Adaptive and Neutral Theory of Evolution Strategies for Large Language Models

Soma Yokoi[*]
Department of Computer Science
The University of Tokyo
yokoi@aoni.waseda.jp

Issei Sato
Department of Computer Science
The University of Tokyo
sato@g.ecc.u-tokyo.ac.jp

Abstract

Evolution strategies (ES) have recently emerged as a promising alternative to gradient-based reinforcement learning for fine-tuning large language models, but classical zeroth-order optimization theory does not systematically account for the empirical observations driving this success. We formulate a finite-time weak approximation of Z-score-normalized ES as a stochastic differential equation (SDE) and analyze Adaptive and Neutral Regimes. This framework explains (1) the information-geometric optimality of the Z-score heuristic for a population-level shaping ratio under a Gaussian reward model, (2) the role of basin-local low-rank geometry in ES with populations far smaller than the parameter dimension, and the mechanisms linking parameter displacement and task curvature to (3) the rise-then-decay of the training reward and (4) catastrophic forgetting of pre-training capabilities. The regime structure also motivates early-stopping and compute-budget guidance by comparing further improvement with accumulated displacement. Synthetic experiments compare closed-form coefficient models of the SDE with ES, and evaluations on seven language models across four tasks show approximately linear full-parameter squared displacement and task-dependent training-reward trajectories.

1 Introduction

Aligning large language models (LLMs) with human values is a paramount challenge. Gradient-based reinforcement learning (RL) methods such as proximal policy optimization [Schulman et al., 2017] are the de facto choice for fine-tuning, but suffer from reward hacking, training instability, and hyperparameter sensitivity [Casper et al., 2023, Qiu et al., 2026]. Evolution strategies (ES) [Salimans et al., 2017] have emerged as a compelling derivative-free alternative; Qiu et al. [2026] report higher Countdown accuracy at matched evaluation budgets and greater training stability in conciseness fine-tuning than the RL baselines.

This empirical success presents three theoretical puzzles unexplained by classical zeroth-order theory [Nesterov and Spokoiny, 2017, Vemula et al., 2019]. (a) The *dimensionality paradox*: how can ES make useful progress with a population size $N \approx 30$ in $D \approx 10^9$ dimensions? (b) *Training-reward rise-then-decay*: Liang et al. [2026] document a characteristic improvement-then-decline pattern in the ES training reward. (c) *Catastrophic forgetting*: Abdi et al. [2026] report that HellaSwag accuracy keeps declining after the training-task accuracy converges, with parameter drift orders of magnitude larger than with group relative policy optimization (GRPO).

We formulate a finite-time weak approximation of the ES dynamics as a stochastic differential equation (SDE) and decompose the parameter space, via a basin-local low-rank Hessian assumption,

[*]Present address: Center for Data Science, Waseda University, Tokyo, Japan.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

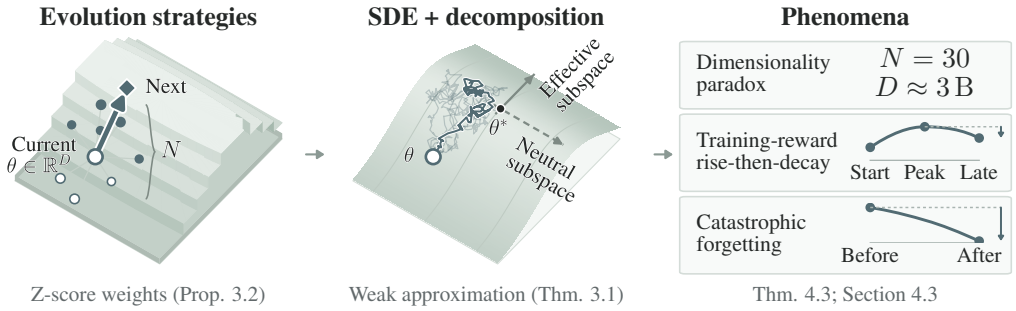


Figure 1: **Overview.** Left: ES makes a discrete update from N perturbations of $\theta \in \mathbb{R}^D$. Center: numerical SDE trajectories are shown on the smoothed reward landscape, with one selected realization highlighted and fixed effective and neutral directions anchored at the local maximizer θ^*. Right: the three phenomena analyzed with this decomposition, namely the dimensionality paradox of progress with $N = 30$ perturbations in $D \approx 3 \times 10^9$ dimensions, training-reward rise-then-decay, and catastrophic forgetting.

into an effective subspace where local progress occurs and a neutral subspace where parameter displacement accumulates. This geometric decomposition relates the three puzzles to local improvement, displacement, and task curvature (Figure 1).

Our contributions are threefold.

(C1) SDE approximation and closed-form coefficient models. We derive a finite-time first-order weak SDE approximation of Z-score-normalized ES with exact finite-batch coefficients (Theorem 3.1). Euler–Maruyama integration of closed-form coefficient models, which approximate the exact coefficients, reproduces ES trajectories on smooth and oscillatory landscapes with pointwise relative differences no larger than approximately 6.6% and 4.2%, respectively, in mean squared neutral radius in the tested settings (Appendix I.3).

(C2) Information-geometric optimality of Z-score shaping. Under a Gaussian reward model, Z-score shaping attains the Fisher-information upper bound on the population-level shaping ratio (Proposition 3.2), providing a principled justification for the shaping choice empirically adopted in Qiu et al. [2026] and Abdi et al. [2026].

(C3) Regime-specialized analysis yields progress and displacement estimates and motivates stopping and budget-allocation guidance. We analyze local adaptive progress and finite-time neutral displacement (Definition 4.1). Puzzle (a) is examined through the effective-subspace noise baseline, proportional to d/N for the effective dimension d, and coefficient residuals (Theorem 4.3); puzzles (b) and (c) are analyzed by evaluating the effect of parameter displacement on the training and evaluation rewards (Section 4.3 and Appendix H). The stopping and budget-allocation guidance is based on comparing further improvement with accumulated displacement (Remark 4.6).

2 Related Work

ES for LLM fine-tuning. ES [Rechenberg, 1973, Schwefel, 1977, Salimans et al., 2017] are zeroth-order methods. Classical analyses yield dimension-dependent complexity bounds under their respective oracle and regularity assumptions [Jamieson et al., 2012, Duchi et al., 2015, Nesterov and Spokoiny, 2017]; yet MeZO [Malladi et al., 2023] demonstrates forward-only fine-tuning of billion-parameter LLMs. Qiu et al. [2026] report less reward hacking than GRPO in their conciseness fine-tuning experiments.

Continuous-time dynamics and intrinsic dimensionality. Continuous-time SDE models have been used to analyze first-order optimization dynamics [Li et al., 2017, Chaudhari and Soatto, 2018, Zhu et al., 2019]; high-dimensional limits of population Gaussian ES updates have recently been characterized under shrinking perturbations [Sarkar et al., 2026], and Fornasier et al. [2026] derive a

form of ES from consensus-based optimization through time rescaling and asymptotic limits; we study Z-score-normalized ES with fixed finite batches. The low-rank geometry assumed in our analysis (Assumption 3.3) is motivated by evidence for low-dimensional structure from intrinsic-dimension probes of neural networks and language-model fine-tuning [Li et al., 2018, Aghajanyan et al., 2021], the "bulk + outliers" Hessian spectrum observed in classification networks [Gur-Ari et al., 2018, Sagun et al., 2018, Ghorbani et al., 2019], the success of low-rank adaptation [Hu et al., 2022] and gradient low-rank projection [Zhao et al., 2024], and curvature measurements of ES-smoothed rewards on Qwen2.5 by Liang et al. [2026].

Recent work on anisotropic ES geometry. Liang et al. [2026] develop dedicated mathematical models for the unshaped ES variant: multi-time discrete Ornstein–Uhlenbeck (OU)-like dynamics for training-reward rise-then-decay and a one-step static best-of-N accessibility model for the dimensionality paradox. The remaining phenomenon, catastrophic forgetting, is empirically documented by Abdi et al. [2026]. Building on a similar low-dimensional curvature hypothesis, we analyze local progress and displacement using a finite-time weak SDE approximation of the practically deployed Z-score-normalized ES. This framework relates training-reward rise-then-decay and catastrophic forgetting to task-dependent displacement contributions and identifies the information-geometric optimality of Z-score shaping (Proposition 3.2); Section 6 compares the results (Table 2).

3 Formulation and Modeling

We separate the mean ES update from sampling fluctuations to study reward improvement and displacement in effective and neutral subspaces (Figure 1); Table 3 at the beginning of the appendix lists the conditions of each result, their basis, and their empirical status.

3.1 Problem Setting

We formulate LLM fine-tuning as a bounded zeroth-order optimization problem: $\max_{\theta\in\mathbb{R}^D} R(\theta)$ where only the oracle $\theta \mapsto R(\theta)$ is available and $|R(\theta)| \le M$ for some $M > 0$. Following standard practice [Qiu et al., 2026, Sarkar et al., 2026], $R(\theta)$ is defined via deterministic greedy decoding, so exploration occurs in parameter space. Since LLM rewards may be sparse or discontinuous, we adopt the Gaussian-smoothing surrogate [Nesterov and Spokoiny, 2017],

$$R_\sigma(\theta) := \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I_D)}[R(\theta + \sigma\epsilon)]. \tag{1}$$

For every fixed $\sigma > 0$, R_σ is C^∞ for bounded Borel-measurable R. As $\sigma \downarrow 0$, it converges to R at every Lebesgue point of R, hence almost everywhere. Gradients and Hessians throughout refer to R_σ; we abbreviate $\mathcal{H}(\theta) := \nabla^2 R_\sigma(\theta)$.

3.2 Standard ES and Continuous-Time SDE Model

We analyze standard ES [Qiu et al., 2026] with isotropic covariance $\sigma^2 I$ and Z-score fitness shaping:

$$\theta_{t+1} = \theta_t + \alpha\,\hat{g}(\theta_t), \quad \hat{g}(\theta_t) = \frac{1}{\sigma N}\sum_{n=1}^{N} F(R_n)\,\epsilon_n, \quad F(R_n) = \frac{R_n - \mu_R}{s_R}, \tag{2}$$

where $N \ge 2$, $R_n := R(\theta_t + \sigma\epsilon_n)$, $\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_D)$ across individuals and iterations, and μ_R and s_R are the batch mean and standard deviation, with $s_R^2 = N^{-1}\sum_n (R_n - \mu_R)^2$. We set $F(R_n) = 0$ when $s_R = 0$. For $s_R > 0$, the normalization yields $\frac{1}{N}\sum_n F(R_n) = 0$ and $\frac{1}{N}\sum_n F(R_n)^2 = 1$, making exploration invariant to reward scale. We define the population reward standard deviation as

$$\sigma_R(\theta) := \mathrm{Std}_{\epsilon\sim\mathcal{N}(0,I_D)}[R(\theta + \sigma\epsilon)]. \tag{3}$$

We assume $\sigma_R(\theta) > 0$ throughout the domain of interest; Proposition A.4 provides the bound $\sigma_R \ge \sigma\|\nabla R_\sigma\|$. We define the exact finite-batch coefficients by $b_N(\theta) := \mathbb{E}[\hat{g}(\theta)]$ and $\Sigma_N(\theta) := \mathrm{Cov}[\hat{g}(\theta)]$, with expectations over a fresh batch.

Theorem 3.1 (Continuous-time model of standard ES)**.** *Under the reward and batch assumptions above, the standard ES update rule admits the first-order weak SDE approximation:*

$$d\Theta_t^\alpha = b_N(\Theta_t^\alpha)\,dt + \sqrt{\alpha}\,\Sigma_N(\Theta_t^\alpha)^{1/2}\,dW_t, \tag{4}$$

where W_t is a D-dimensional standard Wiener process, $\Theta_0^\alpha = \theta_0$, and θ_k denotes the k-th iterate of Equation (2) *with step size α and initial state θ_0. Fix $R, D, N, \sigma, T, \theta_0$ independently of α, with R Borel measurable, $\sigma, T > 0$, and θ_0 deterministic. For every $\varphi \in C^2(\mathbb{R}^D)$ whose derivatives of orders zero through two have polynomial growth,*

$$\max_{0 \le k \le \lfloor T/\alpha \rfloor} |\mathbb{E}\varphi(\theta_k) - \mathbb{E}\varphi(\Theta_{k\alpha}^\alpha)| = O(\alpha) \qquad \text{as } \alpha \downarrow 0. \tag{5}$$

On the time scale $t = k\alpha$, the SDE approximates the mean and second moments of ES with error $O(\alpha)$. This absolute error bound does not identify the diffusion covariance; matching moments to second order does (Appendix A.3). The proof appears in Appendix A. Under the coefficient approximations $b_N \approx \nabla R_\sigma/\sigma_R$ and $\Sigma_N \approx I_D/(\sigma^2 N)$, the factor $1/\sigma_R$ in the drift acts as an *adaptive learning rate*; in iteration time $s = t/\alpha$, the diffusion is isotropic with scale $\eta := \alpha/(\sigma\sqrt{N})$.

The finite quadratic decomposition of σ_R^2 (Proposition A.4) gives the two-term approximation

$$\sigma_R(\theta)^2 \approx \sigma^2\|\nabla R_\sigma(\theta)\|^2 + \frac{\sigma^4}{2}\|\mathcal{H}(\theta)\|_F^2, \tag{6}$$

whose nonnegative residual $\mathbb{E}r_2^2$ is the variance of the reward beyond its linear and quadratic components (Proposition A.4). The gradient-dominant first-order approximation $\sigma_R \approx \sigma\|\nabla R_\sigma\|$ (Definition 4.1) is accurate when $\nabla R_\sigma \neq 0$ and $(\sigma^4/2)\|\mathcal{H}\|_F^2 + \mathbb{E}r_2^2 \ll \sigma^2\|\nabla R_\sigma\|^2$. In the diagonal-quadratic experiments, the Hessian-norm contribution is small relative to the gradient contribution $\sigma^2\|\nabla R_\sigma\|^2$ (Figure 4).

3.3 Information-Geometric Optimality of Z-Score Shaping

Proposition 3.2 (Optimality of Z-score shaping)**.** *For any shaping Ψ satisfying the conditions below, the ratio A_Ψ^2/B_Ψ^2 satisfies the Fisher-information upper bound*

$$A_\Psi^2(\theta)/B_\Psi^2(\theta) \le \mathcal{I}(f_R(\cdot;\theta)), \tag{7}$$

where $Y = R(\theta + \sigma\epsilon)$ has density $f_R(\cdot;\theta)$, $A_\Psi := \mathbb{E}[\Psi'(Y)]$, $B_\Psi^2 := \mathbb{E}[\Psi(Y)^2]$, and $\mathcal{I}(f_R) := \mathbb{E}[s(Y)^2]$ with $s = -\partial_r \log f_R$. Assume that Y lies almost surely in a union of disjoint open intervals, that f_R is positive and locally absolutely continuous on each interval, and that $0 < \mathcal{I}(f_R) < \infty$. Require Ψ to be C^1 on each interval, $|A_\Psi| < \infty$, $0 < B_\Psi^2 < \infty$, and $\Psi f_R \to 0$ at every interval endpoint, including infinite endpoints. Equality holds iff $\Psi = cs$ almost surely for a single constant $c \neq 0$ common to all intervals. For the Gaussian law $Y \sim \mathcal{N}(R_\sigma, \sigma_R^2)$, Z-score shaping $\Psi_Z(r) = (r - R_\sigma)/\sigma_R$ attains equality.

The ratio compares squared mean sensitivity at a fixed second moment of the shaping weights. A bounded nonconstant reward cannot have an exactly Gaussian law, so the Gaussian equality case is an auxiliary model. The proof and a conditional efficiency formula for non-Gaussian rewards appear in Appendix B.

3.4 Subspace Decomposition

Building on the intrinsic low-rank geometry reviewed in Section 2, we formulate the landscape's effective subspace via a spectral gap on the Hessian:

Assumption 3.3 (Approximately low-rank curvature)**.** Let θ^* be a local maximizer of R_σ, and let $\mathcal{H}(\theta^*)$ have eigendecomposition with eigenvalues ordered by $|\lambda_1| \ge \cdots \ge |\lambda_D|$ and corresponding eigenvectors $\xi_1, \ldots, \xi_D$. Define the d-dimensional effective subspace $\mathcal{S} := \operatorname{span}\{\xi_1, \ldots, \xi_d\}$, the fixed orthogonal projector $P := \sum_{i=1}^d \xi_i\xi_i^\top$ onto $\mathcal{S}$, and the Hessian blocks $H_u(\theta) := P\mathcal{H}(\theta)P$, $H_v(\theta) := (I - P)\mathcal{H}(\theta)(I - P)$. Here $\lambda_{\max}(|\cdot|) = \|\cdot\|_{\mathrm{op}}$ denotes the largest absolute eigenvalue of a symmetric matrix. Let $\mathcal{U}(\theta^*)$ be a neighborhood of θ^*. We assume: (i) **Spectral gap:**

$$\rho_H := \sup_{\theta \in \mathcal{U}(\theta^*)} \frac{\lambda_{\max}(|H_v(\theta)|)}{\lambda_{\max}(|H_u(\theta)|)}, \qquad 0 < \rho_H \ll 1; \tag{8}$$

(ii) **Low effective dimension:** $d \ll D$. On $\mathcal{U}(\theta^*)$, the cross terms $P\mathcal{H}(\theta)(I-P)$ remain $O(\rho_H \lambda_{\max}(|H_u(\theta)|))$ in operator norm (see Appendix C).

Assumption 3.3 is an idealization of the empirical "bulk + outliers" Hessian spectrum of overparameterized networks [Sagun et al., 2018, Ghorbani et al., 2019, Liang et al., 2026].

The ambient space splits as $\mathbb{R}^D = \mathcal{S} \oplus \mathcal{S}^\perp$ with the *neutral subspace* $\mathcal{S}^\perp$ of dimension $D - d \approx D$. We use coordinates centered at the local maximizer:

$$u = P(\theta - \theta^*), \qquad v = (I-P)(\theta - \theta^*), \qquad \theta = \theta^* + u + v. \tag{9}$$

Thus $u^* = v^* = 0$. Here ∇_u and ∇_v denote the corresponding projected gradients. For an observation window starting at t_0, the neutral displacement is $v_t - v_{t_0} = (I-P)(\theta_t - \theta_{t_0})$, whose squared norm equals $\|v_t\|^2$ when $v_{t_0} = 0$.

In iteration time $s = t/\alpha$, let $Y_s = \Theta^\alpha_{\alpha s}$ and $\widetilde{W}_s = \alpha^{-1/2} W_{\alpha s}$. Then Equation (4) becomes $dY_s = \alpha b_N(Y_s)\, ds + \alpha A_N(Y_s)\, d\widetilde{W}_s$, where $A_N = \Sigma_N^{1/2}$ and $\widetilde{W}$ is a standard Wiener process. Since P and θ^* are fixed, the coordinates of Y_s satisfy

$$\begin{aligned} du_s &= \alpha P b_N(Y_s)\, ds + \alpha P A_N(Y_s)\, d\widetilde{W}_s, \\ dv_s &= \alpha (I-P) b_N(Y_s)\, ds + \alpha (I-P) A_N(Y_s)\, d\widetilde{W}_s. \end{aligned} \tag{10}$$

The neutral noise covariance per unit s is $\alpha^2 (I-P)\Sigma_N(Y_s)(I-P)$.

For the coefficient model below, we write iteration time s as t and denote the parameter path by θ_t. Under the coefficient approximations $b_N \approx \nabla R_\sigma/\sigma_R$ and $\Sigma_N \approx I_D/(\sigma^2 N)$, Equation (10) gives the model

$$du_t = \frac{\alpha}{\sigma_R(\theta_t)} \nabla_u R_\sigma(\theta_t)\, dt + \eta\, dW_t^u, \qquad dv_t = \frac{\alpha}{\sigma_R(\theta_t)} \nabla_v R_\sigma(\theta_t)\, dt + \eta\, dW_t^v, \tag{11}$$

with $\eta = \alpha/(\sigma\sqrt{N})$, $dW_t^u = P\, dW_t$, and $dW_t^v = (I-P)\, dW_t$ for a standard Wiener process in iteration time. In this model, the neutral noise energy per unit time is $\eta^2(D-d)$, concentrating stochastic exploration on $\mathcal{S}^\perp$. This geometric separation is formalized in Section 4.

