19.6 | 0.60 | 1.017 |
| + antithetic pair | 19.9 | 0.57 | 1.940 |
| + feedback controller, no pair | 19.8 | 0.59 | 1.026 |
| Independent two-draw average | 19.7 | 0.59 | 1.941 |
| Full without KL | 19.8 | 0.57 | 1.952 |
| Two-pass extension[‡] | 20.5 | 0.55 | 1.955 |

Table 16: **One-pass and extension ablations.** Per-example Gaussian draws keep the same marginal variance but change the across-example covariance. The independent two-draw average uses two Gaussian draws and no Kullback–Leibler (KL) consistency term.

| Optimizer / LR | Schedule signal | Fixed AE-2 | Coupled AE-2 | Paired $\Delta$ [95% CI] |
|---|---|---|---|---|
| SGD / inverse square root | nominal LR | 13.5 | 14.5 | +1.04 [0.56, 1.52] |
| SGD / inverse square root | lagged update norm | 13.5 | 14.6 | +1.09 [0.72, 1.46] |
| SGD / cosine | nominal LR | 13.8 | 15.1 | +1.33 [0.92, 1.73] |
| SGD / cosine | lagged update norm | 13.8 | 15.2 | +1.38 [1.02, 1.74] |
| AdamW / cosine | nominal LR | 15.0 | 19.5 | +4.48 [4.04, 4.91] |
| AdamW / cosine | lagged update norm | 15.0 | 16.7 | +1.71 [1.38, 2.04] |

Table 17: **Within-optimizer matched-setting comparisons.** The lagged realized-update-norm rows use the previous update magnitude as a causal schedule signal and normalize it to the same peak variance. Intervals are 95% intervals.

### E.6 One-pass method sensitivity

| Exponent | Selection loss | AE-2 | Sharp. |
|---|---|---|---|
| $\beta = 0.5$ | 1.636 | 17.8 | 0.61 |
| $\beta = 1$ (selected) | 1.599 | 19.5 | 0.58 |
| $\beta = 1.5$ | 1.622 | 19.4 | 0.55 |
| $\beta = 2$ | 1.642 | 18.3 | 0.58 |

Table 18: **Exponent selection at $\sigma_{\max} = 10$.** Selection uses clean loss on the selection split. Benchmark columns are reported only after the choice.

| Comparison | Mean AE-2 ($\beta = 1$) | Mean AE-2 ($\beta = 1.5$) | Paired $\Delta$ [95% CI] |
|---|---|---|---|
| Ten matched seeds | 19.83 | 19.67 | +0.16 [-0.18, 0.49] |

Table 19: **Direct $\beta = 1$ versus $\beta = 1.5$ sensitivity on the same ten seed identities used by the central schedule confirmation.** The interval includes zero, so the downstream benchmark does not sharply separate the two nearby exponents.

| Method | AE-2↑ | Sharp.↓ |
|---|---|---|
| $\beta = 0.5$, $E_{\mathrm{var}}^{\mathrm{dec}}$-matched | 18.2 | 0.61 |
| AdaNoiseSFT | 19.5 | 0.58 |

Table 20: **Exposure-matched shape check.** For $\beta = 0.5$, matching $E_{\mathrm{var}}^{\mathrm{dec}}$ to $\beta = 1$ multiplies peak variance by $\pi/4$, equivalently multiplying peak standard deviation by $\sqrt{\pi}/2$.

| Method | AE-2↑ | IFEval↑ | Arena-H↑ | Carl@50 (%)↓ |
|---|---|---|---|---|
| NEFTune | 15.1 | 46.2 | 16.9 | 11.7 |
| NEFTune (non-padding norm) | 14.5 | 45.8 | 16.1 | 11.8 |

Table 21: **NEFTune normalization robustness on Llama-3.1-8B.** The primary row uses the common fixed-length tensor dimension. The non-padding row recomputes the normalization from each example's active token count.

### E.7 Non-member extraction control

| Method | Eligible probes per seed | Exact matches by seed | Mean (%) |
|---|---|---|---|
| Vanilla SFT | 2500 | 12, 9, 14, 11, 13 | 0.47 |
| NEFTune | 2500 | 10, 13, 9, 11, 8 | 0.41 |
| Fixed Gaussian noise ($\sigma = 10$) | 2500 | 11, 9, 10, 12, 8 | 0.40 |
| AdaNoiseSFT | 2500 | 9, 10, 8, 11, 7 | 0.36 |
| Two-pass extension[‡] | 2500 | 8, 7, 9, 6, 8 | 0.30 |

Table 22: **Non-member exact-continuation control with a common per-seed denominator.**

### E.8 Data-regime sweep

| Training examples | Epochs | Fixed AE-2 | Coupled AE-2 | Paired 95% CI |
|---|---|---|---|---|
| 10,000 | 1 | 9.1 | 9.8 | [0.37, 1.05] |
| 10,000 | 3 | 10.3 | 11.7 | [0.97, 1.84] |
| 10,000 | 5 | 10.6 | 12.3 | [1.26, 2.14] |
| 50,000 | 1 | 11.7 | 13.0 | [0.85, 1.71] |
| 50,000 | 3 | 13.0 | 15.1 | [2.00, 2.35] |
| 50,000 | 5 | 13.2 | 15.7 | [2.33, 2.67] |
| 100,000 | 1 | 12.9 | 14.8 | [1.56, 2.29] |
| 100,000 | 3 | 14.2 | 17.5 | [3.02, 3.68] |
| 100,000 | 5 | 14.4 | 17.8 | [3.11, 3.68] |

Table 23: **Nested Tulu training subsets on Llama-3.1-8B from a separate 125k-example Tulu-3-SFT source pool.** With the same 80/5/5/10 partition rule, the full training partition contains 100k examples, whereas Table 11 starts from a 100k source pool and therefore contains 80k training examples. Fixed denotes fixed Gaussian noise at $\sigma = 10$. The sweep varies distinct-example count and repeated exposure.

| Policy | Diag. loss | Gap | Clean grad. | Carl@50 (%) | Sharp. | AE-2 |
|---|---|---|---|---|---|---|
| *(a) Shared checkpoint at $\tau = 3/4$ of cosine decay* | | | | | | |
| Shared checkpoint | 1.694 | 0.314 | 0.367 | 10.8 | 0.68 | 14.6 |
| Fixed Gaussian noise ($\sigma = 10$) | 1.707 | 0.341 | 0.420 | 11.2 | 0.70 | 15.0 |
| Fixed $\to 0$ at $\tau = 3/4$ (shared checkpoint) | 1.662 | 0.304 | 0.320 | 10.5 | 0.65 | 16.2 |
| Fixed $\to$ AdaNoiseSFT at $\tau = 3/4$ | 1.633 | 0.290 | 0.337 | 10.0 | 0.62 | 17.9 |
| *(b) Full-run schedule controls in the same model-data setting* | | | | | | |
| Fixed Gaussian noise ($\sigma = 10$) | 1.707 | 0.341 | 0.420 | 11.2 | 0.70 | 15.0 |
| Half-level variance | 1.669 | 0.322 | 0.372 | 10.7 | 0.67 | 15.6 |
| Linear taper | 1.653 | 0.304 | 0.354 | 10.5 | 0.65 | 18.1 |
| **AdaNoiseSFT** | 1.610 | 0.255 | 0.342 | 9.1 | 0.58 | 19.5 |

Table 24: **Schedule controls on Llama-3.1-8B + Alpaca-cleaned.** The checkpoint row precedes the three matched final-quarter branches.

# F Extensions: Antithetic Pair and Online Controller

The main text studies the one-SFT-pass AdaNoiseSFT rule. This appendix defines the optional antithetic and feedback components that form the AdaNoiseSFT + two-pass extension. They are reported separately because they increase compute and are not required by the one-pass schedule.

## F.1 Definitions

**Antithetic pair and consistency KL.** At update $t$, sample one $\xi_t \sim \mathcal{N}(0, \sigma_t^2/(n_{\text{tok}} d_E)\, I)$ and reuse it with opposite signs. Two model passes use $E_t + \xi_t$ and $E_t - \xi_t$, producing next-token distributions $\pi_+$ and $\pi_-$. The loss is

$$\begin{aligned}\mathcal{L}_t^{\text{pair}} &= \tfrac{1}{2}\big[\mathcal{L}_{\text{SFT}}(\phi_t, \mathcal{B}_t, E_t + \xi_t) + \mathcal{L}_{\text{SFT}}(\phi_t, \mathcal{B}_t, E_t - \xi_t)\big] \\ &\quad + \tfrac{\lambda_{\text{cons}}}{2}\big[\text{KL}(\pi_+ \| \pi_-) + \text{KL}(\pi_- \| \pi_+)\big], \qquad \lambda_{\text{cons}} = 0.1.\end{aligned}$$

Here $\lambda_{\text{cons}}$ is the consistency-loss weight, and $\mathcal{L}_{\text{cons}} = \frac{1}{2}[\text{KL}(\pi_+ \| \pi_-) + \text{KL}(\pi_- \| \pi_+)]$ is averaged over unmasked response tokens and examples. The symmetric construction treats the perturbations as exchangeable views. Proposition 5 shows cancellation of the leading random term for the paired SFT-gradient estimator. The consistency contribution is separate. Under locally bounded softmax derivatives and token probabilities bounded away from zero, its expected gradient is $O(s_t)$ and can be absorbed into the perturbation-bias envelope through a larger $C_{\text{b}}$.

