Reward-Aware Population Scaling of Evolutionary Strategies in LLM Fine-Tuning

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Using Evolutionary Strategies (ES) for fine-tuning large language models is attractive because it is memory-efficient, parallel, and compatible with black-box or discrete rewards. Yet its population-size conclusions conflict sharply: fine-tuning with cross-entropy (CE) reward succeeds with $N = 1$ [7], while binary-reward training often needs $N \approx 30$ [9]. We show this gap is largely about reward design and normalization, not population size. Binary accuracy reward induces a zero-advantage probability q that depends in closed form on base accuracy, batch size, and intra-pair correctness correlation; a zero-training probe on Qwen2.5-Instruct/GSM8K matches the formula with mean absolute error 0.020 across 12 configurations and finds the availability threshold N_{avail} to be small in this capable-model regime. In this regime, z-score advantage normalization—not population size—can cause $N = 2$ to fail. Disabling normalization lets binary-reward ES with $N = 2$ improve on both GSM8K and TREC, where the normalized variant collapses or degrades. The implication is not that $N = 2$ is universally sufficient, but that small-population failure in capable-model binary ES can be an implementation artifact rather than an intrinsic population limit.

1. Introduction

Zeroth-order and ES methods are appealing for LLM fine-tuning because they avoid backpropagation through long sequences, parallelize across perturbations, and accept black-box or discrete rewards. Yet even the most essential hyperparameter, population size N, looks contradictory across prior work. Malladi et al. showed ES fine-tuning can work with one antithetic pair (i.e., $N = 1$) with CE reward function [7], whereas Qiu et al. claimed binary-reward ES often appears to need much larger populations (i.e. $N \approx 30$) [9]. We argue this is not a paradox but a *scaling law*: N is the scaling variable, and reward granularity together with normalization choices fix where on the law a given setup sits.

N controls how many reward-carrying perturbation directions are available per step. Below an availability threshold N_{avail}, almost no pair carries a usable signal and training stalls; reward granularity sets where N_{avail} lies, and advantage normalization can shift the threshold by erasing reward-scale information at small N.

Reward sparsity is not unique to ES; first-order LLM RL has motivated dense process and token-level reward schemes precisely because terminal binary supervision is weak [1, 2, 5, 6, 12, 13]. ES exposes a sharper form: because the estimator sees only scalar reward differences, quantization creates literal ties.

We make two contributions:

© .

1. **Reward granularity determines the availability threshold.** CE has $q = 0$ (trivial limit, $N_{\text{avail}} = 1$); binary has $q > 0$ with $N_{\text{avail}} \sim \log \delta / \log q$. We give an explicit binary approximation and validate it with a zero-training probe on Qwen2.5-Instruct (0.5B/1.5B/7B) on GSM8K (mean absolute error 0.020 across 12 configurations).

2. **Normalization—not population size—causes small-N failure.** z-score advantage normalization erases reward-scale information at small N (exactly so at $N = 2$). On Qwen2.5-1.5B with both GSM8K and TREC, simply disabling normalization recovers a binary-reward $N = 2$ run that otherwise collapses—the apparent need for large populations is, in this regime, the implementation rather than the population.

Our claims are intentionally narrow. We do not claim $N = 2$ always works, nor that reward degeneracy alone explains every observed $N \approx 30$ threshold. The point is the joint mechanism—reward sparsity together with normalization—read inside a scaling-law frame.

2. Setup: ES as Smoothed High-Dimensional Dynamics

Notation. Let $\theta \in \mathbb{R}^d$ be model parameters, $\mathcal{B}$ a batch of size B, N the number of antithetic perturbation pairs, and $\varepsilon_i \sim \mathcal{N}(0, I_d)$. The pair-level reward difference is $A_i = R(\theta + \sigma\varepsilon_i, \mathcal{B}) - R(\theta - \sigma\varepsilon_i, \mathcal{B})$.

Raw and normalized dynamics. The raw ES estimator is

$$\hat{g}_{\text{raw}} = \frac{1}{N\sigma} \sum_{i=1}^{N} A_i \, \varepsilon_i, \tag{1}$$

and the normalized estimator divides the centered advantages by their empirical standard deviation:

$$\hat{g}_{\text{norm}} = \frac{1}{N\sigma} \sum_{i=1}^{N} \frac{A_i - \bar{A}}{s_A + \varepsilon_0} \, \varepsilon_i, \qquad \bar{A} = \frac{1}{N} \sum_{i=1}^{N} A_i, \tag{2}$$

where s_A is the empirical standard deviation of $\{A_i\}_{i=1}^{N}$ and $\varepsilon_0 \geq 0$ is an optional floor. Our experiments use the pair-level convention with $\varepsilon_0 = 0$.