4 Theoretical Analysis

Definition 4.1 (Dynamic regimes of standard ES)**.**
We organize the analysis into three regimes; local progress and finite-time displacement are analyzed with the exact-coefficient SDE, and coefficient models are used where specified:

Adaptive Regime. The effective-subspace gradient dominates:

$$\sigma_R \approx \sigma \|\nabla_u R_\sigma\|, \qquad \|\nabla_u R_\sigma(\theta_t)\| \gg \|\nabla_v R_\sigma(\theta_t)\|. \tag{12}$$

The normalized gradient flow in Theorem 4.3 describes local progress at constant speed α/σ.

Neutral Regime. The effective-subspace gradient does not dominate:

$$\sigma_R \gg \sigma \|\nabla_u R_\sigma\|. \tag{13}$$

That is, the effective-subspace slope explains little of the reward spread across perturbations, so in the coefficient model of Equation (11), Z-score normalization reduces the effective-subspace drift speed far below the Adaptive speed α/σ. Finite-time growth of the neutral mean squared displacement (MSD) is considered in Proposition 4.4.

Saturation Regime. With constant-speed inward radial drift and isotropic noise, the neutral radius has an invariant law in which drift and diffusion balance; the neutral noise energy accumulated over h iterations, $\eta^2(D-d)h$, matches the second moment of this law when the total number of perturbations is $Nh \approx (D-d)/4$, a quarter of the neutral dimension (Appendix F). The toy experiments show MSD plateaus (Section 5), whereas the LLM experiments considered here use far fewer perturbations, $Nh \ll D - d$.

The Adaptive and Neutral Regimes classify states by the effective-subspace contribution to σ_R. Since Proposition 4.4 places no condition on σ_R, neutral displacement also accumulates during the Adaptive Regime. Local effective-coordinate improvement offers an analogy with Darwinian positive selection, while neutral displacement recalls Kimura's neutral theory [Kimura, 1968] (Appendix G).

4.1 Analysis of the Adaptive Regime

We analyze local progress using the fixed rank-d projector P and local maximizer θ^* defined in Assumption 3.3. Let $F(u) := R_\sigma(\theta^* + u)$ be R_σ on the effective slice $v = v^*$, and define $V_u(u) := F(u^*) - F(u)$ and $\bar{V}_u(\theta) := V_u(P(\theta - \theta^*))$; derivatives of F are taken on $\mathcal{S}$. Although LLM landscapes are globally multimodal, we localize to an open neighborhood $\mathcal{B}(u^*) \subset \mathcal{S}$ of u^*, in which F satisfies the standard regularity conditions:

Assumption 4.2 (Local geometry in the effective subspace)**.** On $\mathcal{B}(u^*)$, $V_u \geq 0$, $\|\nabla^2 F(u)\|_{\mathrm{op}} \leq L$, and the Polyak–Łojasiewicz condition $\frac{1}{2}\|\nabla F(u)\|^2 \geq \mu V_u(u)$ holds with $\mu > 0$. On $\mathcal{U}(\theta^*)$, the Hessian is L_3-Lipschitz in Frobenius norm, $\|\mathcal{H}(\theta) - \mathcal{H}(\theta')\|_F \leq L_3\|\theta - \theta'\|$. The constants L, L_3, μ refer to R_σ and may depend on σ and D.

For the local stochastic analysis, write the SDE of Theorem 3.1 in iteration time as $\theta_s := \Theta^\alpha_{\alpha s}$, with drift αb_N and covariance rate $\alpha^2\Sigma_N$. At states with $u \in \mathcal{B}(u^*)$ and $\nabla F(u) \neq 0$, define the projected coefficient residuals

$$e_N(\theta) := Pb_N(\theta) - \frac{\nabla F(u)}{\sigma\|\nabla F(u)\|}, \qquad E_N(\theta) := P\Sigma_N(\theta)P - \frac{P}{\sigma^2 N}. \tag{14}$$

The drift residual also includes the error of evaluating the gradient on the effective slice $v = v^*$ instead of at the full state θ. The following theorem compares local progress with projected noise and coefficient residuals, and analyzes a separate normalized gradient flow (proof in Appendix D).

Theorem 4.3 (Local progress in the Adaptive Regime)**.** *Under Assumption 4.2, the iteration-time generator $\mathcal{L}_N$ of the exact-coefficient SDE satisfies, at every state used in Equation* (14)*,*

$$\mathcal{L}_N\bar{V}_u(\theta) \leq -\frac{\alpha}{\sigma}\|\nabla F(u)\| + \gamma_u + r_N(\theta) \leq -2\beta\sqrt{V_u(u)} + \gamma_u + r_N(\theta), \tag{15}$$

where $\beta := \alpha\sqrt{2\mu}/(2\sigma)$, $\gamma_u := Ld\alpha^2/(2\sigma^2 N)$, and $r_N(\theta) := \alpha\|\nabla F(u)\|\,\|e_N(\theta)\| + \frac{\alpha^2}{2}|\operatorname{Tr}(\nabla^2 F(u)E_N(\theta))|$. For the independent normalized gradient flow

$$\dot{u}^{\mathrm{ref}}_s = \frac{\alpha}{\sigma}\frac{\nabla F(u^{\mathrm{ref}}_s)}{\|\nabla F(u^{\mathrm{ref}}_s)\|}, \qquad u^{\mathrm{ref}}_0 = u_0 := P(\theta_0 - \theta^*), \tag{16}$$

let $V_0 := V_u(u_0) > 0$ and $T_A := \sqrt{V_0}/\beta$. While u^{ref}_s remains in $\mathcal{B}(u^*)$ with nonzero gradient and $0 \leq s < T_A$,

$$V_u(u^{\mathrm{ref}}_s) \leq (\sqrt{V_0} - \beta s)^2. \tag{17}$$

In Theorem 4.3, the improvement coefficient α/σ is independent of d and D, whereas the noise baseline γ_u has an explicit d/N factor rather than D/N. Hence the population needed for the improvement term to exceed this baseline grows with d rather than D, which addresses the dimensionality paradox at the level of this term; L, $\|\nabla F\|$, and r_N may still depend on D.

4.2 Analysis of the Neutral Regime

We focus on the MSD of $v_s \in \mathcal{S}^\perp$ over a deterministic window of h iterations. Let $c_N(\theta)$ be the probability of a nonzero batch standard deviation at θ, and let $\delta_m(c_N(\theta))$ be the bound in Proposition A.5 on the relative error of the projected covariance trace, where $m = D - d$.

Proposition 4.4 (Total neutral MSD)**.** *Let Y_s solve the exact-coefficient SDE in Equation* (10) *on a deterministic interval $[a, b]$, where R is bounded and Borel measurable and $d < D$. Set $h = b - a > 0$, $q_B = (N-1)h/(D-d)$, and $\bar{c} = h^{-1}\mathbb{E}\int_a^b c_N(Y_s)\,ds$, and suppose $\delta_m(c_N(\theta)) \leq \varepsilon_\Sigma$ on reached states with $c_N(\theta) > 0$. If $0 \leq \varepsilon_\Sigma, q_B \leq 1$, then*

$$\left|\frac{\mathbb{E}\|v_b - v_a\|^2}{\eta^2(D-d)h} - 1\right| \leq 4\left(1 - \bar{c} + \varepsilon_\Sigma + \sqrt{q_B}\right). \tag{18}$$

Hence, when batches rarely have all rewards equal ($1 - \bar{c} \ll 1$), the projected covariance trace is close to $(D-d)c_N/(\sigma^2 N)$ ($\varepsilon_\Sigma \ll 1$), and the window is short compared with $(D-d)/N$ iterations ($Nh \ll D - d$), the neutral MSD grows approximately linearly in the number of iterations, $\mathbb{E}\|v_b - v_a\|^2 \approx \alpha^2(D-d)h/(\sigma^2 N) = \eta^2(D-d)h$. The estimate includes accumulated drift, noise, and their cross term; Appendix E derives Equation (18) from a sharper bound.

4.3 Application: Stopping and Budget Allocation

Training-reward rise-then-decay [Liang et al., 2026] and catastrophic forgetting [Abdi et al., 2026] motivate examining the training-task contribution of neutral displacement and the total parameter displacement (Appendix H). We use the effective-slice gap V_u and the neutral quadratic contribution $V_v(v) := -\frac{1}{2}v^\top H_v(\theta^*)v$. Write $\lambda_u := \lambda_{\max}(|H_u(\theta^*)|)$.

Proposition 4.5 (Training-task contribution and parameter displacement)**.** *(i) Under Assumption 3.3, for a random neutral coordinate $v_s \in \mathcal{S}^\perp$ with $\mathbb{E}\|v_s\|^2 < \infty$, the quadratic contribution at θ^* satisfies*

$$\begin{aligned} 0 \le \mathbb{E}V_v(v_s) &= -\tfrac{1}{2}\operatorname{Tr}(H_v(\theta^*)\operatorname{Cov}(v_s)) - \tfrac{1}{2}(\mathbb{E}v_s)^\top H_v(\theta^*)(\mathbb{E}v_s) \\ &\le \tfrac{1}{2}\rho_H\lambda_u\,\mathbb{E}\|v_s\|^2. \end{aligned} \tag{19}$$

(ii) For $T > 0$ iterations, write $\mathbb{E}\|v_T - v_0\|^2 = \eta^2(D-d)T(1+\delta_T) > 0$ and $\mathbb{E}\|u_T - u_0\|^2 = \varepsilon_T\mathbb{E}\|v_T - v_0\|^2$ with finite second moments. Then the root-mean-square (RMS) displacement is

$$\left(\mathbb{E}\|\theta_T - \theta_0\|^2\right)^{1/2} = \sqrt{\alpha^2(D-d)T/(\sigma^2 N)}\,\sqrt{(1+\delta_T)(1+\varepsilon_T)}. \tag{20}$$

By (i), the neutral coordinates add a nonnegative quadratic training loss with curvature at most $\rho_H\lambda_u$, whereas V_u has negative expected drift where the improvement term in Theorem 4.3 exceeds $\gamma_u + r_N$. Since the training gap $R_\sigma(\theta^*) - R_\sigma(\theta)$ equals $V_u + V_v$ up to a third-order Taylor remainder (Equation (45)), the training reward can rise and then decay once the growth of V_v exceeds the decrease of V_u, which addresses training-reward rise-then-decay.

Part (ii) is an orthogonal decomposition valid for any process, including measured ES runs, whereas Proposition 4.4 with $[a, b] = [0, T]$ bounds $|\delta_T|$ only for the exact-coefficient process; ε_T needs a separate estimate. When both are small, the RMS displacement is close to $\sqrt{\alpha^2(D-d)T/(\sigma^2 N)}$, which grows with the neutral dimension $D - d$ rather than d. An evaluation task need not share the ρ_H-suppressed curvature on $\mathcal{S}^\perp$, so this displacement can impair its performance, which addresses catastrophic forgetting at the level of displacement scale and curvature.

Remark 4.6 (Stopping and budget allocation). When effective-coordinate improvement becomes small while neutral displacement continues to accumulate, early stopping can be considered by comparing the benefit of further training with its effect on the training and evaluation tasks. Under a fixed budget $C = N \cdot T$ at fixed α and σ, the nominal neutral MSD $\eta^2(D-d)T = \alpha^2(D-d)C/(\sigma^2 N^2)$ decreases with N, whereas the improvement coefficient α/σ does not; among allocations with $T = C/N$ long enough for effective-coordinate improvement to become small, a larger N lowers this scale.

Appendix H proves Proposition 4.5 and discusses Remark 4.6 with fixed-budget LLM runs.

5 Numerical Validation

We compare ES with the models of Sections 3 and 4 in synthetic and LLM experiments, following the three parts of Section 4. Appendix I.3 compares the closed-form coefficient models with ES on smooth, multimodal, discontinuous, and oscillatory toy landscapes. The synthetic experiments use a diagonal quadratic toy with effective and neutral curvature magnitudes $\lambda_u = 1$ and $\lambda_v = \rho_H$ realizing Assumption 3.3 ($d = 10$, $\alpha = 5 \times 10^{-4}$, $\sigma = 10^{-3}$, $N = 30$, $\eta \approx 9.13 \times 10^{-2}$). The LLM experiments fine-tune seven models from five families with standard ES on Countdown [Gandhi et al., 2024, Pan et al., 2025], GSM8K [Cobbe et al., 2021], MATH [Hendrycks et al., 2021], and SVAMP [Patel et al., 2021] ($N = 30$, $T = 500$, two or three seeds per model–task cell), vary the population size $N \in \{8, 15, 30, 60, 120\}$ and the horizon up to 2000 iterations, and measure Gram-matrix displacement in one Countdown run of each Qwen2.5-{0.5B, 3B, 7B}-Instruct model [Qwen Team, 2025] (setups in Appendix J).

5.1 Adaptive Regime: Progress across Dimensions

In the toy with $\rho_H = 10^{-6}$ and $D \in \{10^3, 10^4, 10^5\}$ (20 trials), the curves $\|u_t - u^*\|^2$ overlap across D (Figure 2a), consistent with Theorem 4.3, whose improvement coefficient and noise baseline do not depend on D in this toy. In the LLM runs, the normalized reward gaps are described by fitted curves of the quadratic reference form in Equation (17) across the approximately $15.4\times$ parameter-count range of Qwen2.5 (Figure 3a; fits in Appendix J.2), and across seven models the median over retained tasks (two cells excluded) of the seed-mean R^2 of this form lies between 0.955 and 0.978 (Table 1).

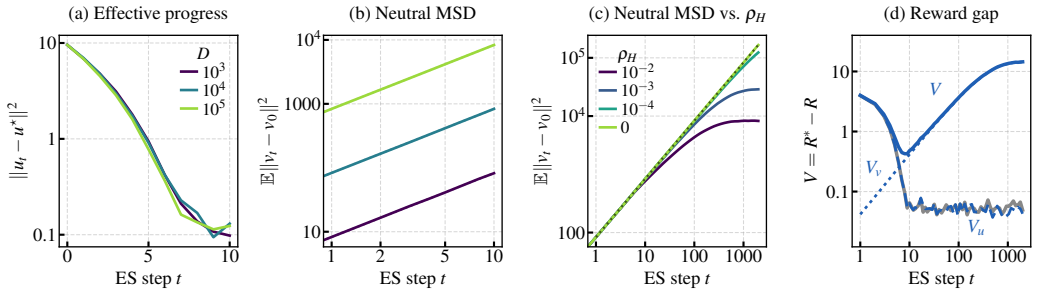


Figure 2: **Synthetic experiments.** (a) The effective-subspace squared distance $\|u_t - u^*\|^2$ is nearly identical across D. (b) Neutral MSD overlaps the nominal $\eta^2(D-d)t$. (c) It saturates for larger ρ_H (dotted: nominal). (d) Growing V_v turns the reward gap upward (blue: $\rho_H = 10^{-3}$; grey: $\rho_H = 0$). In (b,c), the ± 1 s.e. bands are narrower than the lines.

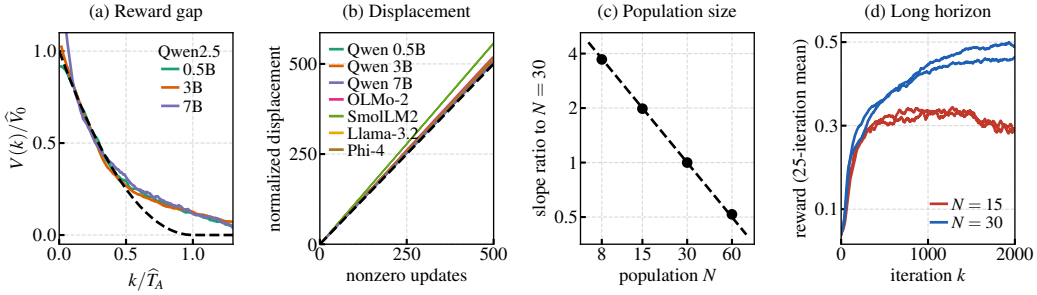


Figure 3: **LLM experiments** (dashed: reference curves). (a) Normalized Qwen2.5 reward gaps overlap and follow the fitted quadratic reference form early on. (b) Squared displacement over $\eta^2 D$ grows nearly linearly, up to 12% above unit rate, across seven models. (c,d) Qwen2.5-0.5B on Countdown: ratios of mean slopes are close to $30/N$; the reward declines after its peak at $N = 15$ but rises at $N = 30$.

5.2 Neutral Regime: Displacement

In the same toy, fitted neutral MSD slopes (Proposition 4.4) match $\eta^2(D-d)$ within 0.2% (ratios $\{1.002, 1.001, 1.000\}$; Figure 2b), and fitting $C\,D^p$ to the full-parameter MSD at $D \in \{10^4, 10^5, 10^6\}$ ($T = 1000$, five trials) gives $p = 0.999$ and $C = 8.43$ (nominal: 1 and 8.33). For the three Qwen2.5 runs, the Gram-matrix MSD of the intended updates, reconstructed from the perturbation seeds, has slopes within 0.2% of the full-parameter nominal rate $\eta^2 D$ (Appendix J.2). Across the seven models and four tasks, the squared displacement of the realized weights grows approximately linearly in the number of nonzero updates, with cell-mean slopes 0.5–12% above $\eta^2 D$ (Figure 3b; task means in Table 1). On Qwen2.5-0.5B with Countdown, the slopes scale approximately as $1/N$ for $N \in \{8, 15, 30, 60\}$ (Figure 3c).

Table 1: **Per-model summary of the LLM experiments.** Displacement slopes per nonzero update exceed $\eta^2 D$ by 1–12% (seed and task means), the quadratic reference form fits retained tasks, and HellaSwag accuracy has a negative slope against $\sqrt{\text{updates}}$ in most runs.

	Qwen2.5			OLMo-2	SmolLM2	Llama-3.2	Phi-4
	0.5B	3B	7B	1B	1.7B	3B	mini
D (10^9)	0.49	3.09	7.62	1.48	1.71	3.21	3.84
Slope / nominal	1.009	1.015	1.008	1.038	1.118	1.010	1.035
Median seed-mean R^2	0.955	0.962	0.956	0.978	0.978	0.959	0.960
Neg. acc. slopes / runs	10/10	8/8	8/8	7/8	5/8	9/10	9/10

Table 2: **Comparison with prior ES studies.** This work treats the three puzzles in one Z-score-normalized model.

	Liang et al. [2026]	Abdi et al. [2026]	This work
Shaping	Unshaped (analysis)	Z-score	Z-score; attains the Fisher bound under an auxiliary Gaussian law (Proposition 3.2)
Dimensionality paradox	Low-dimensional curvature; best-of-N accessibility	—	Noise baseline $\propto d/N$ (Theorem 4.3)
Rise-then-decay	Heterogeneous curvature under fixed stochasticity	—	Neutral contribution V_v (Section 4.3)
Forgetting	—	Observed (HellaSwag)	Displacement scale (Proposition 4.5) and task curvature (Appendix H.5)
Analysis	Discrete AR(1), local OU	Empirical	Finite-time weak SDE

5.3 Application: Rise-then-Decay and Forgetting

At $D = 10^4$ with $\rho_H \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$ (15 trials), the neutral MSD grows approximately linearly and, for $\rho_H \in \{10^{-3}, 10^{-2}\}$, plateaus within the displayed horizon (Figure 2c); the isotropic radial model in Appendix F.2 provides a comparison, but its closed-form prefactor differs from the observed one (Appendix I.2). Through V_v, this displacement shapes the reward gap. The U-shaped gap in Figure 2d (10 trials) matches the $V_u + V_v$ decomposition of Section 4.3: V_u decreases similarly for $\rho_H = 0$ and $\rho_H = 10^{-3}$, while only the latter retains a V_v contribution that bends V upward. In the LLM experiments, held-out HellaSwag [Zellers et al., 2019] accuracy has a negative slope against the square root of completed updates in 56 of 62 runs (Table 1). On Qwen2.5-0.5B with Countdown, the final-50 mean of the training reward lies below the peak 25-iteration mean by about 13% on average at $N = 8$ within 500 iterations and by about 15% at $N = 15$ over 2000 iterations, whereas the $N = 30$ runs remain near their peaks at 2000 iterations (Figure 3d); at a fixed budget $NT = 15000$, larger N reduces displacement and the HellaSwag drop, while the terminal reward is highest at $N = 30$ (Appendix J.3). These task- and horizon-dependent observations are consistent with the trade-off in Remark 4.6.