**Online gradient-alignment controller.** Every $N_{\text{ctrl}} = 50$ updates after controller warmup $T_{\text{ctrl}} = 200$, the extension computes noise-free gradients on one current training audit microbatch and one fresh controller microbatch,

$$g_t^{\text{train}} = \nabla_\phi \mathcal{L}_{\text{SFT}}(\phi_t, \mathcal{B}_t, E_t), \qquad g_t^{\text{ctrl}} = \nabla_\phi \mathcal{L}_{\text{SFT}}(\phi_t, \mathcal{C}_t, E_t^{\text{ctrl}}), \qquad \mathcal{C}_t \sim \mathcal{D}_{\text{ctrl}}.$$

Here $E_t^{\text{ctrl}}$ is the unperturbed embedding tensor of the controller microbatch $\mathcal{C}_t$ drawn from the controller split $\mathcal{D}_{\text{ctrl}}$. Define the disagreement score $\zeta_t = -\cos(g_t^{\text{train}}, g_t^{\text{ctrl}})$. With the lagged exponential moving mean $\mu_{t-1}$ and lagged deviation $\sigma_{t-1}^{\zeta} = \sqrt{\max(\nu_{t-1} - \mu_{t-1}^2, 0)}$,

$$\widehat{\zeta}_t = \frac{\zeta_t - \mu_{t-1}}{\sigma_{t-1}^{\zeta} + \epsilon_c},$$

and define the controller multiplier

$$\omega_t = \text{sigmoid}(c_{\text{ctrl}} \widehat{\zeta}_t), \qquad c_{\text{ctrl}} = 1.5.$$

The applied variance is

$$\sigma_t^2 = \sigma_{\max}^2 \left(\frac{\eta_t}{\eta_{\max}}\right)^{\beta} \omega_t.$$

The nominal scale $\sigma_{\max}$ is unchanged, but $\omega_t \in (0, 1)$ reduces the realized variance whenever feedback is active. We therefore separate state-dependent feedback from variance attenuation with an open-loop replay control.

**Matched-envelope replay.** For each feedback run, we record the realized time-indexed multiplier sequence $\{\omega_t\}$. A paired replay run starts from the same initialization and uses the same future mini-batch order but applies that recorded multiplier sequence without recomputing gradient alignment. The replay therefore matches the feedback run's perturbation envelope step by step while removing dependence on the current model state. The paired feedback-minus-replay contrast isolates online adaptation after matching the realized noise envelope.

## F.2 Pair/controller extension algorithm

Algorithm 2 uses update-level shorthand: each SFT gradient averages the same eight accumulation microbatches as Algorithm 1, with one shared perturbation and global-norm clipping before the optimizer step. It makes the additional work explicit. Disabling the controller sets $\omega_t = 1$ and recovers the one-pass scale coupled to the LR. Enabling the antithetic pair replaces one perturbed model pass with two sign-flipped passes.

**Algorithm 2. Optional pair and controller**

**Require:** $\phi_1, o_1$, schedule $\eta(t)$, $\sigma_{\max}, \beta, T$, two-pass extension flags, $N_{\mathrm{ctrl}}, T_{\mathrm{ctrl}}$
**Ensure:** $\phi_{T+1}$
$\mu_0 \leftarrow 0, \nu_0 \leftarrow 1, \omega_0 \leftarrow 1, \gamma \leftarrow 0.9, \epsilon_c \leftarrow 10^{-6}$
**for** $t = 1, \ldots, T$ **do**
  obtain $\mathcal{B}_t$, set $\eta_t \leftarrow \eta(t)$ and $\omega_t \leftarrow \omega_{t-1}$
  **if** controller enabled, $t \geq T_{\mathrm{ctrl}}$, and $t \bmod N_{\mathrm{ctrl}} = 0$ **then**
    compute $g_t^{\mathrm{train}}$ and $g_t^{\mathrm{ctrl}}$ on the two audit microbatches
    **if** $\|g_t^{\mathrm{train}}\| \, \|g_t^{\mathrm{ctrl}}\| > 0$ **then**
      $\zeta_t \leftarrow -\cos(g_t^{\mathrm{train}}, g_t^{\mathrm{ctrl}})$
    **else**
      $\zeta_t \leftarrow 0$
    **end if**
    $\widehat{\zeta}_t \leftarrow (\zeta_t - \mu_{t-1}) / (\sqrt{\max(\nu_{t-1} - \mu_{t-1}^2, 0)} + \epsilon_c)$
    $\omega_t \leftarrow \mathrm{sigmoid}(1.5\widehat{\zeta}_t)$
    $\mu_t \leftarrow \gamma\mu_{t-1} + (1-\gamma)\zeta_t$, and $\nu_t \leftarrow \gamma\nu_{t-1} + (1-\gamma)\zeta_t^2$
  **else**
    $\mu_t \leftarrow \mu_{t-1}$, and $\nu_t \leftarrow \nu_{t-1}$
  **end if**
  $\sigma_t^2 \leftarrow \sigma_{\max}^2 (\eta_t / \eta_{\max})^{\beta} \omega_t$, then draw update-level $\xi_t$
  **if** pair enabled **then**
    $\hat{g}_t \leftarrow \nabla_\phi [\frac{1}{2}\mathcal{L}_{\mathrm{SFT}}(E_t + \xi_t) + \frac{1}{2}\mathcal{L}_{\mathrm{SFT}}(E_t - \xi_t) + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}}]$
  **else**
    $\hat{g}_t \leftarrow \nabla_\phi \mathcal{L}_{\mathrm{SFT}}(E_t + \xi_t)$
  **end if**
  $g_t^{\mathrm{clip}} \leftarrow \mathrm{Clip}(\hat{g}_t, 1.0)$
  $(\phi_{t+1}, o_{t+1}) \leftarrow \mathsf{OptStep}(\phi_t, o_t, g_t^{\mathrm{clip}}, \eta_t)$
**end for**
**return** $\phi_{T+1}$

### F.3 Controller and estimator controls

| Method | AE-2↑ | Cost↓ |
|---|---|---|
| NEFTune | 15.1 | 1.011 |
| NEFTune (85 percent training) | 15.1 | 1.072 |
| SymNoise‡ | 15.3 | 1.945 |
| SymNoise (85 percent training) | 16.0 | 2.058 |
| AdaNoiseSFT | 19.5 | 1.018 |
| Two-pass extension‡ | 20.5 | 1.955 |
| Training-data controller | 20.0 | 1.949 |

Table 25: **Controller-data controls.** The 85 percent rows use the controller partition as additional supervised data, so their training cost scales with the larger supervised set.

| Method | AE-2↑ | Sharp.↓ | Cost↓ |
|---|---|---|---|
| One-way KL | 19.6 | 0.57 | 1.951 |
| Two-pass extension‡ | 20.5 | 0.55 | 1.955 |
| Negative dot product | 19.8 | 0.59 | 1.950 |
| Raw cosine threshold | 19.8 | 0.59 | 1.951 |

Table 26: **Alternative consistency and controller choices under the same full-extension configuration.**

Table 18 varies one-pass hyperparameters, while Table 26 changes extension components under the same compute accounting.

**Estimator comparison.** The antithetic calculation in Proposition 5 conditions on a fixed mini-batch and shared non-embedding randomness. The paired SFT views therefore use identical dropout masks when isolating this estimator effect. Independent dropout would add another variance source.

The paired-versus-independent two-draw comparison in Table 16 keeps two SFT model passes and disables consistency KL in both estimators.

**Quadratic-gradient illustration.** As a reproducible estimator check, let $q(\xi) = a + J\xi + \frac{1}{2}H(\xi^{\odot 2})$, with output dimension 8 and perturbation dimension 12. Fixed Gaussian coefficient arrays and 20,000 standard-normal draws use seed 24631. This construction isolates embedding randomness with all other quantities fixed, and the paired estimator cancels $J\xi$ exactly.

| Estimator | Variance slope | CV at scale $\lambda = 0.1$ | CV at scale $\lambda = 0.007$ |
|---|---|---|---|
| Single draw | 2.001 | 0.4334 | 0.03052 |
| Antithetic pair | 4.000 | 0.0347 | 0.00017 |

Table 27: **Quadratic-gradient estimator illustration.** Slope is the fitted log-variance versus log-standard-deviation slope. The coefficient of variation (CV) is the square root of total coordinate variance divided by the norm of the mean gradient.

| Variant | AE-2 | Δ vs AdaNoiseSFT [95% CI] | Sharp. | Cost |
|---|---|---|---|---|
| AdaNoiseSFT | 19.50 | N/A | 0.576 | 1.018 |
| Per-example draw | 19.64 | +0.15 [0.06, 0.23] | 0.598 | 1.017 |
| Antithetic pair | 19.88 | +0.38 [-0.06, 0.83] | 0.573 | 1.940 |
| Feedback controller | 19.76 | +0.26 [0.16, 0.36] | 0.588 | 1.026 |
| Open-loop matched envelope | 19.49 | -0.01 [-0.10, 0.08] | 0.595 | 1.021 |
| Independent two-draw | 19.72 | +0.23 [-0.09, 0.54] | 0.595 | 1.941 |
| Full without KL | 19.76 | +0.26 [-0.01, 0.53] | 0.571 | 1.952 |
| Pair/controller, fixed envelope | 19.43 | -0.07 [-0.80, 0.66] | 0.629 | 1.950 |
| Controller, fixed envelope | 19.48 | -0.02 [-0.36, 0.31] | 0.660 | 1.024 |
| One-way KL | 19.59 | +0.09 [-0.39, 0.57] | 0.572 | 1.951 |
| Negative dot product | 19.79 | +0.29 [-0.18, 0.76] | 0.586 | 1.950 |
| Raw cosine threshold | 19.80 | +0.30 [0.10, 0.49] | 0.589 | 1.951 |
| Pair/controller extension | 20.46 | +0.96 [0.52, 1.39] | 0.553 | 1.955 |

Table 28: **Complete component ablation with five paired seeds.** The open-loop row replays a matched noise envelope without using the current alignment signal. Intervals are 95% CIs. N/A marks the reference row of the paired difference.

**Matched-envelope feedback contrast.** Live feedback minus open-loop replay is +0.27 [0.19, 0.36] AE-2, directly comparing the two trajectories after matching the realized noise envelope.

### F.4 Hyperparameter sensitivity

The reference extension uses $\sigma_{\max} = 10$, $\beta = 1$, $c_{\text{ctrl}} = 1.5$, $\lambda_{\text{cons}} = 0.1$, $N_{\text{ctrl}} = 50$, and controller warmup 200. Table 18 varies one-pass schedule parameters, whereas Table 26 changes extension components. The open-loop replay, controller audits, and paired-pass costs remain explicit in the compute comparisons.

## G Extended Limitations

**Scope of the evidence.** The complete $4 \times 4$ studies use Alpaca-cleaned with Llama-3.1-8B and Mistral-7B-v0.3. Outside that setting, the comparison with the selected taper is repeated under cosine LR in five more settings of model and corpus, under linear LR in four, and under step and WSD in three each (Appendix Table 40). These replications span four model configurations and three corpora, and they do not provide a full matrix in the other settings. The dropout study uses one model and one corpus, and the step LR ends at a nonzero plateau, so it tests schedule shape rather than a vanishing bias. The perturbation theory covers the LoRA setting with frozen embeddings. It relies on the tail condition and the local smoothness assumptions of Section 4 and Appendix B. The Adam bound does not prove that Adam converges. Evaluation uses English data only, scores from LLM judges depend on the pipeline, and Appendix G lists further limits.