Rewards and smoothing. We compare two rewards:

$$R_{\text{acc}} = \tfrac{1}{B} \textstyle\sum_j \mathbf{1}[\hat{y}_\theta(x_j) = y_j], \qquad R_{\text{CE}} = \tfrac{1}{B} \textstyle\sum_j \log p_\theta(y_j \mid x_j),$$

where R_{acc} is quantized in steps of $1/B$ and R_{CE} is continuous in the logits and requires white-box access. Earlier black-box prompt-tuning work observed a similar sparsity gap in derivative-free prompt search [11]; we extend it from prompt space to full-parameter ES.

ES optimizes the Gaussian-smoothed objective $f_\sigma(\theta) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, I_d)}[f(\theta + \sigma\varepsilon)]$, which is L_σ-smooth with $L_\sigma = 2/\sigma^2$ even when f is discontinuous (as for binary accuracy reward) [8]. Once the update is divided by the random statistic s_A, the estimator changes qualitatively.

3. A Reward-Aware Population Scaling Law

We treat N as a scaling variable controlling how many reward-carrying perturbation directions are available per ES update, with an availability threshold N_{avail} below which the update typically contains no reward-carrying direction and training fails. The position of the threshold is set by reward granularity: $N_{\text{avail}} = 1$ for dense (CE) reward, where Proposition 1 below records the resulting moment-form N-cancellation invariance; nontrivial for sparse (binary) reward, where Proposition 3 gives an explicit approximation.

Availability. The natural state variable is not just N, but the number of perturbation directions that survive degeneracy. Writing $q(\theta, \mathcal{B}, \sigma) = \mathbb{P}(A = 0)$ for the seed-level zero-advantage probability, we have

$$K_N = \sum_{i=1}^{N} \mathbf{1}[A_i \neq 0], \qquad \mathbb{P}(K_N = 0) = q^N, \qquad N_{\text{avail}}(\delta) = \left\lceil \frac{\log \delta}{\log q} \right\rceil. \tag{3}$$

Under the iid-seed approximation, $\mathbb{E}[K_N] = N(1-q)$. This turns population size into an availability question before it becomes a variance question: if $K_N = 0$, the update contains no reward-carrying direction at all, and $N_{\text{avail}}(\delta)$ is the smallest population for which at least one non-degenerate seed appears with probability at least $1 - \delta$.

Dense limit and N-cancellation. When $q = 0$ the availability threshold collapses to $N_{\text{avail}} = 1$, so any $N \geq 1$ is above threshold. There, the cumulative drift and diffusion across T steps become N-independent under the fixed-budget learning-rate scaling.

Proposition 1 (CE population indifference) *Let $G_i = \sigma^{-1}\big(R(\theta + \sigma\varepsilon_i) - R(\theta - \sigma\varepsilon_i)\big)\varepsilon_i$ and write the raw estimator (1) as $\hat{g}_N = N^{-1}\sum_{i=1}^{N} G_i$ to make the population dependence explicit. Under any reward continuous in θ (in particular CE),*

$$\mathbb{E}[\hat{g}_N] = 2\,\nabla f_\sigma(\theta), \qquad \mathrm{Cov}(\hat{g}_N) = \Sigma(\theta)/N,$$

with $\Sigma(\theta) = \mathrm{Cov}(G_i)$. Consequently, at fixed total perturbation-pair budget $M = NT$ and learning-rate scaling $\eta_N = N\eta_0$, the cumulative drift and diffusion across T steps,

$$\mathbb{E}[\Delta\theta_N] = 2\,M\eta_0\,\nabla f_\sigma(\theta), \qquad \mathrm{Cov}(\Delta\theta_N) = M\eta_0^2\,\Sigma(\theta),$$

are both independent of N in the local linearization around θ.

The proof is in Appendix C.

Corollary 2 (No availability threshold under CE) *Under standard smooth-network assumptions (Appendix C), $K_N = N$ almost surely whenever $\nabla_\theta R_{\text{CE}}(\theta, \mathcal{B}) \neq 0$, and Proposition 1 applies seedwise; see Appendix F for matching CE-reward scaling curves.*

Sparse regime. Binary reward makes $q > 0$, so the threshold is nontrivial and K_N becomes a genuine random count rather than identically N.