6 Discussion

Dimensionality paradox. Classical zeroth-order analyses give dimension-dependent bounds [Jamieson et al., 2012, Duchi et al., 2015, Nesterov and Spokoiny, 2017]. In contrast, the rate of zeroth-order stochastic gradient descent in MeZO depends on a local effective rank of the Hessian rather than on the parameter count [Malladi et al., 2023], and Liang et al. [2026] attribute the success of small-population ES to low-dimensional curvature that keeps improving perturbations accessible (Table 2). For Z-score-normalized ES, our local progress estimate (Theorem 4.3) requires improvement to exceed a noise baseline proportional to d/N, with no explicit dependence on D. The synthetic and LLM observations in Section 5.1 are consistent with this requirement.

Training-reward rise-then-decay. Liang et al. [2026] report a population-dependent decay on GSM8K and Countdown, with larger N (10–30) delaying its onset and reducing its depth, and explain the decay by heterogeneous curvature under fixed stochasticity in an unshaped quadratic model; Appendix H.4 contrasts this with the V_v mechanism of Section 4.3 under Z-score shaping. Our runs show a qualitatively similar dependence: on Countdown the decay appears at $N = 8$ within 500 iterations and at $N = 15$ within 2000, and a GSM8K run at $N = 30$ also declines (Appendix J.3). At $N = 30$, the final-50 reward means of our two Countdown runs are within 2% of their peak 25-iteration means at 2000 iterations. In both views, stochastic displacement along weakly curved directions accumulates faster at smaller N.

Catastrophic forgetting. Abdi et al. [2026] observe the forgetting in puzzle (c) on HellaSwag while fine-tuning Qwen2.5-1.5B-Instruct on Countdown with ES for 500 iterations. For their setup, the leading factor of Equation (20) is $\sqrt{\eta^2(D-d)T} \approx \sqrt{\eta^2 DT} \approx 80$ (Appendix J.1), compared with a Frobenius norm of approximately 88 read from their Fig. 3; this compares a single realized

norm with an RMS scale and assumes small $|\delta_T|$ and ε_T in Proposition 4.5 (ii). Our runs across seven models and four tasks match their N, σ, and α and also evaluate on HellaSwag: for each model, the mean squared-displacement slope per nonzero update is 1–12% above $\eta^2 D$, and HellaSwag accuracy has a negative slope in 56 of 62 runs (Table 1). These observations are consistent with forgetting that accompanies displacement at the predicted full-parameter scale (Section 4.3); the size of the accuracy loss depends on the evaluation-task curvature (Appendix H.5).

Limitations. We assume a static effective subspace $\mathcal{S}$ for the basin-local analysis and use a small-step-size weak approximation on fixed finite horizons; finite-step effects such as edge-of-stability behavior in full-batch gradient descent [Cohen et al., 2021] require separate analysis. Isotropic diffusion is a condition of the coefficient and radial models, while the exact finite-batch SDE retains its full covariance; extensions to covariance-adapted ES [Hansen and Ostermeier, 2001] remain untested. Table 1 uses two or three seeds per model–task cell, the population and horizon comparisons cover only Qwen2.5-0.5B, and forgetting is measured only on HellaSwag. Neither the fitted reward curves nor the measured full-parameter displacement verify the local geometry, a convergence time, or directional isotropy. Mode hopping between distinct local optima lies outside the basin-local regime analysis.

7 Conclusion

Our SDE approximation provides a unified continuous-time treatment of Z-score-normalized ES over fixed finite horizons. Z-score shaping attains the Fisher-information bound on the population-level shaping ratio under the auxiliary Gaussian law and conditions of Proposition 3.2. In the local progress estimate (Theorem 4.3), the noise baseline scales with d/N, whereas the improvement coefficient depends on neither d nor D; hence the population needed for improvement to exceed this baseline grows with d rather than D, which addresses the dimensionality paradox at the level of these terms. The framework also relates improvement and neutral displacement to the training-reward rise-then-decay of Liang et al. [2026] and the catastrophic forgetting of Abdi et al. [2026], and motivates conditional guidance for stopping and budget allocation (Remark 4.6) by comparing further improvement with accumulated displacement.

Across seven LLMs and four tasks, squared displacement grows approximately linearly, with model-mean slopes per nonzero update up to 12% above the full-parameter nominal rate, and held-out HellaSwag accuracy has a negative slope against the square root of completed updates in 56 of 62 runs. On Countdown, the training reward of Qwen2.5-0.5B declines after its peak at $N = 8$ and 15, whereas the $N = 30$ runs remain near their peaks at 2000 iterations. These observations are consistent with the displacement mechanisms but do not measure d or isolate the neutral directions; estimating d and the evaluation-task curvature in LLMs would help test these mechanisms.

Acknowledgments and Disclosure of Funding

This research was conducted using the Supermicro ARS-111GL-DNHR-LCC and FUJITSU Server PRIMERGY CX2550 M7 (Miyabi) at the Joint Center for Advanced High Performance Computing (JCAHPC). The authors received no third-party funding for this work and declare no competing interests.

References

Immanuel Abdi, Akshat Gupta, Micah Mok, Alexander Lu, Nicholas Lee, and Gopala Anumanchipalli. Evolutionary strategies at scale lead to catastrophic forgetting. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 194–204, 2026. doi: 10.18653/v1/2026.acl-short.18.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, 2021. doi: 10.18653/v1/2021.acl-long.568.

Allen Institute for AI. OLMo-2-0425-1B-Instruct: Model card. Hugging Face, 2025. Revision 48d788ec.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL `https://arxiv.org/abs/2110.14168v1`. Version 1, October 27, 2021. Data source used: `https://huggingface.co/datasets/openai/gsm8k`, training split.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Pierre Del Moral and Angèle Niclas. A Taylor expansion of the square root matrix function. *Journal of Mathematical Analysis and Applications*, 465(1):259–266, 2018. doi: 10.1016/j.jmaa.2018.05.005.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256.

Massimo Fornasier, Hui Huang, Jona Klemenc, and Greta Malaspina. From consensus-based optimization to evolution strategies: Proof of global convergence. *arXiv preprint arXiv:2602.11677*, 2026. URL `https://arxiv.org/abs/2602.11677v1`. Version 1, February 12, 2026.

Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D. Goodman. Stream of search (SoS): Learning to search in language. In *First Conference on Language Modeling*, 2024.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241, 2019.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018. URL `https://arxiv.org/abs/1812.04754v1`. Version 1, December 12, 2018.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001. doi: 10.1162/106365601750190398. URL `https://www.cmap.polytechnique.fr/~nikolaus.hansen/cmaartic.pdf`. Author version with errata.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. Data source used: `https://huggingface.co/datasets/EleutherAI/hendrycks_math`, numeric-answer training subset.

Edward 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.

HuggingFaceTB. SmolLM2-1.7B-Instruct: Model card. Hugging Face, 2025. Revision 31b70e2e.

Kevin Jamieson, Robert Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, volume 25, pages 2672–2680, 2012.

Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968. doi: 10.1038/217624a0.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2101–2110, 2017.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40): 1–47, 2019.

Qiyao Liang, Jinyeop Song, Yizhou Liu, Jeff Gore, Ila Fiete, Risto Miikkulainen, and Xin Qiu. The blessing of dimensionality in LLM fine-tuning: A variance-curvature perspective. *arXiv preprint arXiv:2602.00170*, 2026. URL `https://arxiv.org/abs/2602.00170v1`. Version 1, January 30, 2026.

Vadim Linetsky. The spectral representation of Bessel processes with constant drift: Applications in queueing and finance. *Journal of Applied Probability*, 41(2):327–344, 2004. doi: 10.1239/jap/1082999069.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, pages 53038–53075, 2023. doi: 10.52202/075280-2308.

Meta. Llama-3.2-3B-Instruct: Model card. Hugging Face, 2024. Revision 0cb88a4f.

Microsoft. Phi-4-mini-instruct: Model card. Hugging Face, 2025. Revision cfbefacb.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. doi: 10.1007/s10208-015-9296-2.

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. TinyZero. `https://github.com/Jiayi-Pan/TinyZero`, 2025. GitHub repository.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, 2021. doi: 10.18653/v1/2021.naacl-main.168. Data source used: `https://huggingface.co/datasets/ChilleD/SVAMP`, training split.

Xin Qiu, Yulu Gan, Conor F. Hayes, Qiyao Liang, Yinggan Xu, Roberto Dailey, Elliot Meyerson, Babak Hodjat, and Risto Miikkulainen. Evolution strategies at scale: LLM fine-tuning beyond reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 101987–102015, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL `https://arxiv.org/abs/2412.15115v2`. Version 2, January 3, 2025.

Ingo Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, Stuttgart, 1973.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: Extreme singular values. *arXiv preprint arXiv:1003.2990*, 2010. URL `https://arxiv.org/abs/1003.2990v2`. Version 2, April 7, 2010.

Levent Sagun, Utku Evci, V. Uğur Güney, Yann Dauphin, and Léon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. In *International Conference on Learning Representations, Workshop Track*, 2018.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017. URL `https://arxiv.org/abs/1703.03864v2`. Version 2, September 7, 2017.

Bidipta Sarkar, Mattie Fellows, Juan Agustin Duque, Alistair Letcher, Antonio León Villares, Anya Sims, Clarisse Wibault, Dmitry Samsonov, Dylan Cope, Jarek Liesen, Kang Li, Lukas Seier, Theo Wolf, Uljad Berdica, Valentin Mohl, Alexander David Goldie, Aaron Courville, Karin Sevegnani, Shimon Whiteson, and Jakob N. Foerster. Evolution strategies at the hyperscale. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 107670–107741, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL `https://arxiv.org/abs/1707.06347v2`. Version 2, August 28, 2017.

Hans-Paul Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie: Mit einer vergleichenden Einführung in die Hill-Climbing- und Zufallsstrategie*. Birkhäuser, Basel, 1977.

Anirudh Vemula, Wen Sun, and J. Bagnell. Contrasting exploration in parameter and action space: A zeroth-order optimization perspective. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2926–2935, 2019.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019. doi: 10.18653/v1/P19-1472. Data source used: `https://huggingface.co/datasets/Rowan/hellaswag`, validation split.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 61121–61143, 2024.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7654–7663, 2019.

Contents of the Appendix

A SDE Approximation (Theorem 3.1) **15**

A.1 Exact Coefficients and Coefficient Models 15

A.2 Coefficient Regularity and Proof of Theorem 3.1 16

A.3 Drift-Corrected Second-Order Approximation 18

A.4 Finite Quadratic Decomposition of Reward Variance 19

A.5 Finite-Batch Covariance Trace Bound 20

B Optimality of Z-Score Shaping (Proposition 3.2) **21**

C Cross-Hessian Bounds (Assumption 3.3) **22**

D Proof of Theorem 4.3 **23**

E Proof of Proposition 4.4 **24**

F Isotropic Radial Model and Invariant Radius Statistics **25**

F.1 Invariant-Law Calculation 25

F.2 Scale Matching and Displacement 26

G Evolutionary Analogies **26**

H Task Projections, Stopping, and Budget Allocation (Section 4.3) **26**

H.1 General Local Expansion 26

H.2 Training-Gap Decomposition 27

H.3 Proof of Proposition 4.5 27

H.4 Training Task: Rise-then-Decay 27

H.5 Evaluation Task: Catastrophic Forgetting 28

H.6 Stopping and Budget Allocation (Remark 4.6) 28

I Synthetic Experiments **28**

I.1 Validity of the Finite Quadratic Approximation 28

I.2 First- and Second-Order σ_R Models and Radial Scales 29

I.3 Landscape Dependence: Smooth, Multimodal, Discontinuous, and Oscillatory . . . 30

I.4 Computational Resources 31

J LLM Experiments **32**

J.1 Experimental Setup 32

J.2 Measurements and Diagnostics of the Three Qwen2.5 Trajectories 33

J.3 Models, Tasks, and Training Budgets 34

Table 3: Conditions of the main results, their basis, and their empirical status. Full hypotheses are stated at the cited locations.

Conditions (location)	Results	Basis	Empirical status
Bounded rewards, i.i.d. Gaussian perturbations; fixed N, σ; $\alpha \downarrow 0$ (Theorem 3.1)	Fixed-horizon weak approximation	Task scores are bounded, perturbations are Gaussian by construction, and N and σ are held fixed during training	Proved for the update of Equation (2); the error constant at the LLM step size is not estimated
Density, endpoints, integrability (Proposition 3.2)	Fisher bound; Gaussian equality	Z-score shaping is used in practice [Qiu et al., 2026, Abdi et al., 2026]; the Gaussian law attains the bound, and the efficiency $1/(1+J)$ in Equation (35) measures the loss otherwise	Density conditions and J are not verified for LLM rewards (Remark B.1)
Coefficient and variance approximations (Equations (6) and (11))	Closed-form coefficient models	The projected covariance trace is close to the isotropic value $m/(\sigma^2 N)$ when $c \approx 1$ and $(N + \log(1/c))/m \ll 1$ (Proposition A.5); the two terms of Equation (6) are the linear and quadratic components of σ_R^2 (Proposition A.4)	Close to ES on smooth and oscillatory toy landscapes, with larger differences on multimodal and discontinuous ones (Appendix I); not tested on LLMs
Fixed effective subspace, local geometry (Assumptions 3.3 and 4.2)	Local progress; normalized gradient flow (Theorem 4.3)	Bulk-plus-outlier Hessian spectra, low intrinsic dimension, low-rank adaptation, and curvature measurements of ES-smoothed rewards on Qwen2.5 (Section 2); a PL condition holds near a nondegenerate local maximizer	LLM reward gaps are described by the quadratic reference form in the three Qwen2.5 trajectories (Table 7) and in retained cells (Table 1); d, the PL constant, and the residuals are not measured
Fixed window, projected moment bounds (Propositions 4.4 and A.5)	Total neutral MSD; RMS displacement (Proposition 4.5 (ii))	For $d \ll D$, using $m \approx D$ gives $q_B \lesssim 1.2 \times 10^{-4}$ and, at $c = 1$, $\delta_m \lesssim 5 \times 10^{-3}$ for the reported LLM settings	Full-parameter squared displacement grows approximately linearly, with cell-mean slopes 0.5–12% above the nominal rate (Table 8); neutral projections and directional isotropy are not measured
Inward radial drift, isotropic noise (Theorem F.1)	Invariant Gamma law (Proposition F.2)	Motivated by the coefficient model of Equation (11) for an isotropic quadratic neutral reward centered at the origin, with the neutral gradient dominating σ_R and cross terms neglected; the radial model is specified separately (Appendix F)	Toy MSD plateaus, but the closed-form prefactor differs (Appendix I.2); the LLM runs have $Nh \ll D - d$ and show no saturation

A SDE Approximation (Theorem 3.1)

This appendix proves Theorem 3.1, gives its drift-corrected second-order refinement, and collects the reward-variance decomposition and finite-batch coefficient bounds used in the main text.

A.1 Exact Coefficients and Coefficient Models

By exchangeability of the samples, $b_N(\theta) = \sigma^{-1}\mathbb{E}[F(R_1)\epsilon_1]$. This expectation does not in general factor, because $F(R_1)$ depends on ϵ_1 through $R_1 = R(\theta + \sigma\epsilon_1)$ and on the other samples through μ_R and s_R. The closed-form coefficient models replace μ_R and s_R by $R_\sigma(\theta)$ and $\sigma_R(\theta)$; Stein's lemma $\mathbb{E}_\epsilon[R(\theta+\sigma\epsilon)\epsilon] = \sigma\nabla R_\sigma(\theta)$ [Nesterov and Spokoiny, 2017] then gives $b_N \approx \nabla R_\sigma/\sigma_R$, and the isotropic model uses $\Sigma_N \approx I_D/(\sigma^2 N)$. For the exact coefficients, the Z-score normalization imposes the batch identity $\frac{1}{N}\sum_n F(R_n)^2 = \mathbf{1}_{\{s_R>0\}}$, so Cauchy–Schwarz gives

$$\|\hat{g}(\theta)\|^2 \le (\sigma^2 N)^{-1} \sum_n \|\epsilon_n\|^2, \tag{21}$$

and hence $\mathbb{E}\|\hat{g}(\theta)\|^2 \le D/\sigma^2$.

A.2 Coefficient Regularity and Proof of Theorem 3.1

Lemma A.1 (Regularity of the SDE coefficients)**.** *In the setting of Theorem 3.1, the entries of b_N and Σ_N are smooth with bounded derivatives of every order, and $\Sigma_N(\theta)$ is positive definite at every $\theta \in \mathbb{R}^D$. Its principal square root $A_N = \Sigma_N^{1/2}$ is smooth and has bounded derivatives on each compact set. Moreover,*

$$\sup_\theta \|b_N(\theta)\| \le \frac{\sqrt{D}}{\sigma}, \qquad \sup_\theta \|A_N(\theta)\|_F^2 \le \frac{D}{\sigma^2}.$$

Proof. Set $z_n = \theta + \sigma\epsilon_n$ and write the moment integrals against the joint Gaussian density $p_\theta(z)$, where $z = (z_1, \dots, z_N)$. The batch weights $F_n(z) := F(R(z_n))$ are bounded Borel functions of z, and $\sum_n F_n(z) = 0$ gives

$$\hat{g}(\theta; z) = \frac{1}{\sigma^2 N} \sum_n F_n(z) z_n.$$

For fixed z, the displayed expression for $\hat{g}(\theta; z)$ is independent of θ, so differentiation of the moment integrals acts only on p_θ. On bounded sets of θ, every differentiated first- and second-moment integrand is dominated by a polynomial in $\|z\|$ times $\exp\{-\|z\|^2/(4\sigma^2)\}$. This justifies differentiation under the integral. After differentiating the density and substituting $z_n = \theta + \sigma\epsilon_n$, each density derivative divided by p_θ is a polynomial in the Gaussian samples with coefficients independent of θ. The pointwise bound in Equation (21) and Gaussian polynomial moments therefore bound each moment derivative uniformly in θ. The product rule gives uniformly bounded derivatives of Σ_N after subtracting $b_N b_N^\top$.

Translated nondegenerate Gaussian measures are mutually equivalent, so the assumed positive population variance at one state implies $\sigma_R(\theta)^2 > 0$ at every θ. To prove positive definiteness, fix $v \neq 0$, write $U_i = v^\top \epsilon_i$, and set $H = \sum_i F(R_i) U_i$. If $\mathrm{Var}(H) = 0$, then $H = c$ almost surely. However,

$$|H| \le \sqrt{N} \left\{ \sum_i U_i^2 \right\}^{1/2},$$

and the right-hand side is arbitrarily small with positive probability, so $c = 0$. Let $C = \sum_i (R_i - \mu_R) U_i$. When $s_R > 0$, $C = s_R H$, while both sides vanish when $s_R = 0$; hence $C = 0$ almost surely. Expanding C^2 using independence of the sampled pairs (R_i, U_i) across i and discarding nonnegative terms gives

$$\mathbb{E}C^2 \ge \frac{N-1}{N} \sigma_R(\theta)^2 \|v\|^2 > 0.$$

This contradicts $C = 0$. Thus $v^\top \Sigma_N v = \mathrm{Var}(H)/(\sigma^2 N^2) > 0$.