**Optimization assumptions.** The clean-objective result assumes smoothness, finite moments, local derivative bounds along the trajectory, and conditionally centered stochastic gradients. At the working peak $\sigma_{\max} = 10$, Assumption 1 is a trajectory-local bounded-region condition. The general lemma guarantees a bias envelope that contracts with variance. The fixed-noise persistence result additionally uses the tail-nondegeneracy assumption in Proposition 1, including $Ks_{\max} \leq h_0/2$. The $O(s_t^2)$ remainder uses the fourth-order derivative envelope in Assumption 1. The perturbation lemma and stationarity theorem cover the main LoRA setting with a frozen input-embedding map. The full parameter SmolLM2 check is empirical and lies outside that theorem scope. Clean gradient norm is used only as a mechanism diagnostic. Generalization is evaluated by downstream metrics. Epoch-wise reshuffling contributes a separate sampling-bias term, while gradient clipping can change the conditional mean before adaptive moments are formed. Appendix C preserves current-gradient normalization and momentum through a finite-horizon directional remainder. Controlling that remainder remains necessary for a full asymptotic Adam result.

**Schedule identification.** The central identification study changes the optimizer schedule across cosine, linear, step, and WSD LR while keeping model, corpus, and selection protocol fixed. For each LR family, distinct time tapers select their peak independently on the same selection split. The time profile identical to $\beta = 1$ LR coupling is the same method and is marked as such. The full held-out matrix is disclosed, while the compact summary uses the non-matching profile selected by minimum clean selection loss before test AE-2 is evaluated. Tail-swap hybrids alter the final-quarter trajectory continuously, and the equal-area stress family changes tracking mismatch at fixed peak, endpoint, and mean exposure. These controls identify a local schedule effect within the tested families. The lagged realized-update-norm control is nearly tied with nominal-LR coupling under the two SGD settings by point estimate, but it is 2.8 AE-2 lower under AdamW/cosine. Thus the tested AdamW result favors the nominal scheduler signal over this lagged effective-step proxy, without establishing that nominal LR is universally the right control variable.

**Weaker settings and intervals crossing zero.** The ten seeds cosine confirmation has Arena-H intervals crossing zero for Qwen3-14B-Base + Alpaca-cleaned (+0.70 [-0.11, 1.50]), Llama-3.1-8B + UltraChat-50k (+0.14 [-0.66, 0.94]), and Llama-3.1-8B + Tulu-3-100k (+0.36 [-0.45, 1.16]). The linear-LR comparison crosses zero for Mistral-7B-v0.3 + UltraChat-50k (+0.23 [-0.13, 0.58]). The prompt-bootstrap check also crosses zero for Qwen3-14B-Base on AE-2, IFEval, and Arena-H, and for the UltraChat-50k/Tulu-3-100k Arena-H rows (Appendix Table 42). These settings are reported as uncertainty and are not counted as uniformly positive evidence.

**Empirical coverage.** The main experiments use English instruction-tuning data and primarily LoRA adaptation at 7B–14B scale, with supplementary 32B LoRA and 1.7B full parameter checks. Judge-based metrics inherit evaluator-version sensitivity. Carl@50 measures exact continuation behavior and the LoRA-space sharpness diagnostic depends on factor parameterization. Training-seed and prompt-level intervals quantify different sources of variation. The seed intervals condition on paired decoding seeds. Appendix Table 8 reports a five-decoding-seed generation-only sensitivity on one representative checkpoint pair, while broader decoding variability remains separate from training-seed uncertainty. The exploratory one-sided envelope diagnostic also lacks a prospective clipped-schedule test. Broader replication is needed across model families, languages, adaptation methods, and optimizer schedules.

# H Additional Identification Controls

## H.1 Compact schedule ablation

The five seeds schedule ablation is reported in the main text as Table 3.

## H.2 Same-distribution and strength controls

The half-level control uses the same warmup to $\sigma_{\max}^2$ and then holds $\sigma_t^2 = \sigma_{\max}^2/2$ for the remainder of training. With the selected $\sigma_{\max} = 10$, its post-warmup variance is therefore 50, with standard deviation $\sqrt{50}$. This setting differs from a fixed $\sigma = 5$ control.

| Method | AE-2↑ | Carl@50 (%)↓ | Sharp.↓ | Clean grad. |
|---|---|---|---|---|
| Tuned fixed Gaussian noise | 15.6 | 11.3 | 0.68 | 0.364 |
| Half-level variance | 15.6 | 10.7 | 0.67 | 0.372 |
| Fixed Gaussian noise ($\sigma = 10$) | 15.0 | 11.2 | 0.70 | 0.420 |
| Fixed Gaussian, $\sigma = 15$ | 14.0 | 10.0 | 0.73 | 0.463 |
| Fixed Gaussian, $\sigma = 20$ | 12.7 | 9.4 | 0.75 | 0.492 |
| AdaNoiseSFT | 19.5 | 9.1 | 0.58 | 0.342 |

Table 29: **Fixed Gaussian noise strength grid, half-level control, and AdaNoiseSFT.** Same-peak fixed Gaussian noise versus AdaNoiseSFT isolates schedule shape at $\sigma_{\max} = 10$.

| Peak scale | Fixed loss | Linear-taper loss | AdaNoiseSFT loss |
|---|---|---|---|
| 5 | 1.658 | 1.628 | 1.625 |
| 10 | 1.682 | 1.608 | 1.599 |
| 15 | 1.720 | 1.622 | 1.613 |
| 20 | 1.761 | 1.649 | 1.635 |

Table 30: **Clean selection losses for fixed noise, the linear taper, and AdaNoiseSFT.** Each family selects its own peak before benchmark evaluation.

The tuned fixed Gaussian noise comparator, the linear taper, and AdaNoiseSFT each select their peak independently on the same selection split. The main same-peak fixed control still uses $\sigma = 10$ to isolate schedule shape. The primary linear taper control uses its independently selected peak.

| Corpus | Family | $\sigma = 5$ | $\sigma = 10$ | $\sigma = 15$ | $\sigma = 20$ | Selected |
|---|---|---|---|---|---|---|
| UltraChat-50k | linear taper | 1.622 | **1.598** | 1.606 | 1.630 | 10 |
| UltraChat-50k | AdaNoiseSFT | 1.604 | **1.583** | 1.591 | 1.617 | 10 |
| Tulu-3-100k | linear taper | 1.615 | **1.590** | 1.600 | 1.626 | 10 |
| Tulu-3-100k | AdaNoiseSFT | 1.596 | **1.574** | 1.582 | 1.608 | 10 |

Table 31: **Corpus-specific clean selection losses for the two schedule families used in the transfer controls.** Both families select $\sigma_{\max} = 10$ in each corpus, so the fixed $\sigma = 10$ rows remain same-peak controls relative to the AdaNoiseSFT operating point.

### H.3 Early and late intervention controls

| Method | AE-2↑ | Carl@50 (%)↓ | Clean grad. | Diag. loss↓ |
|---|---|---|---|---|
| Vanilla SFT | 12.2 | 13.9 | 0.305 | 1.749 |
| Early-only fixed noise | 14.0 | 12.5 | 0.369 | 1.682 |
| Late-only fixed noise | 12.1 | 13.4 | 0.437 | 1.716 |
| Fixed Gaussian noise ($\sigma = 10$) | 15.0 | 11.2 | 0.420 | 1.707 |
| Fixed $\to$ 0 at $\tau = 3/4$ (shared checkpoint) | 16.2 | 10.5 | 0.320 | 1.662 |
| Fixed $\to$ AdaNoiseSFT at $\tau = 3/4$ | 17.9 | 10.0 | 0.337 | 1.633 |
| AdaNoiseSFT | 19.5 | 9.1 | 0.342 | 1.610 |

Table 32: **Temporal interventions.** Early-only noise ends at decay fraction $1/4$. Late-only noise begins at $3/4$ from the corresponding vanilla trajectory. Removed and late-coupled branches share the checkpoint trained with fixed noise.

The branch comparison holds the pre-branch state and future data order fixed. The early-only and late-only rows answer a different timing question because their noise budgets differ.

### H.4 TAIL-SWAP HYBRID SCHEDULES

| Policy | AE-2 | Δ vs linear [95% CI] | IFEval | Arena-H | Clean grad. |
|---|---|---|---|---|---|
| Linear taper | 18.33 | N/A | 48.87 | 17.75 | 0.354 |
| Linear taper → tail AdaNoiseSFT | 19.56 | +1.23 [0.97, 1.50] | 49.78 | 18.61 | 0.350 |
| AdaNoiseSFT → linear taper tail | 18.69 | +0.36 [0.07, 0.66] | 49.30 | 18.34 | 0.365 |
| AdaNoiseSFT | 19.83 | +1.50 [1.27, 1.73] | 50.39 | 19.29 | 0.342 |

Table 33: **Ten-seed continuous tail-swap study.** At the first post-switch update at or after $\tau_0 = 3/4$, the post-switch branch is multiplicatively rescaled to equal the pre-switch variance at that same discrete update, so neither hybrid contains a variance jump. Quality metrics use ten matched seeds. Clean grad. uses the five seeds diagnostic set. N/A marks the reference row of the paired difference.

The hybrids exchange final-quarter shapes continuously. Because the early noise level and the amount of decay in the final quarter change together, this comparison identifies the schedule ordering but does not isolate a unique causal contribution for either quantity.

### H.5 CONTINUOUS SCHEDULE DIAGNOSTICS

For normalized decay progress $\tau \in [0, 1]$, let $h(\tau) = (1 + \cos(\pi\tau))/2$ and $s(\tau) = s_{\max} h(\tau)^\beta$. The decay phase mean variance $E^{\text{dec}}_{\text{var}}$ and LR-weighted squared-variance quantity $E^{\text{dec}}_{\text{bias}}$ are defined in Eq. (9). At $\beta = 1$, these are $1/2$ and $5/8$. Here the bias budget is the last term of Eq. (8) normalized by its fixed-noise value $3C_{\text{b}}^2 s_{\max}^2$. A same-peak linear variance taper has the same mean variance but bias budget $1/3+2/\pi^2 \approx 0.536$, smaller than $0.625$, so under cosine LR this diagnostic favors linear taper despite its worse AE-2. Under linear LR, the corresponding values are $1/2$ for LR coupling and $3/8 + 2/\pi^2 \approx 0.578$ for cosine taper, so the ordering reverses. These quantities motivate the cross-LR identification test but are not downstream-quality rankings.