Proposition 3 (Binary degeneracy approximation) *Under a small-perturbation, homogeneous-batch approximation,*

$$q = \mathbb{P}(A_{\text{acc}} = 0) \approx \frac{1}{\sqrt{4\pi B\, p_0(1-p_0)(1-\rho)}}, \tag{4}$$

where p_0 is the base accuracy and ρ is the intra-pair correctness correlation.

The right-hand side is a local-CLT point-density approximation at zero (Appendix D); in the extreme sparse regime (B small or p_0 near $\{0, 1\}$) it can formally exceed 1, in which case we use $\min\{1, \cdot\}$ for numerical work and prefer the empirical probe (Appendix E) for calibration. Above N_{avail}, the effective non-degenerate count $K_N \to N(1-q)$ in expectation and population averaging behaves in the usual way; below N_{avail}, the random count K_N itself is the bottleneck and no learning-rate rescaling can recover a gradient-carrying update from a step where $K_N = 0$. Appendix C pinpoints where the CE argument breaks in this regime.

Empirical probe. A zero-training probe on Qwen2.5-Instruct (0.5B/1.5B/7B) on GSM8K validates (4) with MAE 0.020 across 12 configurations (Appendix E). For Qwen2.5-1.5B at $B = 16$, $\hat{q} \approx 0.19$, giving $\mathbb{P}(K_N = 0) \approx 0.036$ at $N = 2$—so the availability threshold is small here, yet normalized binary-reward ES still fails at $N = 2$. The remaining mechanism is §4.

4. Normalization Distorts the Threshold

Since $\mathbb{P}(K_N = 0) \approx 0.036$ at $N = 2$ for Qwen2.5-1.5B, pure availability cannot explain the $N = 2$ collapse—the remaining mechanism is advantage normalization. Raw ES is *self-annealing*: if reward differences shrink near degeneracy, the update shrinks with them. Z-score normalization replaces that magnitude with a small-N scale estimate computed from a handful of sparse advantages, removing the self-annealing and providing an ES-specific instance of how scale-uncontrolled training signals can create new failure modes [3].

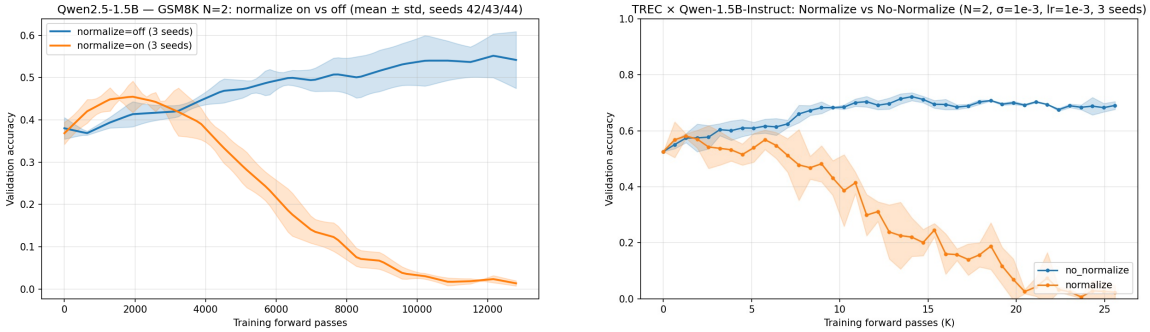
Proposition 4 (Two-sample z-score erases advantage scale) *Let $A_1 \neq A_2$ be pair-level advantages, let $\bar{A} = (A_1 + A_2)/2$, and let s_A be their empirical standard deviation. Then the normalized vector*

$$\left(\frac{A_1 - \bar{A}}{s_A}, \frac{A_2 - \bar{A}}{s_A}\right)$$

is independent of the absolute gap $|A_1 - A_2|$ up to a convention-dependent constant. With population standard deviation it is $(\pm 1, \mp 1)$; with sample standard deviation it is $(\pm 1/\sqrt{2}, \mp 1/\sqrt{2})$.

The proof is in Appendix G. Z-score normalization always removes absolute scale, but at $N = 2$ even ratio information disappears: an infinitesimal advantage gap and a large advantage gap therefore produce the same normalized update magnitude.

Under sparse binary rewards this matters most. Many A_i are exactly zero, and the nonzero ones can be very small near degeneracy: with raw advantages such steps naturally decay, but normalization promotes the one or two surviving nonzero seeds to fixed-size updates while estimating s_A from only a handful of sparse rewards. Normalization does not merely standardize the signal—it changes the stochastic dynamics.