The principal square-root map is smooth on the cone of positive-definite matrices [Del Moral and Niclas, 2018, Theorem 1.1]. Thus A_N is smooth, and its derivatives are bounded on compact sets by continuity. Finally, Jensen's inequality and $\mathrm{Tr}\, \Sigma_N \le \mathbb{E}\|\hat{g}\|^2 \le D/\sigma^2$ give the stated bounds. □

Localization. Work in the setting of Theorem 3.1, with $0 < \alpha \le \min\{1, T\}$, and write $K_\alpha = \lfloor T/\alpha \rfloor$. The preceding lemma gives a unique nonexplosive strong solution of Equation (4), since its coefficients are locally Lipschitz and bounded. Let x_t solve $\dot{x}_t = b_N(x_t)$ with $x_0 = \theta_0$. The bounded, globally Lipschitz drift gives a global solution with $\|x_t - \theta_0\| \le T\sqrt{D}/\sigma$ for $t \le T$. Choose a bounded open neighborhood V of this trajectory and $\rho > 0$ such that $\overline{U}_\rho \subset V$, where $U_\rho = \{z : \mathrm{dist}(z, \{x_t : 0 \le t \le T\}) < \rho\}$. Choose $\chi \in C_c^\infty(V)$ with $0 \le \chi \le 1$ and $\chi = 1$ on a neighborhood of $\overline{U}_\rho$, with V, ρ, and χ fixed independently of α. Define the cut-off update and coefficients by

$$\bar{G} = \chi \hat{g}, \qquad \bar{b} = \chi b_N, \qquad \bar{A} = \chi A_N. \tag{22}$$

At each state, $\mathbb{E}\bar{G} = \bar{b}$ and $\mathrm{Cov}(\bar{G}) = \bar{A}\bar{A}^\top$. The preceding lemma and the compact support of χ make $\bar{b}$, $\bar{A}$ smooth with bounded derivatives on the full space. Let the cut-off processes start from θ_0 and satisfy

$$\bar{\theta}_{k+1} = \bar{\theta}_k + \alpha \bar{G}(\bar{\theta}_k),$$
$$d\bar{\Theta}_t^\alpha = \bar{b}(\bar{\Theta}_t^\alpha)\, dt + \sqrt{\alpha}\, \bar{A}(\bar{\Theta}_t^\alpha)\, dW_t.$$

Couple the two discrete processes by the same fresh batches, and the two SDEs by the same Brownian motion. Each pair agrees until its first exit from U_ρ; the discrete iterates also agree at the exit step.

Exit estimates. For each fixed integer $m \geq 1$, the pointwise bound in Equation (21) gives uniform bounds on the conditional $2m$-th moments of the centered cut-off updates. Independence of fresh batches makes $\mathcal{M}_k = \alpha \sum_{j<k} \{\bar{G}(\bar{\theta}_j) - \bar{b}(\bar{\theta}_j)\}$ a martingale. The discrete Burkholder–Davis–Gundy and power-mean inequalities give

$$\mathbb{E} \max_{k \leq K_\alpha} \|\mathcal{M}_k\|^{2m} \leq C_m \alpha^{2m} K_\alpha^m \leq C_m T^m \alpha^m.$$

Since $\chi(x_t) = 1$, we have $\bar{b}(x_t) = b_N(x_t)$. The one-step ODE remainder is $O(\alpha^2)$, so discrete Gronwall gives the first bound below. For the SDE, $\|\bar{A}\|_F^2 \leq D/\sigma^2$ and the continuous Burkholder–Davis–Gundy and Gronwall inequalities give the second:

$$\begin{aligned} \mathbb{E} \max_{k \leq K_\alpha} \|\bar{\theta}_k - x_{k\alpha}\|^{2m} &\leq C_m \alpha^m, \\ \mathbb{E} \sup_{t \leq T} \|\bar{\Theta}_t^\alpha - x_t\|^{2m} &\leq C_m \alpha^m. \end{aligned} \tag{23}$$

At the first exit from U_ρ, the corresponding process is at least ρ from the deterministic trajectory at that time. Markov's inequality, Equation (23), and the coupling therefore yield

$$\mathbb{P}\{\theta_k \notin U_\rho \text{ for some } k \leq K_\alpha\} + \mathbb{P}\{\Theta_t^\alpha \notin U_\rho \text{ for some } t \leq T\} \leq C_m \alpha^m. \tag{24}$$

The same bound holds for exit from V, since $U_\rho \subset V$. The constants may depend on the fixed parameters, m, V, ρ, and χ, but not on α.

Proof of Theorem 3.1. We compare the cut-off processes with the deterministic trajectory x_t and then remove the cut-off using Equation (24). First, all four processes have uniformly bounded moments on the fixed horizon: for every integer $p \geq 2$,

$$\mathbb{E} \max_{k \leq K_\alpha} \big(\|\theta_k\|^p + \|\bar{\theta}_k\|^p\big) + \mathbb{E} \sup_{t \leq T} \big(\|\Theta_t^\alpha\|^p + \|\bar{\Theta}_t^\alpha\|^p\big) \leq C_p. \tag{25}$$

For the iterates, sum the update bounds in Equation (21) and apply the power-mean inequality: the p-th moment of the total displacement is at most $C_p(K_\alpha \alpha)^p \leq C_p T^p$ for either process, since $0 \leq \chi \leq 1$. For the SDEs, the bounded coefficients in Lemma A.1, $\alpha \leq 1$, and the Burkholder–Davis–Gundy inequality give the same uniform moment bound.

To control the mean deviations, set $e_k = \mathbb{E}\bar{\theta}_k - x_{k\alpha}$ and $e(t) = \mathbb{E}\bar{\Theta}_t^\alpha - x_t$. Taylor expansion of $\bar{b}$ about x_t has a remainder bounded by $C\|y - x_t\|^2$, since its second derivatives are bounded. Using Equation (23) with $m = 1$, the exact conditional mean $\mathbb{E}\bar{G} = \bar{b}$, and the $O(\alpha^2)$ one-step ODE remainder gives

$$\begin{aligned} e_{k+1} &= e_k + \alpha\, \partial\bar{b}(x_{k\alpha}) e_k + O(\alpha^2), \qquad k < K_\alpha, \\ e(t) &= \int_0^t \partial\bar{b}(x_s) e(s)\, ds + O(\alpha), \qquad 0 \leq t \leq T. \end{aligned}$$

Here $\partial\bar{b}$ is the Jacobian, and both remainders are uniform over the indicated indices and times. The second identity follows by taking expectations in the SDE integral equation; boundedness of the coefficients justifies the interchange and makes the stochastic integral mean zero. Since $e_0 = e(0) = 0$, the discrete and continuous Gronwall inequalities yield

$$\max_{k \leq K_\alpha} \|e_k\| + \sup_{t \leq T} \|e(t)\| \leq C\alpha.$$

Choose an integer $q \geq 1$ such that $|\varphi(z)| + \|\nabla\varphi(z)\| + \|\nabla^2\varphi(z)\| \leq C(1 + \|z\|^q)$. For $Y = \bar{\theta}_k$ at $t = k\alpha$, or $Y = \bar{\Theta}_t^\alpha$ at $t \leq T$, the second-order Taylor remainder about x_t is bounded by $C(1 + \|Y\|^q)\|Y - x_t\|^2$, because x_t stays in a fixed bounded set. Cauchy–Schwarz, Equation (25) with $p = 2q$, and Equation (23) with $m = 2$ give

$$\big|\mathbb{E}\varphi(Y) - \varphi(x_t) - \nabla\varphi(x_t)^\top \mathbb{E}(Y - x_t)\big| \leq C\big\{\mathbb{E}(1 + \|Y\|^q)^2\big\}^{1/2} \big\{\mathbb{E}\|Y - x_t\|^4\big\}^{1/2} \leq C\alpha.$$

Together with the mean-deviation bounds, this proves $|\mathbb{E}\varphi(Y) - \varphi(x_t)| \leq C\alpha$ and hence

$$\max_{k \leq K_\alpha} \big|\mathbb{E}\varphi(\bar{\theta}_k) - \mathbb{E}\varphi(\bar{\Theta}_{k\alpha}^\alpha)\big| \leq C\alpha.$$

Finally, each original process agrees with its cut-off counterpart on paths that do not exit U_ρ over the relevant horizon; the discrete pair also agrees at the update that crosses the boundary. For either coupled pair $(X, \bar{X})$ at a comparison time, let E be its exit event in Equation (24). The moment bound with $p = 2q$ and the exit bound with $m = 2$ imply

$$\left|\mathbb{E}\varphi(X) - \mathbb{E}\varphi(\bar{X})\right| \le \left\{\mathbb{E}\left(|\varphi(X)| + |\varphi(\bar{X})|\right)^2\right\}^{1/2} \mathbb{P}(E)^{1/2} \le C\alpha.$$

These bounds are uniform for $k \le K_\alpha$ and $t \le T$. Since $K_\alpha \alpha \le T$, the triangle inequality proves Equation (5), with a constant that may depend on φ and the fixed parameters but not on α. □

A.3 Drift-Corrected Second-Order Approximation

Matching the first two one-step raw moments to order $O(\alpha^3)$ identifies a diffusion covariance and a drift correction. We first give the coefficient calculation under global regularity, which we use below for the cut-off processes.

Proposition A.2 (Identification by one-step moment matching)**.** *Let $G(x)$ have mean $b(x)$ and covariance $\Sigma(x)$, and let $\widetilde{\Delta} = Z_\alpha^\alpha - x$ for the SDE*

$$dZ_t^\alpha = (b + \alpha b_1)(Z_t^\alpha)\, dt + \sqrt{\alpha} A(Z_t^\alpha)\, dW_t, \qquad Z_0^\alpha = x.$$

Suppose b, b_1, A are independent of α, globally Lipschitz, and C^3 with polynomial-growth derivatives. The first and second raw moments of $\Delta = \alpha G(x)$ and $\widetilde{\Delta}$ agree with residuals bounded by $C(1 + \|x\|^r)\alpha^3$, for some C, r independent of x and $0 < \alpha \le 1$, if and only if

$$AA^\top = \Sigma, \qquad b_1 = -\tfrac{1}{2}(\partial b) b. \tag{26}$$

Proof. The discrete moments are $\mathbb{E}\Delta = \alpha b(x)$ and $\mathbb{E}[\Delta\Delta^\top] = \alpha^2\{\Sigma(x) + b(x)b(x)^\top\}$. Writing $a = AA^\top$, the SDE generator is $\mathcal{L}_\alpha = (b + \alpha b_1) \cdot \nabla + (\alpha/2) a : \nabla^2$. The one-step expansions of Li et al. [2019, Lemma 4, under Assumption 4.1] give

$$\begin{aligned} \mathbb{E}\widetilde{\Delta} &= \alpha b(x) + \alpha^2\{b_1(x) + \tfrac{1}{2}(\partial b(x)) b(x)\} + O(\alpha^3), \\ \mathbb{E}[\widetilde{\Delta}\widetilde{\Delta}^\top] &= \alpha^2\{a(x) + b(x)b(x)^\top\} + O(\alpha^3). \end{aligned}$$

Indeed, for $f_i(z) = z_i - x_i$ and $f_{ij}(z) = (z_i - x_i)(z_j - x_j)$, the terms through order α^2 are $\alpha\mathcal{L}_\alpha f + \alpha^2\mathcal{L}_\alpha^2 f/2$, evaluated at x. In particular, $\mathcal{L}_\alpha f_{ij}(x) = \alpha a_{ij}(x)$ and $\mathcal{L}_\alpha^2 f_{ij}(x) = 2b_i(x)b_j(x) + O(\alpha)$. The coefficient assumptions give uniform linear growth, Lipschitz bounds, and polynomial moment bounds for the SDE; the remainders are bounded by $C(1 + \|x\|^r)\alpha^3$. If the moments agree to this order, division of their differences by α^2 and passage to $\alpha \downarrow 0$ at each fixed x gives Equation (26). Conversely, these identities cancel both coefficients of α^2, leaving the stated remainders. □

For ES, Equation (26) gives $A_N A_N^\top = \Sigma_N$ and $b_{1,N} = -\frac{1}{2}(\partial b_N) b_N$ wherever the cut-off equals one. The covariance is identified, while an orthogonal right factor leaves $A_N A_N^\top$ unchanged. We choose the principal square root $A_N = \Sigma_N^{1/2}$ for the following approximation. Define the corrected SDE by

$$dY_t^\alpha = (b_N + \alpha b_{1,N})(Y_t^\alpha)\, dt + \sqrt{\alpha} A_N(Y_t^\alpha)\, dW_t, \qquad Y_0^\alpha = \theta_0. \tag{27}$$

Theorem A.3 (Second-order weak approximation of Z-score-normalized ES)**.** *In the setting of Theorem 3.1, the corrected SDE Equation* (27) *satisfies*

$$\max_{0 \le k \le \lfloor T/\alpha \rfloor} |\mathbb{E}\varphi(\theta_k) - \mathbb{E}\varphi(Y_{k\alpha}^\alpha)| = O(\alpha^2) \qquad \text{as } \alpha \downarrow 0, \tag{28}$$

for every $\varphi \in C^3(\mathbb{R}^D)$ whose derivatives of orders zero through three have polynomial growth. The constant may depend on the fixed $R, D, N, \sigma, T, \theta_0$ and φ, but not on α.

Proof. Use K_α, x_t, U_ρ, and the fixed cut-off χ from the localization above, with $0 < \alpha \le \min\{1, T\}$. Since b_N and all its derivatives are bounded, $b_{1,N}$ is smooth and bounded with bounded derivatives. Thus Equation (27) has a unique nonexplosive strong solution: its coefficients are locally Lipschitz and bounded.

Cut-off correction and moments. For $\bar{b}, \bar{A}, \bar{G}$ in Equation (22), recompute the correction from $\bar{b}$:

$$\bar{b}_1 = -\tfrac{1}{2}(\partial\bar{b})\bar{b} = \chi^2 b_{1,N} - \tfrac{1}{2}\chi\,(b_N \cdot \nabla\chi)b_N.$$

It equals $b_{1,N}$ on a neighborhood of $\overline{U}_\rho$, and generally differs from $\chi b_{1,N}$ elsewhere. Let $\bar{Y}^\alpha$ start from θ_0 with drift $\bar{b} + \alpha\bar{b}_1$ and diffusion factor $\sqrt{\alpha}\bar{A}$. These coefficients are globally Lipschitz with bounded derivatives of every order, uniformly in α.

The bound Equation (21) gives $\|\bar{G}(z)\| \le Q$, where $Q = (\sigma^2 N)^{-1/2}(\sum_n \|\epsilon_n\|^2)^{1/2}$ has moments of every order independent of z. Starting the cut-off chain from any z, Minkowski's inequality gives, for every integer $m \ge 1$,

$$\left(\mathbb{E}\max_{k\le K_\alpha}\|\bar{\theta}^z_k\|^{2m}\right)^{1/(2m)} \le \|z\| + \alpha K_\alpha(\mathbb{E}Q^{2m})^{1/(2m)} \le \|z\| + T(\mathbb{E}Q^{2m})^{1/(2m)}.$$

Here $\bar{\theta}^z$ uses independent fresh batches and has initial value z. The same bound holds for the original chain. Boundedness of the drifts and diffusion factors and the Burkholder–Davis–Gundy inequality also bound all even maximal moments of Y^α and $\bar{Y}^\alpha$ uniformly in α over $[0, T]$.

One-step and finite-time comparison. At every state, $\mathbb{E}\bar{G} = \bar{b}$ and $\mathrm{Cov}(\bar{G}) = \chi^2\Sigma_N = \bar{A}\bar{A}^\top$. Apply Proposition A.2 with $G = \bar{G}$, $b = \bar{b}$, $b_1 = \bar{b}_1$, and $A = \bar{A}$. The first two raw-moment differences are $O(\alpha^3)$ with polynomial bounds in the initial state. For any three coordinate indices i, j, l, the discrete increment satisfies

$$\mathbb{E}|\Delta_i\Delta_j\Delta_l| \le \mathbb{E}\|\alpha\bar{G}\|^3 \le \alpha^3\mathbb{E}Q^3.$$

For the cut-off SDE $\bar{Y}^{\alpha,z}$ started at any z, bounded coefficients and the Burkholder–Davis–Gundy inequality give $\mathbb{E}\|\bar{Y}^{\alpha,z}_\alpha - z\|^p \le C_p\alpha^p$ for every $p \ge 2$, which also bounds its third absolute coordinate-product moments. Thus the cut-off coefficients satisfy Assumption 4.1 of Li et al. [2019] uniformly in α, and their derivatives through order three satisfy the growth bounds used in its Proposition 25. The one-step estimates and the arbitrary-initial-state even-moment bound verify the hypotheses of its Theorem 3 at order two. No mollification is needed, so its mollification residual is zero. Consequently,

$$\max_{k\le K_\alpha}\left|\mathbb{E}\varphi(\bar{\theta}_k) - \mathbb{E}\varphi(\bar{Y}^\alpha_{k\alpha})\right| \le C\alpha^2.$$

Removal of the cut-off. Couple Y^α and $\bar{Y}^\alpha$ with the same Brownian motion; they agree until the first exit from U_ρ. In the comparison of $\bar{Y}^\alpha$ with x_t, the additional drift contributes at most $\alpha T\|\bar{b}_1\|_\infty$, so the argument for Equations (23) and (24) bounds the exit probability of $\bar{Y}^\alpha$ by $C_m\alpha^m$. The discrete exit bound is Equation (24); the original and cut-off chains use the same batches and agree at the exit update, which starts inside U_ρ. Taking $m = 4$ makes both exit probabilities $O(\alpha^4)$, and the uniform moment bounds above, the polynomial growth of φ, and Cauchy–Schwarz on the exit events give

$$\max_{k\le K_\alpha}\left|\mathbb{E}\varphi(\theta_k) - \mathbb{E}\varphi(\bar{\theta}_k)\right| + \sup_{t\le T}\left|\mathbb{E}\varphi(Y^\alpha_t) - \mathbb{E}\varphi(\bar{Y}^\alpha_t)\right| \le C\alpha^2.$$

Combining these estimates with the cut-off comparison proves Equation (28). □

Proposition A.2 identifies the diffusion covariance by matching one-step moments to second order: $A_N A_N^\top = \Sigma_N$. The SDE of Theorem 3.1 already has this covariance, and Theorem A.3 shows that adding the drift correction $\alpha b_{1,N}$ yields second-order weak accuracy.

A.4 Finite Quadratic Decomposition of Reward Variance

Proposition A.4 (Finite quadratic decomposition of population reward variance)**.** *Let $R : \mathbb{R}^D \to \mathbb{R}$ be bounded and Borel measurable, and fix $\theta \in \mathbb{R}^D$ and $\sigma > 0$. For $\epsilon \sim \mathcal{N}(0, I_D)$, write $X = R(\theta+\sigma\epsilon)$, $\mu = R_\sigma(\theta)$, $g = \nabla R_\sigma(\theta)$, and $H = \mathcal{H}(\theta)$, and define*

$$r_2 = X - \mu - \sigma g^\top\epsilon - \frac{\sigma^2}{2}(\epsilon^\top H\epsilon - \mathrm{Tr}\,H).$$

Then

$$\sigma_R(\theta)^2 = \sigma^2\|g\|^2 + \frac{\sigma^4}{2}\|H\|_F^2 + \mathbb{E}r_2^2. \tag{29}$$

The first two terms are the variances of the linear and centered quadratic components of the reward, and $\mathbb{E}r_2^2$ is the remaining variance. In particular, the mean squared residual of the first-order reward approximation is

$$\mathbb{E}[(X-\mu-\sigma g^\top\epsilon)^2] = \frac{\sigma^4}{2}\|H\|_F^2 + \mathbb{E}r_2^2, \qquad \sigma_R(\theta)^2 \geq \sigma^2\|g\|^2. \tag{30}$$

For $g \neq 0$, its relative mean squared residual is the right-hand side of the first equality divided by $\sigma^2\|g\|^2$. For $g = 0$, the identities apply without taking this ratio.