The one-sided envelope summaries $D_+$ and $D_-$ are defined in Appendix H.11.

### H.6 DISCRETE SCHEDULE ACCOUNTING

Let $a_t = \eta_t/\eta_{\max}$ and $v_t = s_t/s_{\max}$. We report $\widehat{E}_{\text{var}} = T^{-1}\sum_t v_t$, $\widehat{E}_{\text{step}} = T^{-1}\sum_t a_t v_t$, and $\widehat{E}_{\text{bias}} = \sum_t a_t v_t^2 / \sum_t a_t$ for each specified schedule.

| LR | Noise policy | $\widehat{E}_{\text{var}}$ | $\widehat{E}_{\text{step}}$ | $\widehat{E}_{\text{bias}}$ | Peak $v_t$ |
|---|---|---|---|---|---|
| Cosine | Fixed | 0.9746 | 0.4915 | 0.9744 | 1.0000 |
| Cosine | Half-level variance | 0.5003 | 0.2544 | 0.2632 | 1.0000 |
| Cosine | Linear taper | 0.5000 | 0.3504 | 0.5341 | 1.0000 |
| Cosine | Linear taper → coupled tail | 0.4906 | 0.3500 | 0.5339 | 1.0000 |
| Cosine | Coupled → linear tail | 0.5055 | 0.3731 | 0.6187 | 1.0000 |
| Cosine | AdaNoiseSFT | 0.5000 | 0.3729 | 0.6186 | 1.0000 |
| WSD | Fixed | 0.9746 | 0.7286 | 0.9826 | 1.0000 |
| WSD | Linear taper | 0.5000 | 0.4562 | 0.4359 | 1.0000 |
| WSD | Budget-matched linear | 0.7372 | 0.6757 | 0.9589 | 1.4995 |
| WSD | AdaNoiseSFT | 0.7372 | 0.6693 | 0.8620 | 1.0000 |
| Linear | Cosine taper | 0.5000 | 0.3504 | 0.5737 | 1.0000 |
| Linear | AdaNoiseSFT | 0.5000 | 0.3333 | 0.5000 | 1.0000 |

Table 34: **Discrete accounting for $T = 1950$ and warmup 100.** WSD uses 925 stable and 925 cosine-decay steps.

The WSD budget-matching amplitude multiplier is 1.5003. Because the first post-warmup point occurs after zero decay progress, the discrete peak shown in Table 34 is 1.4995. It matches total variance but changes the peak. Under cosine LR, the linear taper and AdaNoiseSFT match total variance while retaining different step-weighted and squared-bias budgets.

## H.7 FOUR-SHAPE SCHEDULE IDENTIFICATION

| Optimizer LR | Distinct taper | $\sigma = 5$ | $\sigma = 10$ | $\sigma = 15$ | $\sigma = 20$ | Selected |
|---|---|---|---|---|---|---|
| Cosine | linear taper | 1.628 | **1.608** | 1.622 | 1.649 | 10 |
| Cosine | step taper | 1.634 | 1.618 | **1.612** | 1.629 | 15 |
| Cosine | WSD taper | 1.629 | **1.611** | 1.617 | 1.638 | 10 |
| Linear | cosine taper | 1.652 | **1.631** | 1.640 | 1.665 | 10 |
| Linear | step taper | 1.649 | 1.637 | **1.635** | 1.654 | 15 |
| Linear | WSD taper | 1.657 | **1.634** | 1.641 | 1.663 | 10 |
| Step | linear taper | 1.675 | **1.655** | 1.662 | 1.684 | 10 |
| Step | cosine taper | 1.669 | **1.651** | 1.658 | 1.680 | 10 |
| Step | WSD taper | 1.665 | **1.646** | 1.654 | 1.676 | 10 |
| WSD | linear taper | 1.700 | **1.682** | 1.690 | 1.713 | 10 |
| WSD | cosine taper | 1.693 | **1.672** | 1.679 | 1.704 | 10 |
| WSD | step taper | 1.689 | 1.676 | **1.674** | 1.696 | 15 |

Table 35: **Clean selection losses for every distinct non-matching time taper in the four-shape study.** Each taper selects its peak independently on the same disjoint selection split. Comparator identity is fixed by the lowest selected loss within each optimizer-LR family before held-out benchmark evaluation.

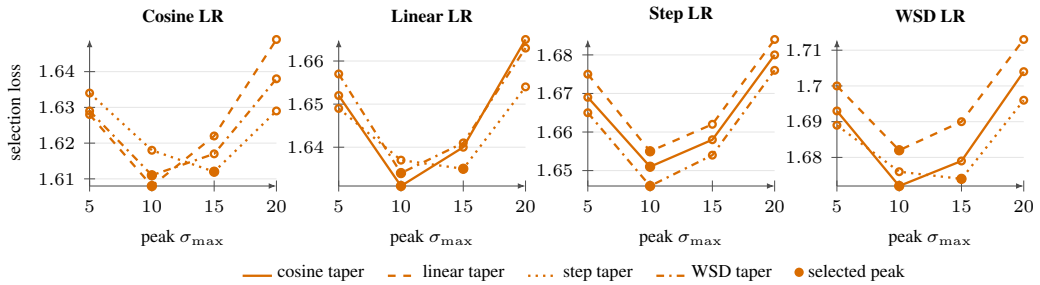



Figure 5: **Peak selection for the mismatched tapers.** Clean selection loss against the peak noise scale for each time taper under each LR schedule. Filled points mark the selected peaks of Table 2.

| Setting | AE-2 Δ [95% CI] | IFEval Δ [95% CI] | Arena-H Δ [95% CI] |
|---|---|---|---|
| *Cosine LR: AdaNoiseSFT minus linear taper* | | | |
| Llama-3.1-8B / Alpaca | +1.50 [1.27, 1.73] | +1.52 [1.12, 1.91] | +1.54 [1.11, 1.98] |
| Mistral-7B-v0.3 / Alpaca | +1.11 [0.62, 1.60] | +1.39 [1.02, 1.76] | +1.54 [0.66, 2.41] |
| Qwen3-8B-Base / Alpaca | +0.87 [0.63, 1.11] | +0.94 [0.37, 1.52] | +1.23 [0.69, 1.76] |
| Qwen3-14B-Base / Alpaca | +0.67 [0.48, 0.86] | +0.92 [0.47, 1.38] | +0.70 [-0.11, 1.50] |
| Llama-3.1-8B / UltraChat-50k | +1.73 [1.43, 2.04] | +1.22 [0.68, 1.76] | +0.14 [-0.66, 0.94] |
| Llama-3.1-8B / Tulu-3-100k | +1.09 [0.80, 1.38] | +1.33 [0.65, 2.02] | +0.36 [-0.45, 1.16] |
| *Linear LR: AdaNoiseSFT minus cosine taper* | | | |
| Llama-3.1-8B / Alpaca | +1.39 [1.10, 1.68] | +0.94 [0.40, 1.48] | +1.22 [0.75, 1.70] |
| Mistral-7B-v0.3 / UltraChat-50k | +0.84 [0.65, 1.03] | +0.28 [0.07, 0.49] | +0.23 [-0.13, 0.58] |
| Qwen3-8B-Base / Alpaca | +0.97 [0.69, 1.24] | +0.70 [0.30, 1.11] | +0.88 [0.33, 1.43] |
| Llama-3.1-8B / Tulu-3-100k | +1.08 [0.76, 1.40] | +0.72 [0.22, 1.23] | +0.58 [0.12, 1.04] |
| Mistral-7B-v0.3 / Alpaca | +0.79 [0.53, 1.05] | +0.61 [0.19, 1.03] | +0.52 [0.06, 0.98] |

Table 36: **Ten-seed paired schedule comparisons ($n = 10$ per setting).** The upper block changes the variance schedule under cosine LR. The lower block changes the LR shape itself and reverses which time profile is coupled to the LR. Intervals are 95% intervals.

| Optimizer LR | AdaNoiseSFT | linear taper | cosine taper | step taper | WSD taper |
|---|---|---|---|---|---|
| Cosine | 19.83 (10) | 18.33 (10) | ≡ coupled | 17.73 (15) | 17.94 (10) |
| Linear | 17.70 (10) | ≡ coupled | 16.31 (10) | 15.98 (15) | 16.14 (10) |
| Step | 17.31 (10) | 15.89 (10) | 16.05 (10) | ≡ coupled | 16.17 (10) |
| WSD | 16.64 (10) | 15.34 (10) | 15.73 (10) | 15.42 (15) | ≡ coupled |

Table 37: **Full schedule-shape matrix on Llama-3.1-8B + Alpaca-cleaned.** Entries are mean AE-2 with the independently selected peak $\sigma_{\max}$ in parentheses. An equivalence mark denotes the profile that is AdaNoiseSFT under that optimizer schedule. Every distinct profile selects its peak on the selection split, and Table 2 selects comparator identity on that same split before evaluating this held-out matrix. Seed-level records for the matrix are included in the supplementary data directory. Central coupled and selected-comparator cells reuse the records reported in the schedule-confirmation analysis.

| Seed | linear taper AE-2 | AdaNoiseSFT AE-2 | Paired Δ |
|---|---|---|---|
| 11 | 18.280 | 19.600 | 1.320 |
| 22 | 17.630 | 19.300 | 1.670 |
| 33 | 17.740 | 18.810 | 1.070 |
| 44 | 18.430 | 20.150 | 1.720 |
| 55 | 18.300 | 19.630 | 1.330 |
| 66 | 18.895 | 19.868 | 0.973 |
| 88 | 18.133 | 19.988 | 1.855 |
| 99 | 17.819 | 19.591 | 1.772 |
| 111 | 19.419 | 21.273 | 1.854 |
| 122 | 18.655 | 20.090 | 1.435 |

Table 38: **Seed records for the cosine-LR schedule contrast.** Seeds 11–55 are the same primary runs reported in Tables 6 and 9. Their five-run linear taper mean is 18.076, displayed there as 18.08. Seeds 66, 88, 99, 111, and 122 complete the ten seeds confirmation. Across all ten seeds, the means are 18.330 and 19.830. The paired mean is 1.500, paired sample SD is 0.322, and the paired Student-$t$ 95% CI is [1.270, 1.730].