(a) Qwen2.5-1.5B/GSM8K. (b) Qwen2.5-1.5B/TREC

Figure 1: **Turning off advantage normalization recovers small-population ES.** At $N = 2$ with binary reward, standard advantage normalization can collapse or degrade, whereas raw advantages preserve reward-scale information and can improve substantially. Left: Qwen2.5-1.5B on GSM8K. Right: Qwen2.5-1.5B on TREC.

The binary population-scaling curves in Appendix F use the default normalized implementation; Figure 1 is a single fixed-everything ablation that varies only the treatment of advantages. With normalization on, the $N = 2$ run briefly rises and then collapses on GSM8K; with normalization off, the same binary-reward ES setup improves steadily to roughly 0.55 validation accuracy on GSM8K and similarly improves on TREC. We do not claim $N = 2$ is universally sufficient or that normalization explains every small-N failure—but in this setting, the apparent $N = 2$ collapse is the implementation, not the population size. Within the scaling-law frame, normalization is the implementation knob that decides whether N near N_{avail} behaves like the raw self-annealing regime or like a fixed-magnitude random walk that ignores the very scale information N_{avail} was defined to track.

5. Discussion and Limitations

Practical recommendation: estimate N_{avail} via the zero-training probe (Appendix E) and start with the smallest N that clears it—in the Qwen2.5-Instruct/GSM8K regime we measured, $N_{\text{avail}} \leq 4$ across all 12 configurations and $N = 2$ with raw advantages was already viable. Tune upward only if optimizer stability, normalization, or wall-clock parallelism require it; near N_{avail}, raw or clipped-std advantages are safer than z-scoring. This complements curvature-based ES analyses [4]: geometry determines whether useful directions exist; reward sparsity and normalization determine whether scalar feedback reveals them.

Three caveats. (a) The normalization-off recovery of $N = 2$ is one model-task setting (Qwen2.5-1.5B on GSM8K/TREC), not a universal claim. (b) The degeneracy model is an availability bound under homogeneous-batch, small-perturbation assumptions; (4) is an approximation, and we treat ρ as a complementary diagnostic only—it indexes the variance of the reward difference but does not predict whether the surviving differences align with useful parameter directions. (c) Proposition 1's drift–diffusion invariance is a moment-level statement under local linearization at fixed θ; we do not derive a descent-level N-cancellation result and make no claim about whether different N values produce equivalent full training trajectories at fixed compute.

References

[1] Alex James Chan, Hao Sun, Samuel Holt, and Mihaela Van Der Schaar. Dense reward for free in reinforcement learning from human feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6136–6154. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chan24a.html.

[2] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards, 2025. URL https://arxiv.org/abs/2502.01456.

[3] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective rl reward at training time for llm reasoning, 2024. URL https://arxiv.org/abs/2410.15115.

[4] 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, 2026. URL https://arxiv.org/abs/2602.00170.

[5] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

[6] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision, 2024. URL https://arxiv.org/abs/2406.06592.

[7] 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*, 2023.

[8] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[9] 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, 2026. URL https://arxiv.org/abs/2509.24372.

[10] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning, 2017. URL https://arxiv.org/abs/1703.03864.

[11] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on*

Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pages 20841–20855. PMLR, 2022.

[12] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9426–9439. Association for Computational Linguistics, 2024.

[13] Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Gunsoo Han, Daniel Nam, Daejin Jo, Kyoung-Woon On, Mark Hasegawa-Johnson, Sungwoong Kim, and Chang Yoo. TLCR: Token-level continuous reward for fine-grained reinforcement learning from human feedback. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14969–14981, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.889. URL https://aclanthology.org/2024.findings-acl.889/.

Appendix A. Algorithmic Conventions: MeZO and ES-at-Scale

Estimator normalization. The two paradigms our paper engages with differ only in the constant in front of the antithetic difference [7, 10]. MeZO uses

$$\hat{g}_{\text{MeZO}} = \frac{R(\theta + \sigma\varepsilon) - R(\theta - \sigma\varepsilon)}{2\sigma}\,\varepsilon, \tag{5}$$

which satisfies $\mathbb{E}[\hat{g}_{\text{MeZO}}] = \nabla f_\sigma(\theta)$ directly via the antithetic Stein identity (Appendix B). ES-at-scale [9] uses the $1/\sigma$ form of (1) instead, giving $\mathbb{E}[\hat{g}_N] = 2\,\nabla f_\sigma(\theta)$ with the factor of two absorbed into the learning rate. The two are equivalent at the level of the expected update once learning rates are matched: at literal hyperparameter η, the ES estimator moves twice as far per step as MeZO, so ES's effective rate is 2η relative to MeZO. We follow the $1/\sigma$ convention throughout to align with Qiu et al. [9]; Proposition 1 and Appendix B are stated under the same convention.