Proof. Let $p_\theta(y)$ be the density of $\mathcal{N}(\theta, \sigma^2 I_D)$. Its first two derivatives are

$$\partial_{\theta_i} p_\theta(y) = \frac{y_i-\theta_i}{\sigma^2}p_\theta(y), \qquad \partial_{\theta_i\theta_j} p_\theta(y) = \left(\frac{(y_i-\theta_i)(y_j-\theta_j)}{\sigma^4} - \frac{\delta_{ij}}{\sigma^2}\right)p_\theta(y).$$

On each compact neighborhood of θ, these derivatives are bounded by an integrable function of the form $C(1+\|y\|^2)e^{-\|y\|^2/(4\sigma^2)}$. Since R is bounded, differentiation under the integral and $y = \theta+\sigma\epsilon$ give

$$\mathbb{E}[(X-\mu)\epsilon] = \sigma g, \qquad \mathbb{E}[(X-\mu)(\epsilon\epsilon^\top - I_D)] = \sigma^2 H.$$

Set $L = \sigma g^\top\epsilon$ and $Q_2 = (\sigma^2/2)(\epsilon^\top H\epsilon - \operatorname{Tr} H)$. The odd Gaussian moments vanish, and the fourth moments satisfy $\mathbb{E}[\epsilon_i\epsilon_j\epsilon_k\epsilon_l] = \delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}$. Thus L and Q_2 have mean zero and

$$\mathbb{E}[LQ_2] = 0, \qquad \mathbb{E}[L^2] = \mathbb{E}[(X-\mu)L] = \sigma^2\|g\|^2, \qquad \mathbb{E}[Q_2^2] = \mathbb{E}[(X-\mu)Q_2] = \frac{\sigma^4}{2}\|H\|_F^2.$$

It follows that $r_2 = X-\mu-L-Q_2$ has mean zero and is orthogonal to both L and Q_2 in L^2. Expanding $\mathbb{E}[(L+Q_2+r_2)^2]$ and $\mathbb{E}[(Q_2+r_2)^2]$ proves Equations (29) and (30). □

A.5 Finite-Batch Covariance Trace Bound

We evaluate the covariance Σ_N of the full update in Equation (2), using the batch standard deviation s_R and the convention $F(R_n) = 0$ when $s_R = 0$. Write $c_N(\theta) := \Pr(s_R > 0)$ for the probability of a nonzero batch variance at a fixed state.

Proposition A.5 (Finite-batch covariance trace bound)**.** *Let R be bounded and Borel measurable, and fix $\theta \in \mathbb{R}^D$, $N \geq 2$, and $\sigma > 0$. For independent standard Gaussian perturbations, fix an orthogonal projection $\Pi = \Pi^\top = \Pi^2$ independently of the current batch, with $m = \operatorname{rank}\Pi \geq 1$. Write $n = N-1$ and $c = c_N(\theta)$. If $c > 0$, then*

$$\left|\operatorname{Tr}(\Pi\Sigma_N(\theta)\Pi) - \frac{mc}{\sigma^2 N}\right| \leq \frac{mc}{\sigma^2 N}\delta_m(c), \tag{31}$$

where $x_m(c) = (n+\log(2/c))/m$ and $\delta_m(c) = 10\{\sqrt{x_m(c)} + x_m(c)\} + n/m$. The mean satisfies

$$\|b_N(\theta)\|^2 \leq \frac{nc}{\sigma^2 N}. \tag{32}$$

If $c = 0$, the update is zero almost surely and $b_N(\theta) = 0$, $\Sigma_N(\theta) = 0$.

Taking $\Pi = I_D$ gives the total covariance trace. For the fixed neutral projection $Q = I_D - P$ in Assumption 3.3, taking $\Pi = Q$ gives the same bound with $m = D-d \geq 1$. The condition $(N+\log(1/c))/m \ll 1$ is sufficient for a small relative error in Equation (31); the bound need not be small for a low-rank projection. Comparison with the nominal coefficient $m/(\sigma^2 N)$ also includes the zero-variance probability:

$$\left|\frac{\sigma^2 N}{m}\operatorname{Tr}(\Pi\Sigma_N(\theta)\Pi) - 1\right| \leq (1-c) + c\delta_m(c), \qquad c > 0.$$

Proof. Fix θ, write $A = \{s_R > 0\}$, and let $z_i = F(R_i)$ and $E = [\epsilon_1, \ldots, \epsilon_N] \in \mathbb{R}^{D\times N}$. Choose a deterministic matrix $U_0 \in \mathbb{R}^{N\times n}$ whose columns form an orthonormal basis of $\mathbf{1}^\perp$. The batch identities $\mathbf{1}^\top z = 0$ and $\|z\|^2 = N\mathbf{1}_A$ imply, with $G = EU_0$ and $w = U_0^\top z$,

$$z = U_0 w, \qquad \|w\|^2 = N\mathbf{1}_A, \qquad \hat{g}(\theta) = \frac{Gw}{\sigma N}.$$

The entries of G are independent standard Gaussian variables. Writing $\Pi = UU^\top$ with $U^\top U = I_m$, the entries of $U^\top G \in \mathbb{R}^{m\times n}$ are also independent standard Gaussians. The weights w may depend on G.

Second moment. Set $K = G^\top \Pi G$ and $\Delta = \|K - mI_n\|_{\mathrm{op}}$. For every batch,

$$\left| \|\Pi \hat{g}(\theta)\|^2 - \frac{m\mathbf{1}_A}{\sigma^2 N} \right| = \frac{|w^\top (K - mI_n) w|}{\sigma^2 N^2} \le \frac{\mathbf{1}_A \Delta}{\sigma^2 N}.$$

This quadratic form includes all cross-sample terms of the update. For $m \ge n$, the Gaussian singular-value bound of Rudelson and Vershynin [2010, Theorem 2.6 and Eq. (2.3)] places the singular values of $U^\top G$ between $(\sqrt{m} - \sqrt{n} - u)_+$ and $\sqrt{m} + \sqrt{n} + u$ with probability at least $1 - 2e^{-u^2/2}$, where $u \ge 0$ and $a_+ = \max(a, 0)$. For $m < n$, the same upper bound follows by applying the result to the transpose, while the lower eigenvalue bound for K is zero. In both cases, with $r_{m,n}(u) = 2\sqrt{m}(\sqrt{n} + u) + (\sqrt{n} + u)^2$,

$$\Pr\{\Delta > r_{m,n}(u)\} \le 2e^{-u^2/2}.$$

For $c > 0$, let $q_c = \sqrt{2\log(2/c)}$. Using $\mathbf{1}_A \Delta \le \mathbf{1}_A r_{m,n}(q_c) + (\Delta - r_{m,n}(q_c))_+$ and integrating the tail gives

$$\begin{aligned} \mathbb{E}[\mathbf{1}_A \Delta] &\le c\, r_{m,n}(q_c) + \int_{q_c}^{\infty} 2e^{-u^2/2} r'_{m,n}(u)\, du \\ &\le c \left\{ r_{m,n}(q_c) + \frac{2(\sqrt{m} + \sqrt{n})}{q_c} + 2 \right\} =: c B_{m,n}(c). \end{aligned}$$

Here $r'_{m,n}(u) = 2(\sqrt{m} + \sqrt{n} + u)$, $\int_q^\infty e^{-u^2/2}\, du \le e^{-q^2/2}/q$, and $2e^{-q_c^2/2} = c$. This bound retains the dependence between A and K. Consequently, setting $\varepsilon_{m,n}(c) = B_{m,n}(c)/m$ yields

$$\left| \frac{\sigma^2 N \mathbb{E}\|\Pi \hat{g}(\theta)\|^2}{mc} - 1 \right| \le \varepsilon_{m,n}(c).$$

Mean and covariance. For every deterministic unit vector v, $G^\top v \sim \mathcal{N}(0, I_n)$. Cauchy–Schwarz gives

$$|v^\top b_N(\theta)| \le \frac{\mathbb{E}[\|G^\top v\| \|w\|]}{\sigma N} \le \frac{\sqrt{\mathbb{E}\|G^\top v\|^2\, \mathbb{E}\|w\|^2}}{\sigma N} = \frac{\sqrt{nc}}{\sigma \sqrt{N}}.$$

Taking the supremum over v proves Equation (32). Subtracting $\|\Pi b_N(\theta)\|^2$ from the second moment gives the sharper bound

$$-\varepsilon_{m,n}(c) - \frac{n}{m} \le \frac{\sigma^2 N \operatorname{Tr}(\Pi \Sigma_N(\theta) \Pi)}{mc} - 1 \le \varepsilon_{m,n}(c). \tag{33}$$

To obtain Equation (31), put $a = n + \log(2/c)$. Since $q_c > 1$, $a \ge 1$, and $(\sqrt{n} + q_c)^2 \le 3a$, the expression for $B_{m,n}(c)$ gives

$$B_{m,n}(c) \le (2\sqrt{3} + 2)\sqrt{ma} + 7a \le 10(\sqrt{ma} + a).$$

Combining this with Equation (33) proves the claim for $c > 0$. If $c = 0$, then $z = 0$ almost surely, which proves the remaining case. □

B Optimality of Z-Score Shaping (Proposition 3.2)

We prove Proposition 3.2 under its stated density and shaping conditions, with θ fixed.

Integration by parts. Cauchy–Schwarz gives $\mathbb{E}|\Psi(Y)s(Y)| \le (B_\Psi^2 \mathcal{I}(f_R))^{1/2} < \infty$. On each support interval (a, b), local absolute continuity and the vanishing endpoint products therefore give

$$\int_a^b \Psi'(r) f_R(r)\, dr = [\Psi(r) f_R(r)]_a^b - \int_a^b \Psi(r) f'_R(r)\, dr = -\int_a^b \Psi(r) f'_R(r)\, dr.$$

The integrability of Ψ' ($|A_\Psi| < \infty$) and of Ψs makes these integrals absolutely summable over the support intervals, yielding $A_\Psi = \mathbb{E}[\Psi(Y)s(Y)]$.

Fisher-information bound. The Cauchy–Schwarz inequality yields

$$A_\Psi^2 = \mathbb{E}[\Psi(Y)s(Y)]^2 \le \mathbb{E}[\Psi(Y)^2]\,\mathbb{E}[s(Y)^2] = B_\Psi^2 \mathcal{I}(f_R),$$

with equality iff $\Psi = cs$ almost surely for a single nonzero constant common to all intervals. Thus score shaping attains the bound whenever the score itself is admissible.

Gaussian case. For $f_R = \mathcal{N}(R_\sigma, \sigma_R^2)$, $s(r) = (r - R_\sigma)/\sigma_R^2$ and $\mathcal{I}(f_R) = 1/\sigma_R^2$. The score is C^1, its derivative is integrable, and its product with the Gaussian density vanishes at both infinite endpoints. The optimal shaping is $\Psi^* = cs$ for $c \neq 0$; normalizing $B_{\Psi^*}^2 = 1$ gives $c = \pm\sigma_R$. The positive choice yields $\Psi_Z(r) = (r - R_\sigma)/\sigma_R$, while both signs attain the same squared objective. □

Non-Gaussian efficiency. Suppose in addition that $0 < \sigma_R^2 = \mathrm{Var}(Y) < \infty$, and set $Z = (Y - R_\sigma)/\sigma_R$ with density p. The density $p(z) = \sigma_R f_R(R_\sigma + \sigma_R z; \theta)$ inherits the density conditions of Proposition 3.2. Assume $zp(z) \to 0$ at every support endpoint, including infinite endpoints. The identity shaping $z \mapsto z$ is then admissible: it is C^1, its derivative is 1, and $\mathbb{E}Z^2 = 1$. Define the squared score distance by

$$J := \mathbb{E}[(s_p(Z) - Z)^2], \qquad s_p = -\partial_z \log p. \tag{34}$$

The change of scale gives $s_p(Z) = \sigma_R s(Y)$ and $\mathcal{I}(p) = \sigma_R^2 \mathcal{I}(f_R)$. Integration by parts for the identity shaping gives $\mathbb{E}[Z s_p(Z)] = 1$, so expanding the square in Equation (34) yields $J = \mathcal{I}(p) - 1$. Since $A_{\Psi_Z} = 1/\sigma_R$ and $B_{\Psi_Z}^2 = 1$, the efficiency relative to the Fisher upper bound is

$$e_Z := \frac{A_{\Psi_Z}^2 / B_{\Psi_Z}^2}{\mathcal{I}(f_R)} = \frac{1}{\sigma_R^2 \mathcal{I}(f_R)} = \frac{1}{1 + J}. \tag{35}$$

When the score itself is admissible, this is also the ratio to the optimal population value.

Remark B.1 (Conditions across the operating regime). Because Equation (35) holds state by state, Z-score shaping keeps the efficiency $e_Z \ge 1/(1 + \bar{J})$ along any trajectory on which the density and shaping conditions hold at every state and $J \le \bar{J}$; the efficiency approaches one as the standardized reward approaches a Gaussian in the score distance J. The comparison concerns population shaping ratios under the stated conditions, which have not been verified along the LLM trajectories.

Remark B.2 (General shaping extension). Under the integrability conditions below, the update direction of ES with a fixed shaping is an unbiased estimator of the gradient of a smoothed shaped objective, and its covariance has an exact finite-batch form. For a fixed Borel-measurable shaping $\Psi : \mathbb{R} \to \mathbb{R}$, independent of the state and the samples, let $\hat{g}_\Psi(\theta) = \frac{1}{\sigma N} \sum_{n=1}^N \Psi(R_n)\epsilon_n$ with $R_n = R(\theta + \sigma\epsilon_n)$ and $\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_D)$. If $\Psi(R(y))$ times the density of $\mathcal{N}(\theta, \sigma^2 I_D)$ and its first θ-derivatives admit a locally uniform integrable dominator in y, which holds for bounded $\Psi \circ R$, differentiation of the Gaussian density gives, without differentiability of R,

$$\mathbb{E}[\hat{g}_\Psi(\theta)] = \nabla_\theta G_\Psi(\theta), \qquad G_\Psi(\theta) := \mathbb{E}_\epsilon[\Psi(R(\theta + \sigma\epsilon))];$$

for affine $\Psi(r) = ar + b$, this equals $a\nabla R_\sigma(\theta)$ exactly. For $Y = R(\theta + \sigma\epsilon)$, $\mu_\Psi := \mathbb{E}[\Psi(Y)\epsilon]$, and $\mathbb{E}[\Psi(Y)^2\|\epsilon\|^2] < \infty$, independence across samples gives the exact covariance

$$\mathrm{Cov}(\hat{g}_\Psi(\theta)) = \frac{\mathbb{E}[\Psi(Y)^2 \epsilon\epsilon^\top] - \mu_\Psi \mu_\Psi^\top}{\sigma^2 N}.$$

These identities apply to fixed shapings, whereas batch Z-score shaping uses sample-dependent weights. When $B_\Psi^2 < \infty$, the deviation of this covariance from $B_\Psi^2 I_D/(\sigma^2 N)$ depends on the joint distribution of reward and perturbation. The population-level ratio A_Ψ^2/B_Ψ^2 in Proposition 3.2 does not determine this covariance.

C Cross-Hessian Bounds (Assumption 3.3)

Under the Hessian Lipschitz condition of Assumption 4.2 and the radius condition below, the cross terms $P\mathcal{H}(\theta)(I - P)$ remain of order $O(\rho_H)$ relative to $\lambda_{\max}(|H_u(\theta)|)$, as required in Assumption 3.3.

Write $H_{\mathrm{cross}} = P\mathcal{H}(I-P) + (I-P)\mathcal{H}P$ for the off-diagonal blocks of $\mathcal{H}(\theta)$. Since $\mathcal{S}$ is an invariant subspace of $\mathcal{H}(\theta^*)$, $H_{\mathrm{cross}}(\theta^*) = 0$; with $\Delta := \mathcal{H}(\theta) - \mathcal{H}(\theta^*)$, $\|P\|_{\mathrm{op}} = \|I-P\|_{\mathrm{op}} = 1$, and the Hessian Lipschitz condition,

$$\|H_{\mathrm{cross}}(\theta)\|_{\mathrm{op}} = \|P\Delta(I-P) + (I-P)\Delta P\|_{\mathrm{op}} \le 2\|\Delta\|_F \le 2L_3\|\theta-\theta^*\|.$$

Let $\mathcal{U}(\theta^*)$ be the open ball of radius r around θ^*, where $4L_3 r \le \rho_H \lambda_{\max}(|H_u(\theta^*)|)$ and $\rho_H \le 2$. The Lipschitz and radius conditions give $|\lambda_{\max}(|H_u(\theta)|) - \lambda_{\max}(|H_u(\theta^*)|)| \le L_3 r \le \frac{1}{2}\lambda_{\max}(|H_u(\theta^*)|)$ on $\mathcal{U}(\theta^*)$, so

$$\varepsilon_{\mathrm{cross}} := \sup_{\theta\in\mathcal{U}(\theta^*)} \frac{\|H_{\mathrm{cross}}(\theta)\|_{\mathrm{op}}}{\lambda_{\max}(|H_u(\theta)|)} \le \frac{2L_3 r}{\lambda_{\max}(|H_u(\theta^*)|)/2} \le \rho_H.$$

D Proof of Theorem 4.3

We analyze the exact-coefficient SDE in iteration time and the independent normalized gradient flow of Theorem 4.3. All coefficient replacements in the local progress estimate are retained through the residuals in Equation (14).

Local stochastic dynamics. The effective-slice function has derivatives $\nabla\bar{V}_u = -P\nabla F$ and $\nabla^2\bar{V}_u = -P\nabla^2 F P$. Hence the iteration-time generator of Equation (4) is

$$\mathcal{L}_N\bar{V}_u(\theta) = -\alpha\langle\nabla F(u), Pb_N(\theta)\rangle - \frac{\alpha^2}{2}\operatorname{Tr}\big(\nabla^2 F(u)P\Sigma_N(\theta)P\big).$$

Substituting Equation (14) gives

$$\begin{aligned}\mathcal{L}_N\bar{V}_u &= -\frac{\alpha}{\sigma}\|\nabla F\| - \alpha\langle\nabla F, e_N\rangle - \frac{\alpha^2}{2\sigma^2 N}\operatorname{Tr}(\nabla^2 F) - \frac{\alpha^2}{2}\operatorname{Tr}(\nabla^2 F E_N)\\ &\le -\frac{\alpha}{\sigma}\|\nabla F\| + \frac{Ld\alpha^2}{2\sigma^2 N} + r_N.\end{aligned}$$

Here $|\operatorname{Tr}(\nabla^2 F)| \le Ld$ on $\mathcal{B}(u^*)$, and Cauchy–Schwarz bounds the drift residual. The PL condition then establishes Equation (15).

For the stopped expectation, let $\mathcal{G}$ be a relatively compact open full-state region with $P(\overline{\mathcal{G}} - \theta^*) \subset \mathcal{B}(u^*)$ and on which $\nabla F \ne 0$. For deterministic $\theta_0 \in \mathcal{G}$, define $\tau_{\mathcal{G}} := \inf\{s \ge 0 : \theta_s \notin \mathcal{G}\}$ and $V_0 := \bar{V}_u(\theta_0)$. The stopped Itô formula gives, for every finite $s \ge 0$,

$$\begin{aligned}\mathbb{E}\bar{V}_u(\theta_{s\wedge\tau_{\mathcal{G}}}) &= V_0 + \mathbb{E}\int_0^{s\wedge\tau_{\mathcal{G}}} \mathcal{L}_N\bar{V}_u(\theta_q)\,dq\\ &\le V_0 + \mathbb{E}\int_0^{s\wedge\tau_{\mathcal{G}}} \left[-2\beta\sqrt{V_u(u_q)} + \gamma_u + r_N(\theta_q)\right] dq.\end{aligned}$$

The derivatives of F and the coefficients are bounded on the stopped compact region by Gaussian smoothing and the coefficient bounds in Appendix A. In particular,

$$\mathbb{E}\int_0^{s\wedge\tau_{\mathcal{G}}} \alpha^2\nabla F(u_q)^\top P\Sigma_N(\theta_q)P\nabla F(u_q)\,dq < \infty,$$

so the stopped stochastic integral is square integrable and has mean zero.