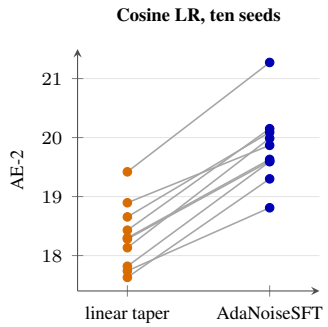



Figure 6: **Paired seeds under cosine LR.** Each line connects the linear taper and AdaNoiseSFT trained with the same seed. All ten seeds improve, by 0.97 to 1.86 AE-2.

Table 2 reports the non-matching time taper with the lowest clean selection loss after each profile selects its peak on the same split. Comparator identity is fixed before test AE-2 evaluation.

Table 36 collects cosine and linear pairwise replications across model families and corpora. The cosine replication fixes linear taper in advance, while the linear replication fixes cosine taper in advance. The central four-shape experiment asks the broader identification question by changing the optimizer schedule on a fixed model and corpus and comparing all distinct time profiles.

### H.8 STEP/WSD REPLICATION BEYOND THE CENTRAL SETTING

| Setting | LR | AdaNoiseSFT | Selected non-matching time taper | Taper AE-2 | Δ [95% CI] |
|---|---|---|---|---|---|
| Mistral-7B-v0.3 / Alpaca-cleaned | step | 15.95 | WSD taper | 15.03 | +0.92 [0.61, 1.23] |
| Mistral-7B-v0.3 / Alpaca-cleaned | WSD | 15.62 | cosine taper | 14.88 | +0.74 [0.45, 1.03] |
| Llama-3.1-8B / UltraChat-50k | step | 18.20 | WSD taper | 17.14 | +1.06 [0.74, 1.38] |
| Llama-3.1-8B / UltraChat-50k | WSD | 17.91 | cosine taper | 17.10 | +0.81 [0.51, 1.11] |
| Qwen3-14B-Base / Alpaca-cleaned | step | 22.08 | WSD taper | 21.36 | +0.72 [0.47, 0.97] |
| Qwen3-14B-Base / Alpaca-cleaned | WSD | 21.55 | cosine taper | 20.99 | +0.56 [0.31, 0.81] |

Table 39: **Cross-backbone and cross-data replication of step/WSD schedule identification.** Each row uses ten matched training seeds. All distinct time tapers select their peak independently on the selection split. The reported comparator is chosen by selection loss before test AE-2 evaluation. These replications extend the fixed Llama-3.1-8B + Alpaca-cleaned four-shape matrix without reusing test AE-2 for comparator selection.

The central $4 \times 4$ matrix keeps model and corpus fixed to identify schedule shape cleanly. Table 39 asks a different question: whether the step/WSD ordering survives one backbone change and one corpus change. Comparator identity is fixed by selection loss before test AE-2 is evaluated. Full time-taper records, selected peaks, and seed-level values are provided in the supplementary material.

**Replication coverage.**

| Setting | Cosine | Linear | Step | WSD |
|---|---|---|---|---|
| Mistral-7B-v0.3 / Alpaca-cleaned | +1.11 [0.62, 1.60] T. 36 | +0.79 [0.53, 1.05] T. 48 | +0.92 [0.61, 1.23] T. 39 | +0.74 [0.45, 1.03] T. 39 |
| Qwen3-8B-Base / Alpaca-cleaned | +0.87 [0.63, 1.11] T. 36 | +0.97 [0.69, 1.24] T. 48 | – | – |
| Qwen3-14B-Base / Alpaca-cleaned | +0.67 [0.48, 0.86] T. 36 | – | +0.72 [0.47, 0.97] T. 39 | +0.56 [0.31, 0.81] T. 39 |
| Llama-3.1-8B / UltraChat-50k | +1.73 [1.43, 2.04] T. 36 | – | +1.06 [0.74, 1.38] T. 39 | +0.81 [0.51, 1.11] T. 39 |
| Llama-3.1-8B / Tulu-3-100k | +1.09 [0.80, 1.38] T. 36 | +1.08 [0.76, 1.40] T. 48 | – | – |
| Mistral-7B-v0.3 / UltraChat-50k | – | +0.84 [0.65, 1.03] T. 48 | – | – |

Table 40: **Coverage of non-central ten seeds AE-2 replications.** Each populated cell reports AdaNoiseSFT minus comparator with a paired 95% CI, followed by the source table. Cosine uses the pre-specified linear taper comparator and linear uses the pre-specified cosine taper comparator. For step and WSD, the displayed non-matching comparator is selected by clean selection loss before test AE-2 evaluation. Dashes denote settings without a ten seeds replication. Of the 24 non-central cells, 13 are populated, and Mistral-7B-v0.3 + Alpaca-cleaned covers all four LR families.

Table 40 assembles the non-central ten seeds coverage in one place. Every populated cell reproduces an AE-2 paired contrast already reported in one of the cited source tables.

### H.9 LEARNING-RATE SELECTION CHECK

| Method | $10^{-4}$ | $2 \times 10^{-4}$ | $3 \times 10^{-4}$ | Selected |
|---|---|---|---|---|
| Vanilla SFT | 1.704 | 1.692 | 1.701 | $2 \times 10^{-4}$ |
| NEFTune | 1.681 | 1.660 | 1.671 | $2 \times 10^{-4}$ |
| Fixed Gaussian noise ($\sigma = 10$) | 1.691 | 1.682 | 1.690 | $2 \times 10^{-4}$ |
| Linear taper | 1.631 | 1.608 | 1.626 | $2 \times 10^{-4}$ |
| AdaNoiseSFT | 1.626 | 1.599 | 1.620 | $2 \times 10^{-4}$ |

Table 41: **Clean selection loss across a common learning-rate grid on Llama-3.1-8B + Alpaca-cleaned.** The same $2 \times 10^{-4}$ LR is selected independently for the displayed baseline families.

| Setting | AE-2 95% CI | IFEval 95% CI | Arena-H 95% CI |
|---|---|---|---|
| Llama-3.1-8B / Alpaca | [0.92, 2.06] | [0.93, 2.11] | [0.88, 2.24] |
| Mistral-7B-v0.3 / Alpaca | [0.74, 1.48] | [0.67, 2.07] | [0.74, 2.32] |
| Qwen3-8B-Base / Alpaca | [0.02, 1.65] | [0.07, 1.82] | [0.35, 2.17] |
| Qwen3-14B-Base / Alpaca | [-0.70, 1.95] | [-0.15, 1.92] | [-0.42, 1.80] |
| Llama-3.1-8B / UltraChat-50k | [1.09, 2.37] | [0.54, 1.88] | [-0.64, 0.93] |
| Llama-3.1-8B / Tulu-3-100k | [0.63, 1.55] | [0.63, 2.04] | [-0.40, 1.09] |

Table 42: **Prompt-level sensitivity.** Prompt-level sensitivity intervals for the AdaNoiseSFT minus linear taper difference, obtained by resampling matched prompt-level evaluator records and recomputing the paired metric within each evaluation set. See Table 36.

### H.10 Uncertainty beyond AE-2

| Metric | Fixed mean | Coupled mean | Paired Δ | Paired 95% CI |
|---|---|---|---|---|
| AE-2↑ | 15.0 | 19.5 | +4.48 | [4.04, 4.91] |
| MT-B↑ | 6.33 | 6.74 | +0.41 | [0.34, 0.48] |
| Arena-H↑ | 16.4 | 19.2 | +2.81 | [2.33, 3.30] |
| IFEval↑ | 46.3 | 50.2 | +3.92 | [3.32, 4.51] |
| MMLU↑ | 61.3 | 61.3 | +0.05 | [-0.12, 0.23] |
| TruthQA↑ | 37.4 | 37.3 | -0.11 | [-0.44, 0.22] |
| BBH↑ | 50.9 | 50.8 | -0.05 | [-0.34, 0.23] |
| GSM8K↑ | 34.5 | 34.4 | -0.07 | [-0.60, 0.45] |
| Inv-IFE↑ | 65.8 | 67.0 | +1.15 | [0.34, 1.96] |
| Carl@50 (%)↓ | 11.2 | 9.1 | -2.09 | [-2.73, -1.45] |
| Sharp.↓ | 0.70 | 0.58 | -0.13 | [-0.16, -0.10] |
| Clean grad. | 0.420 | 0.342 | -0.08 | [-0.10, -0.06] |
| Diag. loss↓ | 1.707 | 1.610 | -0.10 | [-0.12, -0.08] |
| Gap↓ | 0.341 | 0.255 | -0.09 | [-0.09, -0.08] |

Table 43: **Llama-3.1-8B: paired AdaNoiseSFT-minus-fixed comparisons with 95% intervals.**

| Metric | Fixed mean | Coupled mean | Paired Δ | Paired 95% CI |
|---|---|---|---|---|
| AE-2↑ | 13.2 | 17.0 | +3.78 | [3.43, 4.12] |
| MT-B↑ | 6.05 | 6.47 | +0.42 | [0.28, 0.56] |
| Arena-H↑ | 14.2 | 17.0 | +2.79 | [1.97, 3.62] |
| IFEval↑ | 43.3 | 46.4 | +3.11 | [2.29, 3.92] |
| MMLU↑ | 60.2 | 60.2 | -0.05 | [-0.26, 0.15] |
| TruthQA↑ | 39.3 | 39.4 | +0.14 | [-0.37, 0.65] |
| BBH↑ | 49.1 | 49.4 | +0.33 | [-0.02, 0.68] |
| GSM8K↑ | 29.6 | 30.0 | +0.37 | [-0.16, 0.91] |
| Inv-IFE↑ | 62.2 | 62.9 | +0.71 | [0.40, 1.01] |
| Carl@50 (%)↓ | 10.8 | 8.6 | -2.22 | [-2.55, -1.89] |
| Sharp.↓ | 0.75 | 0.64 | -0.11 | [-0.15, -0.07] |
| Clean grad. | 0.426 | 0.361 | -0.06 | [-0.10, -0.03] |
| Diag. loss↓ | 1.736 | 1.672 | -0.06 | [-0.07, -0.06] |
| Gap↓ | 0.366 | 0.288 | -0.08 | [-0.09, -0.07] |