Which quantities are normalized. Our analysis assumes the implementation normalizes the pair-level advantages $A_i = R(\theta + \sigma\varepsilon_i) - R(\theta - \sigma\varepsilon_i)$ across $i = 1, \ldots, N$. Under this convention, the $N = 2$ scale-erasure claim of Proposition 4 follows from two-sample z-scoring of (A_1, A_2); it does *not* require assuming that the algorithm separately normalizes the positive and negative rollouts or that the pair-level advantages are exactly $\{+a, -a\}$. Some MeZO and ES implementations optionally divide A_i by the mini-batch standard deviation; whether this helps depends on the regime, as Section 4 shows for binary reward.

Conceptual difference between the two paradigms. The distinction between MeZO and ES-at-scale is not which prefactor is "correct"—both produce unbiased estimates of ∇f_σ up to a constant factor. The substantive difference is the reward signal:

- **MeZO** uses CE reward, which is white-box and incurs no availability threshold (Corollary 2); $N = 1$ is sufficient.
- **ES-at-scale** uses binary accuracy reward, which is black-box compatible but sparse (Proposition 3); the reported $N \approx 30$ is consistent with the availability threshold being substantial in their base-accuracy regime, though our two-mechanism story suggests normalization (Section 4) likely contributes as well.

Both choices are essentially optimal for their setting. The $N \approx 30$ floor reflects the interaction of binary reward with the base-accuracy regime of their structured-prompt evaluation, rather than a fundamental property of the LLM or task; in our Qwen2.5/GSM8K probe (Appendix E), the same mechanism gives a much smaller N_{avail}.

Clipped-standard-deviation normalization. When we refer to a possible remedy, we mean replacing (2) by

$$\tilde{A}_i = \frac{A_i - \bar{A}}{\max(s_A, \varepsilon_0)}$$

for a floor $\varepsilon_0 > 0$. The experiments in the main text use $\varepsilon_0 = 0$; the clipped version is a proposed intervention rather than a validated result here.

Appendix B. Gaussian Smoothing, Stein Identity, and L_σ-Smoothness

Definition and approximation gap. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the reward objective and define $f_\sigma(\theta) = \mathbb{E}_\varepsilon[f(\theta + \sigma\varepsilon)]$ with $\varepsilon \sim \mathcal{N}(0, I_d)$. For M-Lipschitz f, $|f_\sigma(\theta) - f(\theta)| \le M\sigma$; at $\sigma = 10^{-3}$ this gap is negligible compared to initial suboptimality.

Stein gradient identity. By Gaussian integration by parts, $\nabla f_\sigma(\theta) = \sigma^{-1}\,\mathbb{E}[f(\theta + \sigma\varepsilon)\,\varepsilon]$. For the antithetic pair, $\mathbb{E}[(f(\theta + \sigma\varepsilon) - f(\theta - \sigma\varepsilon))\,\varepsilon] = 2\sigma\,\nabla f_\sigma(\theta)$. The $1/(2\sigma)$-prefactor estimator (e.g., MeZO) is unbiased for ∇f_σ exactly; the $1/\sigma$ convention $\hat{g}_{\text{raw}}$ used in the main text differs by a factor of two absorbed into the learning rate. Normalizing by the data-dependent statistic s_A changes the estimator qualitatively and removes this direct unbiasedness interpretation.

L_σ-smoothness.