Normalized gradient flow. For the independent normalized gradient flow Equation (16), the chain rule gives

$$\frac{d}{ds}V_u(u_s^{\mathrm{ref}}) = -\frac{\alpha}{\sigma}\|\nabla F(u_s^{\mathrm{ref}})\| \le -2\beta\sqrt{V_u(u_s^{\mathrm{ref}})}.$$

Since $V_u \ge 0$ on the open region $\mathcal{B}(u^*)$, a zero of V_u there has zero gradient. Thus while the stated local conditions and nonzero gradient persist, $V_u > 0$ and $\frac{d}{ds}\sqrt{V_u(u_s^{\mathrm{ref}})} \le -\beta$. Integrating over $0 \le s < T_A$ proves Equation (17). For the stochastic process, the stopped drift inequality above retains both the noise baseline γ_u and the coefficient residual r_N, so it holds without a smallness or N-scaling assumption on r_N.

E Proof of Proposition 4.4

We prove the following sharper bound, which implies Equation (18). Assume the hypotheses of Proposition 4.4 except $0 \le \varepsilon_\Sigma, q_B \le 1$, write $Q = I - P$ and $m = D - d$, and let $S_c = \bar{c}\,\alpha^2 mh/(\sigma^2 N)$. If $\bar{c} > 0$, then

$$\left|\mathbb{E}\|v_b - v_a\|^2 - S_c\right| \le S_c \left\{\varepsilon_\Sigma + q_B + 2\sqrt{(1+\varepsilon_\Sigma)q_B}\right\}. \tag{36}$$

If $\bar{c} = 0$, the displacement is zero almost surely.

On the deterministic interval $[a, b]$, integrate the neutral component of Equation (10) to obtain

$$\Delta := v_b - v_a = B + M, \qquad B = \alpha \int_a^b Q b_N(Y_s)\, ds, \qquad M = \alpha \int_a^b Q A_N(Y_s)\, d\widetilde{W}_s.$$

By Equation (32), $\|b_N\|^2 \le (N-1)c_N/(\sigma^2 N) \le (N-1)/(\sigma^2 N)$. Also, Equation (21) gives $\operatorname{Tr}\Sigma_N \le \mathbb{E}\|\hat{g}\|^2 \le D/\sigma^2$. These bounds ensure square integrability of B, M on every finite interval where the solution exists and justify the expectation and time-integral exchanges below. Itô's isometry gives

$$\mathbb{E}\|M\|^2 = \alpha^2 \mathbb{E} \int_a^b \operatorname{Tr}(Q\Sigma_N(Y_s)Q)\, ds. \tag{37}$$

Fubini's theorem gives $\mathbb{E}B = \alpha \int_a^b \mathbb{E}[Q b_N(Y_s)]\, ds$; since the drift integrand depends on Y_s, $\mathbb{E}\|B\|^2$ generally differs from $\|\mathbb{E}B\|^2$. Although $\mathbb{E}M = 0$, B and M need not be orthogonal in expectation, and

$$\mathbb{E}\|\Delta\|^2 = \mathbb{E}\|M\|^2 + \mathbb{E}\|B\|^2 + 2\mathbb{E}\langle B, M\rangle. \tag{38}$$

Apply Proposition A.5 with $\Pi = Q$ at each state. Together with Equation (37), the assumed trace-error bound gives

$$\left|\mathbb{E}\|M\|^2 - S_c\right| \le \varepsilon_\Sigma S_c.$$

Separately, Cauchy–Schwarz for the time integral and Equation (32) imply

$$\mathbb{E}\|B\|^2 \le \alpha^2 h\, \mathbb{E} \int_a^b \|b_N(Y_s)\|^2\, ds \le \frac{\alpha^2 (N-1)h}{\sigma^2 N} \mathbb{E} \int_a^b c_N(Y_s)\, ds = q_B S_c.$$

Consequently,

$$|\mathbb{E}\langle B, M\rangle| \le \sqrt{\mathbb{E}\|B\|^2\, \mathbb{E}\|M\|^2} \le S_c \sqrt{(1+\varepsilon_\Sigma)q_B}.$$

Substitution into Equation (38) proves Equation (36).

If $\bar{c} = 0$, nonnegativity implies $c_N(Y_s) = 0$ for almost every (s, ω). Then $b_N(Y_s) = 0$ and $\Sigma_N(Y_s) = 0$ almost everywhere by Proposition A.5, so $B = M = \Delta = 0$ almost surely.

Relative form. Since $S_c = \bar{c}S$ for $S = \alpha^2 mh/(\sigma^2 N)$, comparison with the nominal linear reference gives

$$\left|\frac{\mathbb{E}\|v_b - v_a\|^2}{S} - 1\right| \le (1 - \bar{c}) + \bar{c}\left\{\varepsilon_\Sigma + q_B + 2\sqrt{(1+\varepsilon_\Sigma)q_B}\right\}.$$

For $0 \le \varepsilon_\Sigma, q_B \le 1$, the bounds $\bar{c} \le 1$, $\sqrt{1+\varepsilon_\Sigma} \le \sqrt{2}$, and $q_B \le \sqrt{q_B}$ make the right-hand side at most $(1-\bar{c}) + \varepsilon_\Sigma + (1+2\sqrt{2})\sqrt{q_B} \le 4(1 - \bar{c} + \varepsilon_\Sigma + \sqrt{q_B})$, which proves Equation (18); if $\bar{c} = 0$, the left-hand side equals one. The estimate uses the actual, possibly random Y_a at the deterministic start time; $v^* = 0$ does not imply $v_a = 0$. The covariance trace is $\operatorname{Tr}\operatorname{Cov}(\Delta) = \mathbb{E}\|\Delta\|^2 - \|\mathbb{E}\Delta\|^2$, where $\mathbb{E}\Delta = \mathbb{E}B$. The same argument with $Q = I_D$ and $m = D$ gives the full-parameter displacement, while the neutral result concerns $Q(Y_b - Y_a)$. □

Scope. The proof treats the exact-coefficient SDE with state-dependent coefficients throughout the window. Applying the bound to a discrete-ES window additionally requires the discrete-ES and SDE window MSDs to differ by a small fraction of S, which the single-time absolute $O(\alpha)$ bound of Theorem 3.1 does not by itself ensure.

F Isotropic Radial Model and Invariant Radius Statistics

We separately specify an inward-drift model on the fixed neutral subspace $\mathcal{S}^{\perp}$ to examine a possible saturation scale. We use iteration time $s = t/\alpha$ and measure the radius from the fixed coordinate origin corresponding to θ^*.

Theorem F.1 (Radial equation for the isotropic reference model)**.** *Let $q = D - d \geq 2$, $a = \alpha/\sigma > 0$, $\eta = \alpha/(\sigma\sqrt{N}) > 0$, and $r_0 = \|v_0\| > 0$. The independently specified vector SDE on $\mathcal{S}^{\perp} \simeq \mathbb{R}^q$,*

$$dv_s = -a\frac{v_s}{\|v_s\|}\,ds + \eta\,dB_s, \tag{39}$$

where B_s is standard q-dimensional Brownian motion, has a unique nonexplosive strong solution from v_0. Its radius $r_s = \|v_s\|$ satisfies

$$dr_s = \left(-a + \frac{(q-1)\eta^2}{2r_s}\right)ds + \eta\,d\beta_s, \tag{40}$$

where β_s is a one-dimensional standard Brownian motion. The origin is an unattainable entrance boundary and infinity is a natural boundary for the radial diffusion.

Radial equation and boundaries. The vector coefficients are locally Lipschitz away from the origin. On each annulus, Itô's formula applies with $\hat{v} = v/\|v\|$, $\nabla r = \hat{v}$, and $\nabla^2 r = (I - \hat{v}\hat{v}^{\top})/r$. The Hessian trace is $(q-1)/r$, and $\beta_s = \int_0^s \hat{v}_\ell^{\top}\,dB_\ell$ has quadratic variation s, giving Equation (40) up to the exit time. For $X_s = r_s/\eta$, the generator is

$$\mathcal{L}f(x) = \tfrac{1}{2}f''(x) + \left(\frac{\nu + 1/2}{x} - \frac{a}{\eta}\right)f'(x), \qquad \nu = \frac{q-2}{2} \geq 0.$$

This is the Bessel process with constant drift of Linetsky [2004, p. 328, Section 1, Eqs. (1)–(3)], whose origin is entrance and whose infinite boundary is natural. Consequently the process starting at $r_0 > 0$ never reaches zero. The bounded vector drift and constant diffusion rule out explosion, extending the local strong solution to all finite times.

Proposition F.2 (Invariant radial law and its moments)**.** *The radial diffusion in Theorem F.1 has invariant density*

$$\pi_r(r) = \frac{\kappa^q}{\Gamma(q)}r^{q-1}e^{-\kappa r}, \qquad r > 0, \tag{41}$$

where $\kappa = 2a/\eta^2 = 2\sigma N/\alpha$ is the rate of the Gamma distribution with shape q. Its mean and second moment are

$$\mathbb{E}_\pi r = \frac{q}{\kappa}, \qquad \mathbb{E}_\pi r^2 = \frac{q(q+1)}{\kappa^2} = \frac{\alpha^2 q(q+1)}{4\sigma^2 N^2}. \tag{42}$$

F.1 Invariant-Law Calculation

Write $b(r) = -a + (q-1)\eta^2/(2r)$ for the drift in Equation (40). The stationary Fokker–Planck equation with zero probability current is

$$b(r)\pi_r(r) - \frac{\eta^2}{2}\pi_r'(r) = 0.$$

Thus $(\log \pi_r)' = -\kappa + (q-1)/r$, and normalization follows from

$$\int_0^\infty r^{q-1}e^{-\kappa r}\,dr = \Gamma(q)\kappa^{-q}.$$

Together with the entrance and natural boundaries, this normalized speed density gives the invariant probability measure. Equivalently, changing variables $r = \eta X$ in Linetsky [2004, Eq. (3)] gives Equation (41). Integrating r and r^2 against this density yields Equation (42).

F.2 Scale Matching and Displacement

Matching the drift-free scale $\eta^2 qh$ over a duration h to the invariant second moment gives

$$h_{\text{match}} := \frac{\mathbb{E}_\pi r^2}{\eta^2 q} = \frac{q+1}{4N} \simeq \frac{q}{4N} \qquad (q \gg 1). \tag{43}$$

Thus the drift-free neutral MSD reaches the invariant second moment once the total number of perturbations Nh reaches $(q+1)/4$, about a quarter of the neutral dimension; this compares scales and is not a saturation arrival time.

G Evolutionary Analogies

Table 4 details the correspondences with positive selection and neutral theory noted in Section 4 and where they break down.

Table 4: Correspondence between the Adaptive and Neutral descriptions and evolutionary theory.

Concept	ES counterpart	Result	Limit of the analogy
Darwinian positive selection	Improvement along the normalized effective-subspace gradient	Improvement coefficient α/σ independent of d and D (Theorem 4.3)	Selection acts through Z-score weighting of perturbations; individuals are not inherited across iterations
Neutral theory [Kimura, 1968]	Displacement on $\mathcal{S}^\perp$ whose nominal rate does not depend on the reward	Neutral MSD grows approximately linearly at $\alpha^2(D-d)/(\sigma^2 N)$ per iteration on short windows (Proposition 4.4)	Neutral alleles are substituted at approximately the rate per individual at which they arise [Kimura, 1968, p. 625], independently of population size, whereas the ES neutral MSD rate decreases as $1/N$

H Task Projections, Stopping, and Budget Allocation (Section 4.3)

This appendix supplies the analysis deferred from Section 4.3.

H.1 General Local Expansion

For an evaluation function J (possibly distinct from the training reward R), Taylor expansion relates the expected reward change to the full displacement $\Delta_t = \theta_t - x$ from a deterministic reference point x:

$$\begin{aligned} \mathbb{E}[J(x) - J(\theta_t)] &= -\nabla J(x)^\top m_t - \tfrac{1}{2}\mathrm{Tr}\big(H_J(x) C_t\big) - \tfrac{1}{2} m_t^\top H_J(x) m_t + r_3(t), \\ |r_3(t)| &\le \tfrac{L_J}{6}\mathbb{E}\|\Delta_t\|^3, \end{aligned} \tag{44}$$

where $m_t = \mathbb{E}\Delta_t$, $C_t = \mathrm{Cov}(\Delta_t)$, $H_J = \nabla^2 J$, and $\mathbb{E}\|\Delta_t\|^3 < \infty$. Here $J \in C^3(O)$ on a deterministic open set O containing $[x, \theta_t]$ almost surely, with $\sup_{y \in O} \|\nabla^3 J(y)\|_{\text{op}} \le L_J < \infty$ in the Euclidean trilinear operator norm. Indeed, the pointwise Taylor remainder is bounded by $L_J\|\Delta_t\|^3/6$; taking expectations and using $\mathbb{E}[\Delta_t \Delta_t^\top] = C_t + m_t m_t^\top$ gives Equation (44). All terms are integrable by the finite third moment, and the identity holds also for indefinite $H_J(x)$.

For the fixed projector P, C_t includes both projected covariance blocks and their cross-covariances. On a window starting at t_0, $(I-P)\Delta_t = (v_t - v_{t_0}) + (I-P)(\theta_{t_0} - x)$, with x still deterministic. The second moment of $(I-P)\Delta_t$ includes the expected squared initial offset and twice the expected inner product of these two terms, in addition to $\mathbb{E}\|v_t - v_{t_0}\|^2$. Proposition 4.4 concerns the increment on a deterministic interval $[a, b]$ in iteration time, corresponding here to $t_0 = a$ and $t = b$, rather than the full displacement from x.

H.2 Training-Gap Decomposition

To relate the effective-slice gap to the full training gap, let $V(\theta_s) := R_\sigma(\theta^*) - R_\sigma(\theta_s)$. In the centered coordinates, $\theta_s - \theta^* = u_s + v_s$ and $u^* = 0$. Taylor expansion around the local maximizer θ^*, with all Hessian blocks below evaluated at θ^*, gives

$$V(\theta_s) = V_u(u_s) + V_v(v_s) + \Delta_{\text{cross}}(\theta_s) + O(L_3\|\theta_s - \theta^*\|^3), \tag{45}$$

where

$$V_u(u_s) = F(u^*) - F(u_s) = -\tfrac{1}{2}u_s^\top H_u u_s + O(L_3\|u_s\|^3), \tag{46}$$

$$V_v(v_s) := -\tfrac{1}{2}\, v_s^\top H_v\, v_s, \tag{47}$$

$$\Delta_{\text{cross}}(\theta_s) := -u_s^\top H_{\text{cross}}\, v_s = 0.$$

The cross block at θ^* vanishes because P is a spectral projector of $\mathcal{H}(\theta^*)$. On a neighborhood where the Taylor segments lie in $\mathcal{U}(\theta^*)$, the remainder in Equation (45) is bounded by $L_3(\|\theta_s - \theta^*\|^3 + \|u_s\|^3)/6$. Thus a bound on V_u alone does not control the full training gap; the neutral quadratic contribution and Taylor remainder must also be assessed.

H.3 Proof of Proposition 4.5

(i) Neutral quadratic contribution. The finite second moment gives $\mathbb{E}[v_s v_s^\top] = \mathrm{Cov}(v_s) + (\mathbb{E}v_s)(\mathbb{E}v_s)^\top$. Substituting this identity into Equation (47) gives the equality in Equation (19). Since θ^* is a local maximizer, $-H_v(\theta^*)$ is positive semidefinite, so $V_v(v_s) \geq 0$. The curvature ratio in Assumption 3.3 gives $\|H_v(\theta^*)\|_{\text{op}} \leq \rho_H \lambda_u$ and hence $V_v(v_s) \leq \frac{1}{2}\rho_H\lambda_u\|v_s\|^2$ pointwise. Taking expectations proves the upper bound.

(ii) Parameter displacement. Orthogonality of the fixed subspaces gives

$$\mathbb{E}\|\theta_T - \theta_0\|^2 = \mathbb{E}\|u_T - u_0\|^2 + \mathbb{E}\|v_T - v_0\|^2.$$

The neutral second moment is $\mathbb{E}\|v_T - v_0\|^2 = \mathrm{Tr}(\mathrm{Cov}(v_T - v_0)) + \|\mathbb{E}(v_T - v_0)\|^2$. Substituting the two moment identities in Proposition 4.5 (ii) and $\eta^2 = \alpha^2/(\sigma^2 N)$, then taking the square root, yields Equation (20). The reward change on another task is evaluated separately by Equation (44), retaining its mean and remainder terms. □

H.4 Training Task: Rise-then-Decay

Liang et al. [2026] identify rise-then-decay of the training reward as a structural consequence of anisotropic curvature interacting with fixed stochasticity (their Fig. 1, §4). Their analysis uses a discrete local AR(1)/OU model applied to unshaped ES (their Algorithm 1 and Eq. 7), in which the update rule carries a raw reward r_k without Z-score normalization.

For Z-score-normalized ES, we use Equation (44) to show how improvement and curvature-weighted displacement can contribute to the rise and the decay.

Apply Equation (44) with $J = R_\sigma$ and $x = \theta^*$ (so $H_R = \mathcal{H}$). At this local maximizer, $\nabla R_\sigma(\theta^*) = 0$ and $-H_R(\theta^*) \succeq 0$, so both quadratic terms contribute nonnegative loss, with the remainder retained. Comparing expected rewards at two times requires subtracting all terms in Equation (44) at this same reference point.

In this expansion, the effective-slice improvement of Theorem 4.3 can contribute to the rise, and the neutral covariance term, bounded with a ρ_H-suppressed curvature as in Proposition 4.5 (i), can contribute to the decay.

Shaping-dependent modifications. In the unshaped analysis of Liang et al. [2026], where $\lambda_i > 0$ denotes an eigenvalue of the negative reward Hessian, each mode's contribution to the expected reward relaxes exponentially at rate $\gamma_i = -2\log|1 - \alpha\lambda_i|$ (their Eq. 21). Three quantities differ between the unshaped and Z-score settings:

- Rise timescale: γ_i^{-1} per mode for the exponential rise (unshaped) vs $\sqrt{V_0}/\beta$ for the deterministic normalized-gradient bound (Z-score, Equation (17)).

- Radial scale comparison: $\gamma_i^{-1} \approx 1/(2\alpha\lambda_i)$ per mode for small $\alpha\lambda_i$ (unshaped) vs $h_{\text{match}} = (D - d + 1)/(4N)$ in iteration time for the isotropic radial model of Equation (40) (see Equation (43)).
- Plateau coefficient: Liang et al. [2026], Eq. 19 vs the invariant radius moments of Equation (40) in Appendix F.1, whose projection onto R_σ depends on H_v.

The balance of improvement and curvature-weighted displacement provides a mechanism for rise-then-decay under Z-score shaping, while its occurrence and quantitative profile depend on these contributions and the remainder.

H.5 Evaluation Task: Catastrophic Forgetting

Let R' evaluate a pre-training capability distinct from R (e.g., HellaSwag in Abdi et al. 2026, Fig. 2). Apply Equation (44) with $J = R'_\sigma := \mathbb{E}_\epsilon[R'(\theta + \sigma\epsilon)]$ and $x = \theta_0$, under its smoothness and moment conditions. Because θ_0 is generally not a critical point of R'_σ, the gradient term need not vanish; the mean quadratic term and expected remainder also contribute.