Table 44: **Mistral-7B-v0.3: paired AdaNoiseSFT-minus-fixed comparisons with 95% intervals.**

| Metric | Fixed mean | Coupled mean | Paired Δ | Paired 95% CI |
|---|---|---|---|---|
| AE-2↑ | 18.8 | 22.0 | +3.18 | [2.96, 3.40] |
| MT-B↑ | 6.65 | 7.07 | +0.43 | [0.37, 0.49] |
| Arena-H↑ | 21.9 | 24.1 | +2.23 | [1.50, 2.97] |
| IFEval↑ | 52.5 | 55.4 | +2.88 | [2.13, 3.64] |
| MMLU↑ | 72.2 | 72.4 | +0.15 | [-0.06, 0.37] |
| TruthQA↑ | 47.4 | 47.4 | -0.01 | [-0.62, 0.61] |
| BBH↑ | 55.2 | 55.4 | +0.19 | [-0.29, 0.67] |
| GSM8K↑ | 47.5 | 47.9 | +0.33 | [-0.03, 0.69] |
| Inv-IFE↑ | 70.3 | 70.7 | +0.34 | [-0.04, 0.72] |
| Carl@50 (%)↓ | 9.4 | 8.2 | -1.24 | [-1.46, -1.02] |
| Sharp.↓ | 0.66 | 0.57 | -0.09 | [-0.12, -0.06] |
| Clean grad. | 0.362 | 0.307 | -0.05 | [-0.08, -0.03] |
| Diag. loss↓ | 1.582 | 1.507 | -0.08 | [-0.09, -0.06] |
| Gap↓ | 0.291 | 0.234 | -0.06 | [-0.07, -0.05] |

Table 45: **Qwen3-8B-Base: paired AdaNoiseSFT-minus-fixed comparisons with 95% intervals.**

Prompt-level intervals complement seed CIs by varying evaluator-sample composition.

### H.11 LR-RELATIVE DIAGNOSTICS AND SECOND-SETUP TEST

*Remark* (Optional step-scale design heuristic). For any chosen constant $B_{\text{step}} > 0$, Lemma 1 implies that $s_t \le (B_{\text{step}}/C_{\text{b}})\eta_t$ is sufficient for $\|\eta_t d_t\| \le B_{\text{step}}\eta_t^2$. Saturating this pointwise envelope at a calibrated peak recovers $s_t \propto \eta_t$. We use this only as a design heuristic. It is not required by Corollary 1 and does not predict which vanishing schedule maximizes downstream quality.

Let $a_t = \eta_t/\eta_{\max}$ and $v_t = s_t/s_{\max}$. On the post-warmup schedule grid, define the weighted mismatch

$$D_{\text{LR}} = \frac{\sum_t a_t(v_t - a_t)^2}{\sum_t a_t^3}.$$

This dimensionless schedule diagnostic vanishes for proportional coupling by construction. Let $T_{\text{dec}} = 1850$ denote the post-warmup grid points, including the zero-LR endpoint. We report $E_{\text{step}}^{\text{dec}} = T_{\text{dec}}^{-1}\sum_t a_t v_t$ on that full grid and define $\text{tail} := \{t : \tau_t \ge 3/4\}$ over post-warmup points. The tail-restricted squared-bias ratio is $E_{\text{bias,tail}} = \sum_{t\in\text{tail}} a_t v_t^2 / \sum_{t\in\text{tail}} a_t$. Ratios $v_t/a_t$ are evaluated only where $a_t > 0$, leaving 1849 positive-LR points. The table's "Last" column therefore uses the final positive-LR point. At equal peak, this ratio characterizes over- or under-exposure relative to LR.

| LR | Variance | $D_{\text{LR}}$ | $E_{\text{step}}^{\text{dec}}$ | $E_{\text{bias,tail}}$ | Last $v_t/a_t$ |
|---|---|---|---|---|---|
| Cosine | Linear taper | 0.0091 | 0.3511 | 0.037278 | 749.7769 |
| Cosine | AdaNoiseSFT | 0.0000 | 0.3747 | 0.009322 | 1.0000 |
| Linear | Cosine taper | 0.0114 | 0.3511 | 0.007338 | 0.0013 |
| Linear | AdaNoiseSFT | 0.0000 | 0.3331 | 0.031250 | 1.0000 |

Table 46: **Post-warmup diagnostics computed from the specified schedules.** Here $\text{tail} = \{t : \tau_t \ge 3/4\}$ denotes the final quarter of post-warmup points. Sum-based quantities use all $T_{\text{dec}} = 1850$ post-warmup grid points including the zero-LR endpoint. The displayed $v_t/a_t$ ratio uses the final one of the 1849 positive-LR points. Appendix B.4 and Eq. (9) give the corresponding continuous-integral exposure and bias quantities. The cosine taper control under linear LR has greater exposure and smaller tail bias than LR coupling.

**Exploratory one-sided envelope diagnostic.** The remark opening Appendix H.11 defines an LR-relative upper envelope. To summarize violations of that envelope without treating under- and over-exposure as symmetric, define

$$D_+ = \frac{\sum_t a_t\,[v_t - a_t]_+^2}{\sum_t a_t^3}, \qquad D_- = \frac{\sum_t a_t\,[a_t - v_t]_+^2}{\sum_t a_t^3}.$$

These deterministic schedule summaries were formulated post hoc after inspecting the control results. They are exploratory and require prospective validation.

| LR | Variance schedule | $D_+$ | $D_-$ |
|---|---|---|---|
| Cosine | Linear taper | 0.0013 | 0.0078 |
| Cosine | $\beta = 0.5$, exposure-matched | 0.0042 | 0.0293 |
| Cosine | $\beta = 1$ | 0.0000 | 0.0000 |
| Cosine | $\beta = 1.5$ | 0.0000 | 0.0124 |
| Cosine | $\beta = 2$ | 0.0000 | 0.0375 |
| Linear | Cosine taper | 0.0088 | 0.0026 |
| Linear | $\beta = 1$ | 0.0000 | 0.0000 |
| Linear | $\delta = -1.2$ equal-area | 0.0027 | 0.0008 |
| Linear | $\delta = +1.2$ equal-area | 0.0008 | 0.0027 |

Table 47: **One-sided LR-relative envelope diagnostics.** $D_+$ records only variance above the proportional envelope, while $D_-$ records only variance below it. The summaries diagnose schedule geometry and are not fitted predictors of AE-2. The $\delta$ rows use the equal-area family $v_\delta(u) = 1 - u + \delta u(1-u)(u-1/2)$, where $u$ is normalized post-warmup progress.

The existing controls motivate a two-sided tradeoff: schedules above the envelope can incur an overshoot penalty, while schedules far below it can lose useful perturbation exposure. A prospective clipped-schedule experiment is required to test that interpretation directly.

| Setting | Metric | Cosine taper | AdaNoiseSFT | Paired $\Delta$ [95% CI] |
|---|---|---|---|---|
| Llama-3.1-8B / Alpaca | AE-2 | 16.31 | 17.70 | +1.39 [1.10, 1.68] |
| Llama-3.1-8B / Alpaca | IFEval | 47.23 | 48.17 | +0.94 [0.40, 1.48] |
| Llama-3.1-8B / Alpaca | Arena-H | 17.70 | 18.92 | +1.22 [0.75, 1.70] |
| Mistral-7B-v0.3 / UltraChat-50k | AE-2 | 15.24 | 16.08 | +0.84 [0.65, 1.03] |
| Mistral-7B-v0.3 / UltraChat-50k | IFEval | 47.04 | 47.32 | +0.28 [0.07, 0.49] |
| Mistral-7B-v0.3 / UltraChat-50k | Arena-H | 16.97 | 17.20 | +0.23 [-0.13, 0.58] |
| Qwen3-8B-Base / Alpaca | AE-2 | 20.56 | 21.53 | +0.97 [0.69, 1.24] |
| Qwen3-8B-Base / Alpaca | IFEval | 53.79 | 54.49 | +0.70 [0.30, 1.11] |
| Qwen3-8B-Base / Alpaca | Arena-H | 22.40 | 23.28 | +0.88 [0.33, 1.43] |
| Llama-3.1-8B / Tulu-3-100k | AE-2 | 17.91 | 18.99 | +1.08 [0.76, 1.40] |
| Llama-3.1-8B / Tulu-3-100k | IFEval | 49.21 | 49.93 | +0.72 [0.22, 1.23] |
| Llama-3.1-8B / Tulu-3-100k | Arena-H | 18.03 | 18.61 | +0.58 [0.12, 1.04] |
| Mistral-7B-v0.3 / Alpaca | AE-2 | 15.08 | 15.87 | +0.79 [0.53, 1.05] |
| Mistral-7B-v0.3 / Alpaca | IFEval | 44.82 | 45.43 | +0.61 [0.19, 1.03] |
| Mistral-7B-v0.3 / Alpaca | Arena-H | 15.21 | 15.73 | +0.52 [0.06, 0.98] |

Table 48: **Five ten seeds linear-LR comparisons.** Equal peak and total post-warmup variance isolate the scheduler relation, while the cosine taper control keeps its cosine time profile. Intervals are 95% intervals.

| $\delta$ | $D_{\mathrm{LR}}$ | AE-2 | $\Delta$ vs proportional [95% CI] |
|---|---|---|---|
| -1.2 | 0.0034 | 17.50 | -0.20 [-0.34, -0.06] |
| -0.6 | 0.0009 | 17.61 | -0.09 [-0.29, 0.11] |
| +0.0 | 0.0000 | 17.70 | N/A |
| +0.6 | 0.0009 | 17.69 | -0.01 [-0.11, 0.09] |
| +1.2 | 0.0034 | 17.31 | -0.39 [-0.56, -0.23] |

Table 49: **Equal-area tracking stress test under linear LR, held out from model selection.** All schedules share peak one, zero endpoint, and mean post-warmup variance $1/2$. Positive and negative perturbations move away from proportional tracking without changing total exposure. N/A marks the reference row of the paired difference.

**Held-out tracking check.** For normalized post-warmup progress $u \in [0, 1]$, the family is $v_\delta(u) = 1-u+\delta u(1-u)(u-1/2)$. It preserves peak one, endpoint zero, and mean exposure $1/2$. Small departures produce uncertain quality changes, whereas the larger mismatches reduce AE-2 on average on both sides of proportional tracking. At $|\delta| = 1.2$, the two distortions have the same $D_{\mathrm{LR}}$ but unequal AE-2 point estimates. Their difference is not tested directly. We therefore use this comparison only to show that $D_{\mathrm{LR}}$ alone does not determine the observed effect magnitude.

| Backbone | Family | 5 | 10 | 15 | 20 | Selected peak |
|---|---|---|---|---|---|---|
| Llama | fixed | 1.658 | 1.682 | 1.720 | 1.761 | 5 |
| Llama | linear taper | 1.628 | 1.608 | 1.622 | 1.649 | 10 |
| Llama | AdaNoiseSFT | 1.625 | 1.599 | 1.613 | 1.635 | 10 |
| Mistral | fixed | 1.715 | 1.738 | 1.777 | 1.817 | 5 |
| Mistral | linear taper | 1.677 | 1.652 | 1.668 | 1.689 | 10 |
| Mistral | AdaNoiseSFT | 1.656 | 1.639 | 1.648 | 1.665 | 10 |
| Qwen3-8B-Base | fixed | 1.558 | 1.576 | 1.619 | 1.659 | 5 |
| Qwen3-8B-Base | linear taper | 1.530 | 1.505 | 1.515 | 1.543 | 10 |
| Qwen3-8B-Base | AdaNoiseSFT | 1.490 | 1.470 | 1.482 | 1.504 | 10 |
| Qwen3-14B-Base | fixed | 1.435 | 1.455 | 1.494 | 1.537 | 5 |
| Qwen3-14B-Base | linear taper | 1.434 | 1.402 | 1.422 | 1.447 | 10 |
| Qwen3-14B-Base | AdaNoiseSFT | 1.395 | 1.380 | 1.385 | 1.406 | 10 |

Table 50: **Clean selection losses across backbones on the shared grid** $\{5, 10, 15, 20\}$**.** The fixed family additionally searches $\sigma \in \{1, 2, 3\}$ in Table 51. The selected peaks follow the documented selection procedure.

| Backbone | $\sigma = 1$ | $\sigma = 2$ | $\sigma = 3$ | $\sigma = 5$ | Selected |
|---|---|---|---|---|---|
| Llama | 1.688 | 1.675 | 1.665 | 1.658 | 5 |
| Mistral | 1.749 | 1.741 | 1.729 | 1.715 | 5 |
| Qwen3-8B-Base | 1.587 | 1.577 | 1.567 | 1.558 | 5 |
| Qwen3-14B-Base | 1.463 | 1.453 | 1.451 | 1.435 | 5 |

Table 51: **Fixed Gaussian noise selection loss on** $\sigma \in \{1, 2, 3, 5\}$**.** The selected value is $\sigma = 5$.

# I Additional Mechanism Diagnostics

## I.1 Exact quadratic least-squares example

The following quadratic least-squares case isolates the clean-gradient residual created by fixed Gaussian smoothing.

Proposition 4, stated and proved in Appendix B.8, gives the exact fixed-noise residual and the vanishing-noise limit used in Figure 7.

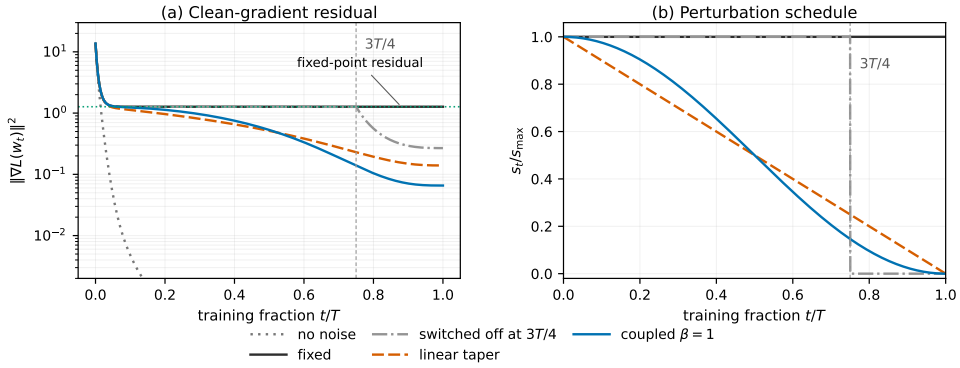



Figure 7: **Illustration of Proposition 4 with** $d_w = 20$**,** $m = 200$**,** $T = 4000$**,** $s_{\max} = 0.5$**, and** $\eta_{\max} = 0.05/\lambda_{\max}(\hat{\Sigma})$**.** Fixed perturbation converges to the residual $s^2\|w_s\|^2$ marked by the horizontal guide. Vanishing schedules remove the floor. The vertical guide marks $3T/4$.

## I.2 Checkpoint bias-to-step audit

To test whether the measured perturbation bias scales with the applied noise variance along the trajectory, as the leading term of Lemma 1 predicts, we estimate it directly at six saved checkpoints using the pre-registered 1,024-example training-audit subset from Appendix D. Clean gradients are

accumulated over the entire subset before the norm is taken. For each 64-example audit block, perturbed gradients use 16 antithetic pairs evaluated with the training microbatch size, and the paired mean is aggregated across blocks to define $\hat{d}_t$. With $v_t = \sigma_t^2/\sigma_{\max}^2$, Table 52 reports the variation of $u_t^{\text{audit}} = \|\hat{d}_t\|/v_t$ after normalization by its first-checkpoint value. The diagnostic tests approximate variance scaling along each observed trajectory. It is not an independent test that LR matching is preferable: for a non-matching schedule, the variance-to-LR geometry $v_t/(\eta_t/\eta_{\max})$ is fixed by construction.

| LR | Selected non-match | Coupled CV($u^{\text{audit}}$)↓ | Non-match CV($u^{\text{audit}}$)↓ | Coupled final/first | Non-match final/first |
|---|---|---|---|---|---|
| Cosine | Linear taper | 0.039 | 0.052 | 0.99 | 1.04 |
| Linear | Cosine taper | 0.030 | 0.047 | 1.01 | 0.96 |
| Step | WSD taper | 0.041 | 0.061 | 0.98 | 1.06 |
| WSD | Cosine taper | 0.033 | 0.054 | 1.00 | 0.95 |

Table 52: **Checkpoint variance-scaling audit.** At six fixed training-progress checkpoints, $\hat{d}_t$ estimates $d_t$ from paired perturbed and clean gradients on the same 1,024-example audit subset. We define $u_t^{\text{audit}} = \|\hat{d}_t\|/v_t$, where $v_t = \sigma_t^2/\sigma_{\max}^2$, normalize $u_t^{\text{audit}}$ by its first-checkpoint value, and report its coefficient of variation and final-to-first ratio. This tests the leading order prediction that perturbation bias is approximately proportional to applied variance along the observed trajectory. It is not an independent test that LR matching is preferable.

### I.3 Matched-trace case study

The case study uses seed 77, which was reserved before training and is excluded from the five- and ten seeds result sets. The fixed, removed, and late-coupled branches start from the seed-77 checkpoint trained with fixed noise at the first update with $\tau \geq 3/4$ and share future minibatches and dropout streams. Figure 8 shows the matched final-quarter schedules and clean-gradient traces computed on the same pre-registered 1,024-example training-audit subset. The endpoint norms are formed after accumulating gradients over all audit examples. Delete-one-audit-block jackknife standard errors are 0.011, 0.009, and 0.010 for fixed, removed, and late-coupled branches, respectively. Downstream quality for the corresponding five seeds branches (seeds 11–55) is reported in Table 32.

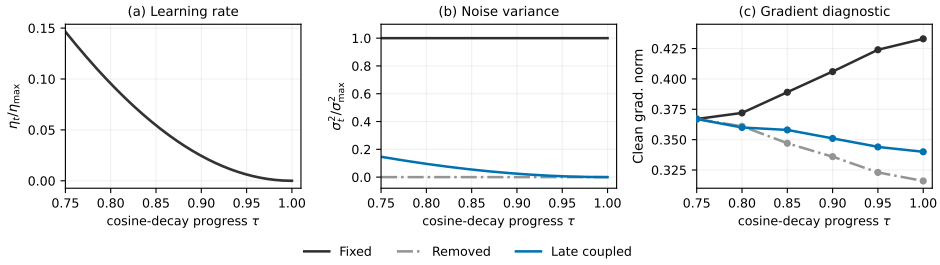



Figure 8: **Final-quarter shared-checkpoint case study.** Panel (b) shows normalized variance $\sigma_t^2/\sigma_{\max}^2$. Markers in panel (c) show the checkpoint values, and lines connect them. Endpoint quality for the corresponding five seeds branches (seeds 11–55) is reported in Table 32.

| Final-quarter policy | AE-2↑ | Clean grad. | Diag. loss↓ |
|---|---|---|---|
| Continue fixed noise | 15.0 | 0.420 | 1.707 |
| Remove noise | 16.2 | 0.320 | 1.662 |
| **Switch to noise coupled to the LR** | 17.9 | 0.337 | 1.633 |

Table 53: **Matched final-quarter branches.** All branches start from the same checkpoint trained with fixed noise at $\tau = 3/4$. Diag. loss is the clean response-token loss on the held-out diagnostic split. Clean gradient norm serves only as a diagnostic associated with residual perturbation bias.

# J ADDITIONAL SCHEDULE EXPERIMENTS

This appendix reports four additional experiments. All of them use the ten matched training seeds 11, 22, 33, 44, 55, 66, 88, 99, 111, and 122, paired Student-$t$ 95% CIs computed from unrounded paired differences, and the selection protocol of Section 5.1 unless stated otherwise.

## J.1 SECOND COMPLETE $4 \times 4$ MATRIX

Table 54 and Figure 9 complete the schedule matrix for Mistral-7B-v0.3 with Alpaca-cleaned. The cosine and linear rows keep the linear taper and cosine taper comparators that the replication protocol fixes in advance. The step and WSD rows keep the comparator selected by clean selection loss. The two remaining mismatched tapers of each row are added with peaks selected on the selection split (Table 55). The matching shape has the highest AE-2 in every row. The comparator fixed before evaluation is the strongest mismatched taper in the linear and WSD rows but not in the cosine and step rows, so its identity does not follow from test AE-2. On IFEval, the step contrast is $+0.51$ [0.07, 0.95] and the WSD contrast is $+0.33$ [$-0.08$, 0.74].