The Hessian $H_{R'} = \nabla^2 R'_\sigma$ is determined by R'_σ, and its restriction to $\mathcal{S}^\perp$ is in general unrelated to the spectral structure of R_σ. In the local quadratic approximation, fluctuations in directions where the pre-training reward curves downward can contribute to performance loss. This mechanism is consistent with the forgetting observed by Abdi et al. [2026] (their Fig. 2–3).

The two phenomena can thus involve the same neutral displacement with different task-dependent curvature weights and mean contributions. The expansion concerns the smoothed evaluation R'_σ; its quantitative application to the reported discrete task scores requires further justification.

H.6 Stopping and Budget Allocation (Remark 4.6)

While its local conditions hold, Equation (17) bounds the remaining effective-slice gap of the normalized gradient flow by $(\sqrt{V_0} - \beta s)^2$, which shrinks to zero by T_A, whereas the nominal neutral MSD $\eta^2(D - d)T$ grows linearly in the number of iterations T. Once the effective-slice gap is small, neutral displacement continues to accumulate and can affect the training and evaluation rewards through the task-dependent curvature, mean, and remainder terms of Equation (44).

In practice, when the training reward levels off, this suggests weighing the benefit of further training against its effects on the training and evaluation tasks before deciding whether to stop. Under a fixed budget $C = N \cdot T$, it also suggests choosing the largest N for which $T = C/N$ still covers the effective-coordinate improvement, because the nominal neutral MSD $\alpha^2(D-d)C/(\sigma^2 N^2)$ decreases with N. Because varying N also changes the exact coefficients b_N and Σ_N, this nominal comparison guides the allocation rather than determining an optimal one.

In the fixed-budget Countdown runs on Qwen2.5-0.5B (Table 9), the measured displacement stays within 11% of the full-parameter comparison scale $\eta^2 DT$, the HellaSwag drop decreases from 16.2 to 2.1 percentage points as N increases from 15 to 60, and the terminal reward is highest at $N = 30$, consistent with the trade-off in Remark 4.6.

I Synthetic Experiments

This appendix compares the closed-form coefficient models with ES on toy landscapes and reports the computational resources of the synthetic experiments. Unless stated otherwise, the reward is the diagonal quadratic of Section 5, $R(\theta) = -\frac{1}{2}\theta^\top \Lambda \theta$, where Λ is diagonal with top-d entries λ_u and bulk entries $\lambda_v = \rho_H \lambda_u$, so $\mathcal{H} = -\Lambda$; the other landscapes are specified in Appendix I.3.

I.1 Validity of the Finite Quadratic Approximation

For the polynomial toy rewards below, the finite quadratic decomposition follows from Gaussian moment calculations; these unbounded rewards are not covered by Proposition A.4. Figure 4 compares the terms of Equation (6) on the diagonal quadratic for $\sigma \in \{5\times10^{-4}, 7.5\times10^{-4}, 10^{-3}, 1.5\times10^{-3}, 2.5\times10^{-3}\}$ (10 trials, 1500 iterations). Let Q_t be the trial mean of the right-hand side of

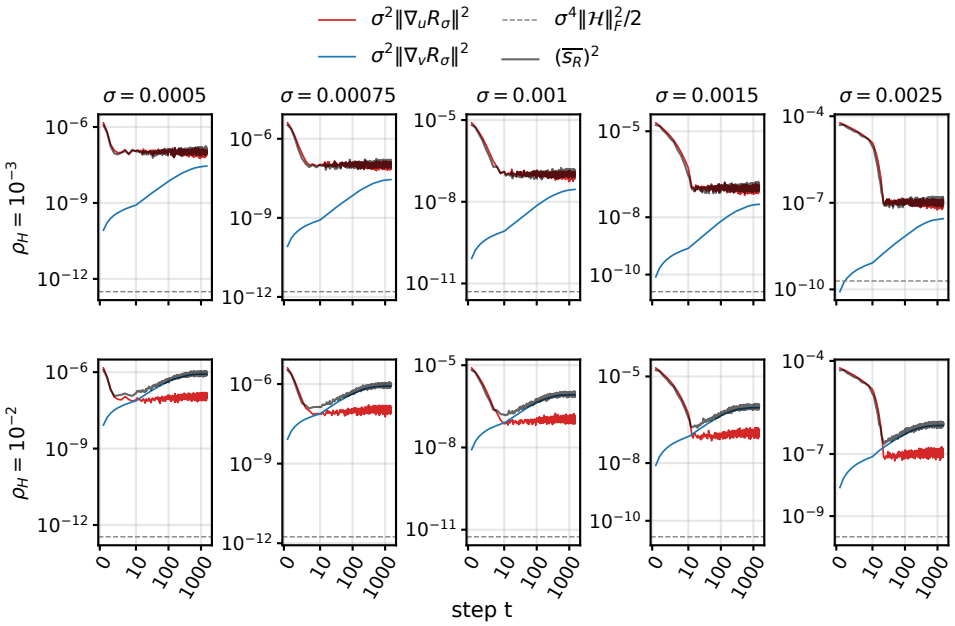


Figure 4: **Gradient and Hessian contributions on the diagonal quadratic.** Rows: $\rho_H \in \{10^{-3}, 10^{-2}\}$; columns: σ. Red and blue: effective- and neutral-gradient contributions; grey: $\sigma^4\|\mathcal{H}\|_F^2/2$; black: squared trial-mean sample standard deviation. The Hessian contribution stays below 0.35% of the larger gradient contribution.

Equation (6) and $O_t = \overline{s_R(t)^2}$ the trial mean of the batch reward variance (divisor N); we report the time average of $|O_t - Q_t|/O_t$, which includes finite-batch estimation effects.

Smooth quartic perturbation. On $R(\theta) = -\frac{1}{2}\theta^\top \Lambda \theta - (c/4)\|\theta\|^4$ with $D = 10^4$, $d = 10$, $\rho_H = 10^{-3}$, $\sigma = 10^{-3}$, $N = 30$, and $c \in \{0, 0.01, 0.1\}$, this relative residual averages approximately 7.2–8.1% over iterations 30–1499 and 10 trials, while the ratio of the time-averaged third-degree variance contribution to the time average of Q_t is below 10^{-13}.

Finite-population contribution. On the diagonal quadratic with $\alpha = 5 \times 10^{-4}$, $\sigma = 10^{-3}$, $D = 10^4$, $d = 10$, and $\rho_H \in \{10^{-3}, 10^{-2}\}$ (five trials, iterations 30–499), the time-averaged relative residual decreases with the population size, consistent with finite-population sampling:

$\mathrm{mean}_t\, \lvert O_t - Q_t\rvert / O_t$	$N = 30$	$N = 100$	$N = 300$	$N = 1000$
$\rho_H = 10^{-3}$	10.9%	5.4%	3.0%	1.6%
$\rho_H = 10^{-2}$	11.3%	5.3%	3.0%	1.6%

Bimodal Gaussian with narrow peaks. The bimodal reward is the sum of two Gaussian functions of the first coordinate in $D = 10^4$ dimensions. With peak widths 10^{-4} and 6×10^{-4}, $\alpha = 2 \times 10^{-7}$, initial first coordinate 0.0025, $N = 100$, five trials, and iterations 30–499, the relative residual averages approximately 66–94% for $\sigma \in \{8 \times 10^{-4}, 10^{-3}, 1.4 \times 10^{-3}, 1.8 \times 10^{-3}\}$.

I.2 First- and Second-Order σ_R Models and Radial Scales

We integrate the coefficient model Equation (11) by the Euler–Maruyama method with $\sigma_R^{(1)} = \sigma\|\nabla R_\sigma\|$ or $\sigma_R^{(2)} = (\sigma^2\|\nabla R_\sigma\|^2 + \frac{1}{2}\sigma^4\|\mathcal{H}\|_F^2)^{1/2}$ on the diagonal quadratic ($D = 10^4$, $d = 10$, $N = 30$, $\alpha = 5 \times 10^{-4}$, unit iteration step, 100 trials per model, 1500 iterations) and compare the two models with ES (10 trials) for $\rho_H \in \{10^{-3}, 10^{-2}\}$ and $\sigma \in \{5 \times 10^{-4}, 10^{-3}\}$ (Figure 5). The time average over iterations 1–1500 of the absolute difference between model and ES mean squared

radii, relative to ES, ranges from approximately 0.7% to 4.1%. The two models give visually close curves in all six conditions, including $\sigma = 2\times10^{-3}$, for which no ES runs exist.

At $\rho_H = 10^{-3}$ and $\sigma = 10^{-3}$, the ES mean squared radius at $s = 1500$ is approximately 28421, and both models give approximately 28137, a difference of approximately 1.0%. These finite-time values differ from the invariant second moment $\mathbb{E}_\pi r^2 = \alpha^2(D-d)(D-d+1)/(4\sigma^2N^2) \approx 6931$ of the isotropic radial model (Equation (42)) and from the equilibrium estimate of approximately 13871 obtained by substituting mean squared coordinates into the two-term coefficient model. The ES runs and both coefficient models start at $v_0 = 0$ and retain the effective-subspace component, whereas Theorem F.1 analyzes $r_0 > 0$.

I.3 Landscape Dependence: Smooth, Multimodal, Discontinuous, and Oscillatory

To probe the first-order σ_R approximation beyond the diagonal-quadratic case, we compare ES with three coefficient models using $\sigma_R^{(1)}$, $\sigma_R^{(2)}$, or a 300-sample Monte-Carlo estimate of σ_R. The four landscapes are: (i) the diagonal quadratic; (ii) a bimodal Gaussian-mixture reward in the effective coordinate; (iii) a radial indicator $\mathbf{1}\{\|u\| < r_{\text{target}}\}$; and (iv) the diagonal quadratic plus $0.05\sum_{i=1}^{d}\sin(\omega_i\theta_i)$, with fixed frequencies drawn from $[20, 100]$. The bimodal and indicator rewards also include the weak neutral quadratic term $-10^{-3}\|v\|^2/2$. Each model uses four trials of 600 iterations at $D = 5000$ and $N = 30$ (Figure 6); these are coefficient approximations, not the exact finite-batch SDE. The first column of the figure uses the zero reference of each landscape, which is not the distance from an optimum for the bimodal and oscillatory rewards; the reward-spread column shows the batch standard deviation for ES, the analytical approximations for the first- and second-order models, and the 300-sample estimate for the Monte-Carlo model.

Smooth single-basin (quadratic) and oscillatory (jagged). For $d = 10$, $\rho_H = 2.5\times10^{-3}$, $\alpha = 10^{-3}$, $\sigma = 10^{-2}$, and $u_{0,i} = 6$, the largest pointwise relative difference of the trial-mean $\langle\|v_k\|^2\rangle$ from ES over $k = 0, \ldots, 600$ is approximately 6.6% on the quadratic landscape (first-order model, $k = 575$) and 4.2% on the oscillatory landscape (second-order model, $k = 28$) across the three models. The shading in these two neutral-coordinate panels is the pointwise mean ±1 standard error over the four trials. Gaussian smoothing multiplies each sine component by $\exp(-\omega_i^2\sigma^2/2)$,

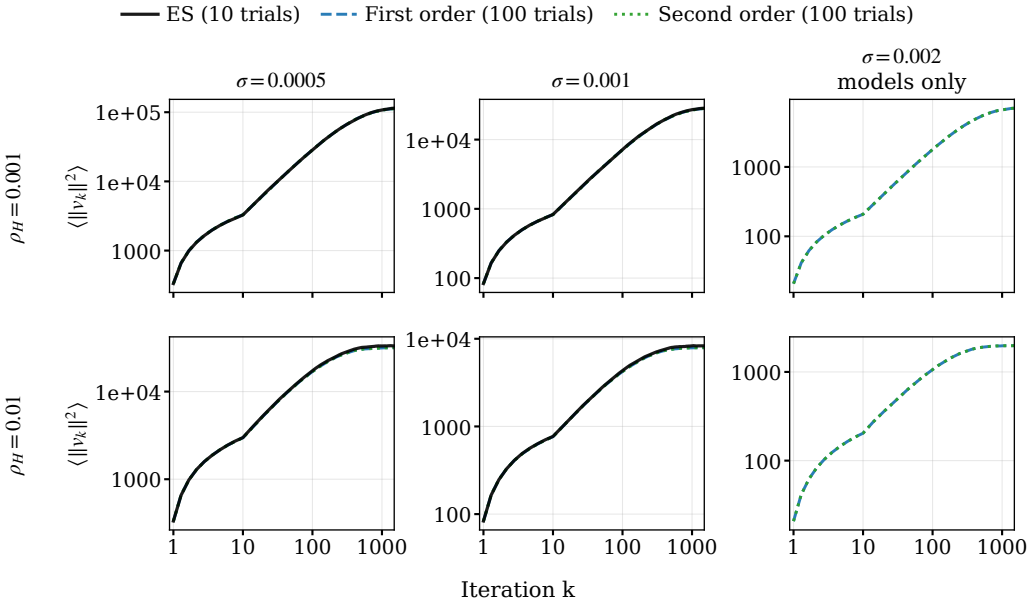


Figure 5: **Mean squared neutral radius of ES and the coefficient models.** Solid black: ES (10 trials; first two columns only); dashed and dotted: models with $\sigma_R^{(1)}$ and $\sigma_R^{(2)}$ (100 trials each). Rows: $\rho_H = 10^{-3}, 10^{-2}$; columns: $\sigma = 5\times10^{-4}, 10^{-3}, 2\times10^{-3}$.

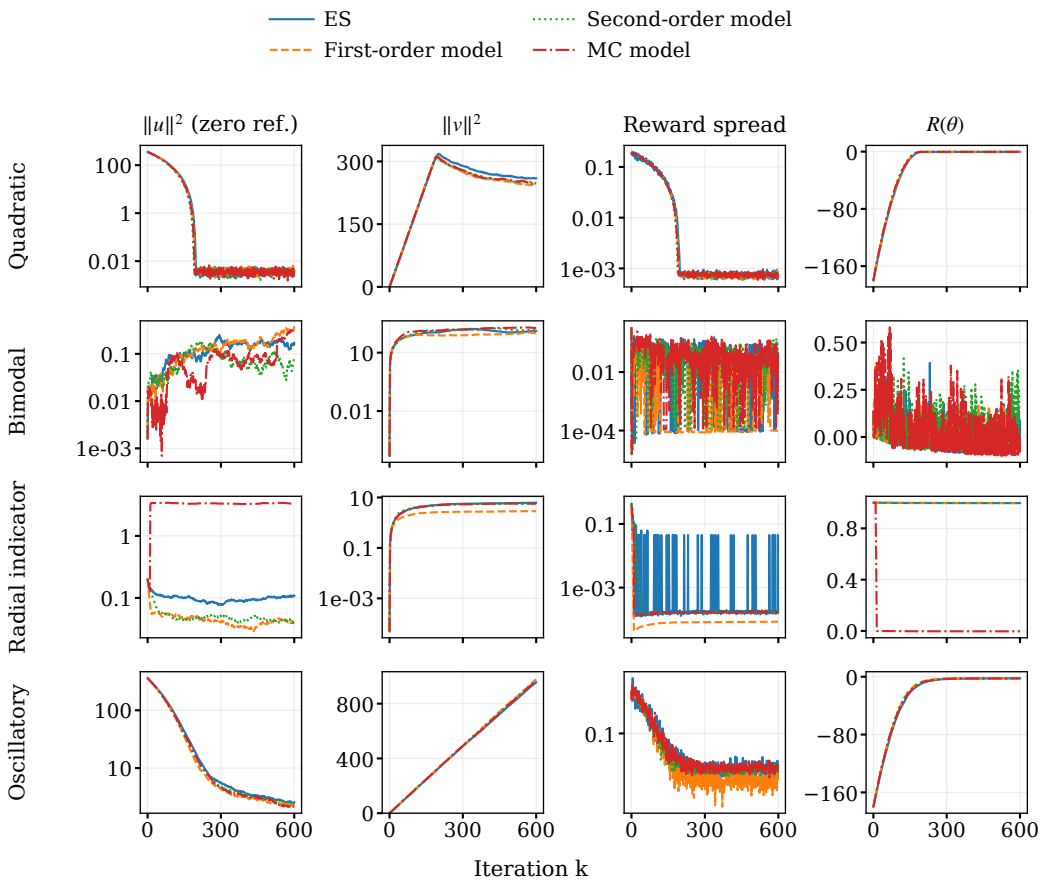


Figure 6: **Four-landscape comparison of ES and three coefficient models.** Rows: quadratic, bimodal, radial indicator, and oscillatory rewards; columns: effective- and neutral-coordinate squares, reward spread, and raw reward. Lines average four trials over 600 iterations.

approximately 0.755 at $\omega_i = 75$, so the smoothed oscillatory landscape retains most of its oscillation amplitude.

Multimodal (bimodal). The initial effective coordinate is $u_0 = 0.04$ with $\sigma = 0.01$, and the plotted u^2 is measured from zero rather than from either peak. At the final iteration, the first-order model has a larger four-trial mean of u^2 than ES, whereas the second-order model has a smaller mean.

Discontinuous (radial indicator). Here $d = 5$, $u_{0,i} = 0.2$, $r_{\text{target}} = 0.5$, and $\sigma = 0.05$; smoothed radial derivatives are estimated by finite differences, and the squared effective Hessian norm is approximated by d times the squared radial second derivative. The model with a Monte-Carlo denominator leaves the indicator support early in all four trials.

I.4 Computational Resources

The synthetic experiments ran on the CPU of an Apple M1 Pro MacBook Pro. For the eight experiments in Table 5, 46 of the 47 timed conditions completed within five minutes and the largest took about 30 minutes; the intervals sum to about 69 minutes. Each wall-clock interval includes all trials within one condition and some aggregation, but excludes program startup, imports, final file saving, and subsequent plotting. The model comparison in Figure 5 reuses the ES trajectories of Figure 4, so their computation is counted only in the Figure 4 row.

Table 5: Approximate synthetic experiment times. Times are summed over conditions and rounded to whole minutes; values below one minute are marked < 1, and the total is computed before rounding. The quartic row includes only the three conditions with $\sigma = 10^{-3}$ and $c \in \{0, 0.01, 0.1\}$, and the Figure 5 row counts model conditions and model trajectories.

Reported result	Conditions	Trials/condition	Approx. minutes
Figure 2a,b	3	20	< 1
Figure 2c	4	15	7
Section 5.2	3	5	32
Figure 4	10	10	8
Appendix I.1, quartic	3	10	3
Appendix I.1, bimodal	4	5	2
Appendix I.1, finite N	8	5	13
Figure 5	12	100	3
Total	47		69

The separate two-condition rise-then-decay simulation (Figure 2d), with 10 trials of 2000 updates per condition, takes an estimated five minutes. The four-landscape comparison used the same CPU class; its per-condition settings are given in Appendix I.3.

J LLM Experiments

This appendix first details the three Qwen2.5-Instruct trajectories with Gram-matrix measurements (Figure 3a, Appendix J.2); Appendix J.3 reports the model, task, and budget comparisons summarized in Tables 1 and 9. These three trajectories fine-tune the models on the Countdown task [Gandhi et al., 2024, Pan et al., 2025] using standard ES with non-antithetic perturbations, following Qiu et al. [2026].

J.1 Experimental Setup

The following setup applies to these three trajectories unless stated otherwise; Appendix J.3 describes the other runs.

Models and datasets. We use three Qwen2.5-Instruct models [Qwen Team, 2025] spanning a factor of approximately 15.4 in parameter count: Qwen2.5-0.5B-Instruct ($D = 494{,}032{,}768$), Qwen2.5-3B-Instruct ($D = 3{,}085{,}938{,}688$), and Qwen2.5-7B-Instruct ($D = 7{,}615{,}616{,}512$). All parameters are included in the ES update (full fine-tuning; no low-rank adaptation or other low-rank parametrization is imposed at optimization time), so Assumption 3.3 remains an assumption about the intrinsic spectrum of the landscape rather than a property imposed by construction. Table 6 lists the models and datasets used in the LLM experiments, including the held-out evaluation data.

Task and reward. In the Countdown arithmetic task [Gandhi et al., 2024, Pan et al., 2025], the model must compose the given integers into an arithmetic expression equal to a target value. In all LLM runs, the training reward is the mean score of the perturbed candidates over 200 fixed prompts before each update, with deterministic greedy decoding (temperature 0, seed 42); the score adds answer correctness to 0.1 times a format score, so the reward is not an accuracy.

ES hyperparameters and comparison scale. The three trajectories follow the ES configuration of Qiu et al. [2026], which corresponds to $\alpha = 5\times10^{-7}$, $\sigma = 10^{-3}$, and $N = 30$ in Equation (2), with $T = 500$ iterations. Perturbations are regenerated on the fly from random seeds as in Qiu et al. [2026], which allows reconstruction of the intended updates without materializing billion-dimensional tensors. The full-parameter comparison scale per iteration is $\eta^2 D = \alpha^2 D/(\sigma^2 N)$, and lags and horizons count iterations. For the Qwen2.5-1.5B-Instruct setup of Abdi et al. [2026] ($D \approx 1.54\times10^9$, $T = 500$, and the same α, σ, and N), the nominal scale is $\sqrt{\eta^2 DT} \approx 80.20$, compared with a Frobenius norm of approximately 88 read from their Fig. 3 (Section 6).

Table 6: **Models and datasets used in the LLM experiments.** Names link to the distribution repositories. Licenses refer to the identified model versions and dataset distribution cards; the Countdown card does not specify a dataset license. The first four datasets provide training subsets; HellaSwag provides held-out evaluation data.

Model or dataset	Reference	License
Models		
Qwen2.5-0.5B-Instruct	[Qwen Team, 2025]	Apache 2.0
Qwen2.5-3B-Instruct	[Qwen Team, 2025]	Qwen Research License
Qwen2.5-7B-Instruct	[Qwen Team, 2025]	Apache 2.0
OLMo-2-0425-1B-Instruct	[Allen Institute for AI, 2025]	Apache 2.0
SmolLM2-1.7B-Instruct	[HuggingFaceTB, 2025]	Apache 2.0
Llama-3.2-3B-Instruct	[Meta, 2024]	Llama 3.2 Community License
Phi-4-mini-instruct	[Microsoft, 2025]	MIT
Datasets		
Countdown-Tasks-3to4	[Pan et al., 2025]	Not specified
GSM8K	[Cobbe et al., 2021]	MIT
SVAMP	[Patel et al., 2021]	MIT
MATH	[Hendrycks et al., 2021]	MIT
HellaSwag	[Zellers et al., 2019]	MIT

J.2 Measurements and Diagnostics of the Three Qwen2.5 Trajectories

Reward-curve fits. We estimate R^* by the mean reward over the final 50 iterations, set $V(k) = R^* - R(k)$, and fit a line to $\sqrt{\max(V(k), 0)}$ on a window selected by maximizing R^2 over five candidate starts with a moving-average endpoint rule (Table 7). The fitted intercept $\sqrt{\widehat{V}_0}$ and the negative of the fitted slope, $\hat{\beta}$, give the quadratic reference form $(\sqrt{\widehat{V}_0} - \hat{\beta}k)_+^2$ normalized in Figure 3a, and $\widehat{T}_A = \sqrt{\widehat{V}_0}/\hat{\beta}$ is an extrapolated zero. The fits describe the similarity of the rescaled curves of these three trajectories.

Table 7: **Reward-curve fits of the three Qwen2.5 trajectories.**

Model	Fit window of k	R^2	$\widehat{T}_A$
Qwen2.5-0.5B	$[10, 130)$	0.991	277
Qwen2.5-3B	$[0, 158)$	0.983	274
Qwen2.5-7B	$[20, 141)$	0.948	317

Gram-matrix MSD. We compute the $T \times T$ Gram matrix $G_{ij} = \langle \Delta\theta_i, \Delta\theta_j \rangle$ of reconstructed intended update vectors ($T = 500$) without forming the $T \times D$ array, using $O(T^2)$ memory. Since $\| \sum_{i=k}^{k+\ell-1} \Delta\theta_i \|^2 = \sum_{i,j \in [k,\, k+\ell)} G_{ij}$, the time-averaged MSD averages block sums of G at lags $\ell = 1, \ldots, 100$ over the starting points $k = 300, \ldots, 499 - \ell$. This measures full-parameter intended-update displacements. The fitted MSD slopes are 1.0003, 1.0008, and 1.0012 times $\eta^2 D$ for Qwen2.5-{0.5B, 3B, 7B}; for Qwen2.5-0.5B, the slope is 4.118 against 4.117. Each model contributes a single trajectory, so the near-perfect linearity ($R^2 \approx 1.0000$) reflects time averaging within that trajectory, not agreement across independent runs.

Distributional and reward-spread diagnostics. For a descriptive view of the displacement, we compute fixed signed feature hashes $Z_k = \Phi\, \Delta\theta_k \in \mathbb{R}^{3000}$, where each column of Φ has one nonzero entry of magnitude $1/\sqrt{3000}$ with a random sign and a uniformly sampled bucket, and the same map is reused across iterations. For the descriptive normal Q–Q plot in the bottom row of Figure 7, we center the 3000 components of $\sum_{k=300}^{349} Z_k$ and divide them by their common standard deviation. We then sample 2000 components without replacement with seed 0 and plot them against normal quantiles. The top row shows the sample reward standard deviation $s_R(k)$ with divisor $N = 30$; over $k = 300, \ldots, 499$, its regression slopes are $\{1.80 \times 10^{-5}, 1.22 \times 10^{-5}, -8.12 \times 10^{-6}\}$ per iteration

with $R^2 = \{0.010, 0.039, 0.029\}$, so a linear trend explains at most 4% of the variation of s_R in this window.

J.3 Models, Tasks, and Training Budgets

We extend the model and task comparison to the seven models and four tasks in Table 8. Besides the three Qwen2.5 models, the checkpoints are OLMo-2-0425-1B-Instruct [Allen Institute for AI, 2025], SmolLM2-1.7B-Instruct [HuggingFaceTB, 2025], Llama-3.2-3B-Instruct [Meta, 2024], and Phi-4-mini-instruct [Microsoft, 2025]. The tasks are Countdown [Gandhi et al., 2024, Pan et al., 2025], the GSM8K [Cobbe et al., 2021] and SVAMP [Patel et al., 2021] training splits, and the numeric-answer training subset of MATH [Hendrycks et al., 2021]; dataset providers are identified with the references. The grid contains 62 trajectories across 28 cells at $N = 30$, with two or three independent seeds per cell. Of these trajectories, 59 contain 500 iterations and three extend to 2000; grid analyses use only the first 500 iterations. The population and long-horizon groups contain 18 and five trajectories, respectively, and overlap the grid: the three groups together contain 74 distinct runs, rather than 85 independent trials. All runs use the same ES with $\alpha = 5 \times 10^{-7}$ and $\sigma = 10^{-3}$. Model identifiers, revisions, parameter counts, and run membership are specified in the experiment manifests of the code repository (`https://github.com/syokoi/es-dynamics-code`): `training/experiments.json` and the group-specific `data/experiments.json` files.

Computational resources. The LLM experiments used NVIDIA GH200 servers with vLLM inference backends, with four GPUs per model for the three Qwen2.5 trajectories of Appendix J.2 and single-GPU training configurations for the 74 runs above. These 74 runs comprise 44,625 completed training iterations, with training-loop times totaling about 1,150 hours. Individual runs took about 2–34 hours for 500 iterations and 9–47 hours for 2000 iterations. For a serial one-GPU rerun on a GH200, the single-GPU timings and matching Countdown configurations for the three Qwen2.5 models give a planning estimate of about 1,200 hours for all 77 training loops, assuming comparable throughput. These times and the rerun estimate exclude initial model loading, replay on restart, separate held-out evaluations, and final saving. Implementation checks and analyses required additional computation beyond the reported experiments.

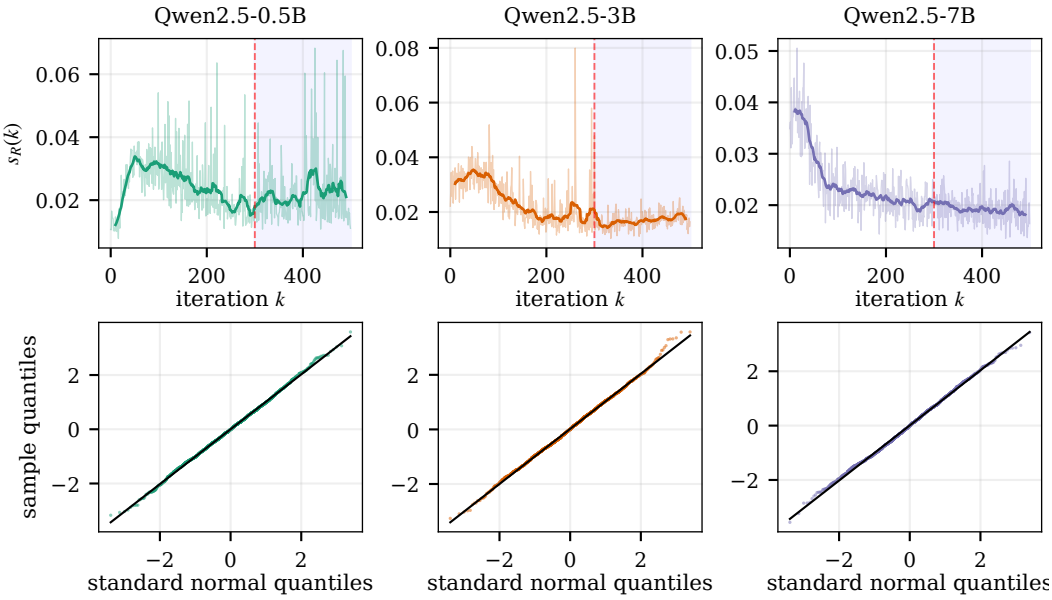


Figure 7: **Diagnostics of the three Qwen2.5 trajectories.** Top: sample reward standard deviation $s_R(k)$ (thin) and its 20-iteration moving average (thick); the shaded window starts at $k = 300$. Bottom: normal Q–Q plot of one fixed-hash intended-update displacement from iteration 300 to 350, with a fitted line.

Table 8: **Realized-weight displacement across models and tasks.** Slopes of squared full-parameter displacement per nonzero update (mean $\pm$ sample standard deviation across seeds; †: three seeds, otherwise two) and the nominal slope $\eta^2 D$.

Model	Nominal	Countdown	GSM8K	MATH	SVAMP
Qwen2.5-0.5B	4.12	$4.16 \pm 0.01^{\dagger}$	$4.15 \pm 0.01^{\dagger}$	4.15 ± 0.02	4.15 ± 0.01
Qwen2.5-3B	25.72	26.15 ± 0.03	26.05 ± 0.08	26.07 ± 0.06	26.16 ± 0.08
Qwen2.5-7B	63.46	64.08 ± 0.20	64.09 ± 0.06	63.77 ± 0.05	63.89 ± 0.31
OLMo-2-0425-1B	12.37	12.86 ± 0.03	12.856 ± 0.001	12.810 ± 0.003	12.86 ± 0.01
SmolLM2-1.7B	14.26	15.89 ± 0.02	15.96 ± 0.06	15.93 ± 0.07	15.97 ± 0.03
Llama-3.2-3B	26.77	$27.21 \pm 0.11^{\dagger}$	$26.95 \pm 0.09^{\dagger}$	26.97 ± 0.09	27.00 ± 0.04
Phi-4-mini	31.97	$33.03 \pm 0.05^{\dagger}$	$33.01 \pm 0.06^{\dagger}$	33.04 ± 0.01	33.31 ± 0.28

Table 9: **Fixed-budget Countdown comparison on Qwen2.5-0.5B** ($NT = 15000$). Reward: mean of the final five population rewards; displacement: $\|\theta_T - \theta_0\|^2/(\eta^2 DT)$; HellaSwag drop: initialization minus final accuracy on the same 500 items, in percentage points. Mean $\pm$ sample standard deviation over n seeds.

(N, T)	n	Terminal reward	Displacement / nominal	HellaSwag drop
$(15, 1000)$	3	0.3350 ± 0.0063	1.002 ± 0.002	16.2 ± 1.9
$(30, 500)$	3	0.3657 ± 0.0023	1.013 ± 0.001	4.1 ± 1.5
$(60, 250)$	3	0.3070 ± 0.0115	1.036 ± 0.002	2.1 ± 0.6
$(120, 125)$	1	0.2561	1.110	-0.8

Full-parameter displacement and reward profiles. These 74 runs record squared displacement of the realized weights from initialization, $\|\theta_k - \theta_0\|^2$ after k completed updates. Table 8 gives its fitted slopes per nonzero update; the fits include an intercept and use iterations 300–499, corresponding to completed updates 301–500, with time counting updates of norm at least 10^{-9} and iterations with update norms below this threshold omitted from the regression. The cell-mean ratios to $\eta^2 D$ range from 1.005 to 1.120, and all 62 within-trajectory regressions have $R^2 > 0.999$.

For each grid trajectory, we fit the descriptive curve $r_i = r_\infty - A(1 - i/h)_+^2$, $A > 0$, the quadratic reference form of Equation (17) in reward units, to the population rewards at iterations $i = 25, \ldots, 499$. For each candidate $h \in \{30, 35, \ldots, 2995\}$, we estimate r_∞ and A by unconstrained least squares and select the h with the smallest residual sum of squares among fits with $A > 0$ and a design matrix of rank two. The reward-profile comparison excludes two cells: Qwen2.5-0.5B on MATH, whose reward turns downward, and OLMo-2-0425-1B on Countdown, which includes a trajectory with 154 zero-update iterations. Both cells remain in the displacement comparison. Among the 26 retained cells, 25 have a seed-mean $R^2 \geq 0.93$; Llama-3.2-3B on MATH has $R^2 = 0.891$. The ratio of the sample standard deviation of fitted h across seeds to its mean has median 9% and maximum 39% across these cells.

Population size and fixed budget. On Qwen2.5-0.5B, the population group contains three Countdown seeds at each $N \in \{8, 15, 30, 60\}$ and one at $N = 120$; GSM8K has one seed each at $N = 15, 60$ and three at $N = 30$. For Countdown at $T = 500$, displacement slopes over iterations 300–499, using raw iteration time, have ratios 3.712, 1.985, 1.000, and 0.518 relative to $N = 30$, compared with nominal ratios 3.750, 2.000, 1.000, and 0.500 (Figure 3c). At the fixed candidate-evaluation budget $NT = 15000$ (Table 9), larger N reduces the accumulated displacement $\|\theta_T - \theta_0\|^2$ before the allocation-specific normalization in the table, but the highest terminal reward among the measured allocations occurs at $N = 30$; these rankings compare seed means. The auxiliary $N = 120$, $T = 125$ trajectory is a complete allocation at this budget and has no between-seed uncertainty estimate. The $N = 8$ runs stop at 500 iterations and do not reach the required $T = 1875$ for this comparison.

Held-out accuracy and observation horizon. The Rowan/hellaswag validation split [Zellers et al., 2019] is evaluated every 25 completed updates on 500 items. Choice log-likelihoods are normalized by the UTF-8 byte length of each space-prefixed choice. Across the 62 grid trajectories, 56 accuracy slopes against the square root of completed updates 25–500 are negative and six are

nonnegative; the four-task mean slope is negative for each model. An individual accuracy has resolution $1/500 = 0.002$ and a binomial standard error of about 0.022 near accuracy $1/2$.

To distinguish sustained changes from local peaks, we compare the final-50 raw-reward mean with the maximum trailing 25-iteration mean over the stated horizon. At $T = 500$, this ratio is 0.867 ± 0.058 for Countdown at $N = 8$, compared with 0.985 ± 0.004, 0.993 ± 0.002, and 0.995 ± 0.001 at $N = 15, 30, 60$ (mean $\pm$ sample standard deviation over three seeds each). The $N = 8$ change in the trailing mean from iteration 262 to 499 is -0.0056 ± 0.0096 (mean $\pm$ sample standard deviation over three seeds), with one seed increasing; a peak-based decline thus does not imply a negative net change for every trajectory on that interval. Over 2000 iterations, the two Countdown trajectories at $N = 15$ have final-to-peak ratios 0.841 and 0.850, whereas the two at $N = 30$ have ratios 0.995 and 0.981 and continue improving well beyond 500 iterations (Figure 3d). The single GSM8K trajectory at $N = 30$ has ratio 0.674. At 2000 iterations, HellaSwag is available for only one Countdown seed at each of $N = 15, 30$, and its drops are 27.4 and 17.6 percentage points, respectively; the GSM8K evaluation stops at 500 iterations. These observations show task- and horizon-dependent reward profiles and a fixed-budget trade-off between training reward and retained accuracy.

Per-model summary. Table 1 reports, for each model, the task mean of the cell-mean slopes in Table 8 divided by the nominal slope, the median over retained cells of the seed-mean reward-profile R^2, and the number of grid runs with a negative slope of HellaSwag accuracy against the square root of completed updates.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and contributions (C1)–(C3) in the introduction state the claims with their main conditions, corresponding to Theorems 3.1 and 4.3, Propositions 3.2, 4.4 and 4.5, and Remark 4.6, with experiments in Section 5; Section 6 and Table 3 give their scope and empirical status.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 states the main modeling assumptions and the scope of the LLM experiments, and Table 3 lists the conditions of each result with its empirical status.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each formal result states or references its assumptions (summarized in Table 3), and proofs are given in Appendices A, B, D to F and H; Remark 4.6 is qualitative guidance rather than a formal result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 and Appendix I specify the synthetic experiments, and Appendix J specifies the LLM models, data subsets, ES hyperparameters, and measurement procedures; the code repository records the model revisions and per-run configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiment code, run manifests, saved measurements, instructions for obtaining the public models and prompt datasets, and commands for LLM training, synthetic simulation, CPU reanalysis, and figure regeneration are available at `https://github.com/syokoi/es-dynamics-code`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and Appendix I give the synthetic settings, and Appendix J gives the LLM tasks, data subsets, rewards, hyperparameters, and analysis windows; the LLM runs use the α and σ of Qiu et al. [2026].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Figure 2b,c and the quadratic and oscillatory neutral-coordinate panels of Figure 6 show pointwise means with one-standard-error bands over independent trials, while Tables 8 and 9 report means and sample standard deviations across seeds. For the three Qwen2.5 trajectories of Appendix J.2, fit statistics describe individual trajectories rather than between-seed variability.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Synthetic conditions run in minutes on a laptop CPU, and LLM runs take hours to about two days on NVIDIA GH200 GPUs; Appendices I.4 and J.3 give the hardware, per-run times, and total budget. Implementation checks and analyses required additional computation beyond the reported experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All authors have reviewed the NeurIPS Code of Ethics and confirm that this research, which involves no human subjects or crowdsourcing, conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This foundational analysis of an existing fine-tuning method introduces no new model, dataset, or application, and we identify no direct path to negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work introduces no new models or datasets requiring special release safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Table 6 credits the models and datasets with their licenses, and the code repository records the model revisions. The LLM training code adapts routines from the ES implementation of Qiu et al. [2026] (`https://github.com/VsonicV/es-at-scale`, Academic Public License), as documented in the repository.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code repository, distributed under the Academic Public License, documents the code and saved measurements, including configurations, dependencies, commands, required external inputs, licenses, and known limitations; no new datasets or models are introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no human subjects research; IRB approval is therefore not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methods do not involve LLMs as important, original, or non-standard components; the pretrained LLMs in Appendices J.1 and J.3 are the models fine-tuned by ES in the experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.