Table 54: **Complete schedule matrix on Mistral-7B-v0.3 with Alpaca-cleaned.** Held-out AE-2 over ten seeds. A dash marks the taper that coincides with LR coupling in that row.

| Optimizer LR | AdaNoiseSFT | linear taper | cosine taper | step taper | WSD taper |
|---|---|---|---|---|---|
| Cosine | **16.90** | 15.79 | – | 15.29 | 15.88 |
| Linear | **15.87** | – | 15.08 | 14.72 | 15.02 |
| Step | **15.95** | 14.69 | 15.12 | – | 15.03 |
| WSD | **15.62** | 14.64 | 14.88 | 14.82 | – |

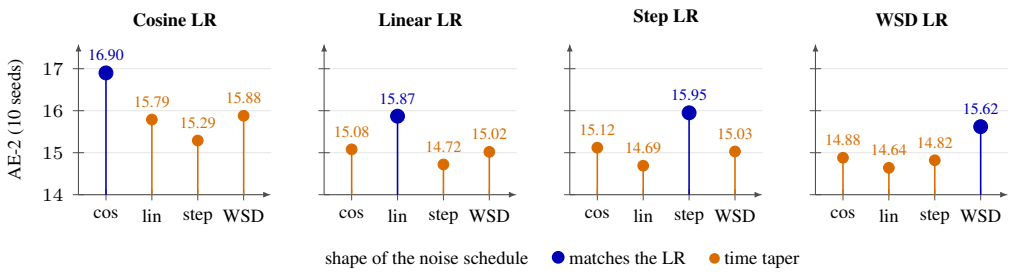



Figure 9: **The second matrix shows the same pattern.** Held-out AE-2 of the four noise shapes under each LR schedule for Mistral-7B-v0.3 with Alpaca-cleaned, over ten seeds.

Table 55: **Added cells of the Mistral matrix.** Paired AE-2 difference between AdaNoiseSFT and each added taper, and the clean selection loss of that taper at each peak.

| LR | Taper | $\Delta$ AE-2 [95% CI] | $\sigma = 5$ | $\sigma = 10$ | $\sigma = 15$ | $\sigma = 20$ |
|---|---|---|---|---|---|---|
| Cosine | step taper | $+1.61$ [1.18, 2.04] | 1.684 | 1.663 | **1.650** | 1.672 |
| Cosine | WSD taper | $+1.02$ [0.64, 1.40] | 1.677 | **1.651** | 1.660 | 1.690 |
| Linear | step taper | $+1.15$ [0.79, 1.51] | 1.689 | 1.682 | **1.668** | 1.695 |
| Linear | WSD taper | $+0.85$ [0.49, 1.21] | 1.702 | **1.676** | 1.681 | 1.694 |
| Step | linear taper | $+1.26$ [0.87, 1.65] | 1.721 | **1.688** | 1.697 | 1.724 |
| Step | cosine taper | $+0.83$ [0.47, 1.19] | 1.711 | 1.687 | **1.684** | 1.713 |
| WSD | linear taper | $+0.98$ [0.65, 1.31] | 1.736 | **1.709** | 1.714 | 1.744 |
| WSD | step taper | $+0.80$ [0.43, 1.17] | 1.725 | 1.713 | **1.697** | 1.718 |

## J.2 QWEN3-14B-BASE UNDER STEP AND WSD LR

The Qwen3-14B-Base setting with Alpaca-cleaned previously covered only cosine LR, where the AdaNoiseSFT contrast is $+0.67$ AE-2. Under step and WSD LR, the peaks of AdaNoiseSFT and of the three mismatched tapers are selected on the selection split (Table 56), and the two arms are then trained with ten seeds. AdaNoiseSFT reaches 22.08 AE-2 against 21.36 for the selected WSD taper under step LR, a paired difference of $+0.72$ [0.47, 0.97]. Under WSD LR it reaches 21.55 against

20.99 for the selected cosine taper, a difference of $+0.56$ [0.31, 0.81]. The IFEval differences are $+0.54$ [0.10, 0.98] and $+0.39$ [$-0.02$, 0.80]. These settings differ from the Llama and Mistral settings in model family as well as size, so they support no statement about model size.

Table 56: **Peak selection on Qwen3-14B-Base.** Clean selection loss at each peak. Bold marks the selected peak.

| LR | Schedule | $\sigma = 5$ | $\sigma = 10$ | $\sigma = 15$ | $\sigma = 20$ |
|---|---|---|---|---|---|
| Step | AdaNoiseSFT | 1.434 | **1.406** | 1.420 | 1.446 |
| Step | linear taper | 1.449 | **1.431** | 1.435 | 1.462 |
| Step | cosine taper | 1.448 | 1.425 | **1.416** | 1.439 |
| Step | WSD taper | 1.445 | **1.414** | 1.429 | 1.454 |
| WSD | AdaNoiseSFT | 1.443 | **1.413** | 1.428 | 1.451 |
| WSD | linear taper | 1.464 | **1.439** | 1.450 | 1.474 |
| WSD | cosine taper | 1.454 | **1.425** | 1.436 | 1.458 |
| WSD | step taper | 1.459 | 1.443 | **1.430** | 1.449 |

### J.3 Controls that move a taper toward the LR

The design rule of Section 4 predicts that a taper loses quality where its variance exceeds the LR envelope. Under step LR, the clipped control uses $v_t = \min(v_t^{\text{WSD taper}}, a_t)$, where $a_t$ is the LR normalized by its peak. Both parent arms already use $\sigma_{\max} = 10$, so the clipped control keeps this peak without retuning. Under WSD LR, clipping the cosine taper would change nothing, because cosine taper never exceeds the WSD LR. The WSD control therefore moves the taper halfway toward the LR, $v_t = \frac{1}{2} v_t^{\text{cosine taper}} + \frac{1}{2} a_t$, again with $\sigma_{\max} = 10$. Table 57 reports both controls.

Table 57: **Controls that move a taper toward the LR** on Llama-3.1-8B with Alpaca-cleaned, over ten seeds.

| LR | Schedule | AE-2 | Paired difference [95% CI] |
|---|---|---|---|
| Step | WSD taper | 16.17 | |
| Step | WSD taper clipped to the LR | 17.05 | $+0.88$ [0.61, 1.15] over WSD taper |
| Step | AdaNoiseSFT | 17.31 | $+0.26$ [0.07, 0.45] over the clipped taper |
| WSD | cosine taper | 15.73 | |
| WSD | halfway to the LR | 16.27 | $+0.54$ [0.29, 0.79] over cosine taper |
| WSD | AdaNoiseSFT | 16.64 | $+0.37$ [0.15, 0.59] over the halfway taper |

### J.4 LoRA dropout coupled to the LR

This experiment applies the same rule to a second regularizer. It uses Llama-3.1-8B with Alpaca-cleaned, turns embedding noise off in every arm, and changes only the LoRA dropout probability $p_t$. The arms are no dropout ($p = 0$), fixed dropout ($p_t = p_{\max}$), dropout coupled to the LR ($p_t = p_{\max}\eta_t/\eta_{\max}$), and three time tapers. Each schedule selects $p_{\max} \in \{0.05, 0.10, 0.20\}$ by clean selection loss (Table 59). The cosine arm without dropout reuses the Vanilla SFT run of Table 1, which uses no LoRA dropout. The comparator taper is linear taper under cosine LR, as in the replication protocol, and the taper with the lowest selection loss under step and WSD LR. Table 58 reports test AE-2. Against the comparator taper, the IFEval differences are $+0.42$ [0.03, 0.81], $+0.32$ [$-0.06$, 0.70], and $+0.27$ [$-0.11$, 0.65] under cosine, step, and WSD LR.

Table 58: **LoRA dropout coupled to the LR** on Llama-3.1-8B with Alpaca-cleaned, over ten seeds. Upper block: test AE-2, with a dash where a taper coincides with LR coupling. Lower block: paired differences of dropout coupled to the LR.

| LR | $p = 0$ | Fixed | LR coupling | linear taper | cosine taper | step taper | WSD taper |
|---|---|---|---|---|---|---|---|
| Cosine | 12.20 | 13.18 | **14.10** | 13.72 | – | 13.28 | 13.78 |
| Step | 11.54 | 12.15 | **13.06** | 12.39 | 12.52 | – | 12.57 |
| WSD | 11.71 | 11.98 | **12.79** | 12.22 | 12.38 | 12.43 | – |

| | vs. comparator taper | vs. fixed dropout | vs. $p = 0$ |
|---|---|---|---|
| Cosine | $+0.38$ [0.13, 0.63] | $+0.92$ [0.57, 1.27] | $+1.90$ [1.43, 2.37] |
| Step | $+0.49$ [0.19, 0.79] | $+0.91$ [0.53, 1.29] | $+1.52$ [1.05, 1.99] |
| WSD | $+0.41$ [0.12, 0.70] | $+0.81$ [0.45, 1.17] | $+1.08$ [0.62, 1.54] |

Table 59: **Selection of the dropout peak.** Clean selection loss at each $p_{\max}$. Bold marks the selected value.

| LR | Schedule | 0.05 | 0.10 | 0.20 | LR | Schedule | 0.05 | 0.10 | 0.20 |
|---|---|---|---|---|---|---|---|---|---|
| Cosine | Fixed | 1.685 | **1.669** | 1.680 | Step | cosine taper | 1.705 | 1.686 | **1.674** |
| Cosine | LR coupling | 1.672 | **1.647** | 1.662 | Step | WSD taper | 1.694 | **1.667** | 1.681 |
| Cosine | linear taper | 1.679 | **1.654** | 1.666 | WSD | Fixed | 1.708 | 1.696 | **1.691** |
| Cosine | step taper | 1.682 | 1.665 | **1.655** | WSD | LR coupling | 1.697 | **1.671** | 1.684 |
| Cosine | WSD taper | 1.675 | **1.658** | 1.671 | WSD | linear taper | 1.707 | 1.688 | **1.676** |
| Step | Fixed | 1.700 | **1.689** | 1.694 | WSD | cosine taper | 1.704 | 1.681 | **1.673** |
| Step | LR coupling | 1.691 | **1.663** | 1.679 | WSD | step taper | 1.701 | **1.675** | 1.686 |
| Step | linear taper | 1.698 | 1.679 | **1.672** | | | | | |