Theorem 5 (Nesterov and Spokoiny 8) *If $|f| \le 1$, then f_σ is L_σ-smooth with $L_\sigma = 2/\sigma^2$.*

Proof The second-order Stein identity gives $[\nabla^2 f_\sigma(\theta)]_{kl} = \sigma^{-2}\,\mathbb{E}_\varepsilon[f(\theta + \sigma\varepsilon)(\varepsilon_k\varepsilon_l - \delta_{kl})]$. For any unit vector $v \in \mathbb{R}^d$,

$$v^\top \nabla^2 f_\sigma\, v = \frac{1}{\sigma^2}\,\mathbb{E}_\varepsilon\big[f(\theta + \sigma\varepsilon)\big((v \cdot \varepsilon)^2 - 1\big)\big]\,.$$

Bounding with $|f| \le 1$ and $\mathbb{E}[(v \cdot \varepsilon)^2] = 1$ gives $|v^\top \nabla^2 f_\sigma\, v| \le \sigma^{-2}\mathbb{E}[(v \cdot \varepsilon)^2 + 1] = 2/\sigma^2$. This holds for all unit v, so $\|\nabla^2 f_\sigma\|_{\text{op}} \le 2/\sigma^2 = L_\sigma$. ■

The crucial feature is that no smoothness assumption on f is required: binary accuracy reward is a step function ($L = \infty$ classically) yet f_σ has finite smoothness. The descent lemma for f_σ is therefore exact:

$$\mathbb{E}[f_\sigma(\theta + \eta\hat{g})] - f_\sigma(\theta) \ge \eta\|\nabla f_\sigma\|^2 - \frac{\eta^2 L_\sigma}{2}\,\mathbb{E}[\|\hat{g}\|^2].$$

Remark 6 (Cocoercivity without PL) *L_σ-smoothness alone gives $\|\nabla f_\sigma(\theta)\|^2 \le 2L_\sigma(f_\sigma^* - f_\sigma(\theta))$: a single ascent step at rate $1/L_\sigma$ from θ improves f_σ by $\|\nabla f_\sigma\|^2/(2L_\sigma)$, and this cannot exceed $f_\sigma^* - f_\sigma(\theta)$. No PL condition is needed.*

Appendix C. CE Population Indifference and Non-Degeneracy

Proof of Proposition 1. With $G_i = \sigma^{-1}(R(\theta + \sigma\varepsilon_i) - R(\theta - \sigma\varepsilon_i))\varepsilon_i$ and $\hat{g}_N = N^{-1}\sum_i G_i$, the antithetic Stein identity from Appendix B gives

$$\mathbb{E}[(R(\theta + \sigma\varepsilon) - R(\theta - \sigma\varepsilon))\,\varepsilon] = 2\sigma\,\nabla f_\sigma(\theta),$$

hence $\mathbb{E}[G_i] = 2\,\nabla f_\sigma(\theta)$ and $\mathbb{E}[\hat{g}_N] = 2\,\nabla f_\sigma(\theta)$. The seeds ε_i are iid, so the G_i are iid with covariance $\Sigma(\theta) = \mathrm{Cov}(G_i)$, and

$$\mathrm{Cov}(\hat{g}_N) = \frac{1}{N}\,\Sigma(\theta).$$

For the budget argument, fix θ and consider one ES step at population N with learning rate $\eta_N = N\eta_0$. The single-step expected drift is

$$\eta_N\,\mathbb{E}[\hat{g}_N] = 2\,N\eta_0\,\nabla f_\sigma(\theta),$$

and the single-step covariance is

$$\eta_N^2 \operatorname{Cov}(\hat{g}_N) = N\eta_0^2 \Sigma(\theta).$$

Across $T = M/N$ independent steps, under the local linearization that θ does not drift far from the reference point so that ∇f_σ and Σ are constant,

$$\mathbb{E}[\Delta\theta_N] = T\,\eta_N\,\mathbb{E}[\hat{g}_N] = 2\,M\eta_0\,\nabla f_\sigma(\theta), \qquad \operatorname{Cov}(\Delta\theta_N) = T\,\eta_N^2 \operatorname{Cov}(\hat{g}_N) = M\eta_0^2\,\Sigma(\theta).$$

Both quantities are independent of N.

Proof of Corollary 2. Fix a batch $\mathcal{B}$ and define

$$h(\varepsilon) = R_{\mathrm{CE}}(\theta + \sigma\varepsilon, \mathcal{B}) - R_{\mathrm{CE}}(\theta - \sigma\varepsilon, \mathcal{B}).$$

Under standard transformer architectures—compositions of analytic operations (matrix multiplication, softmax, smooth activations such as GELU/SwiGLU)—$R_{\mathrm{CE}}(\theta, \mathcal{B})$ is real analytic in θ, so h is real analytic in ε. Its gradient at the origin is

$$\nabla_\varepsilon h(0) = 2\sigma\, g_{\mathrm{CE}}(\theta, \mathcal{B}), \qquad g_{\mathrm{CE}}(\theta, \mathcal{B}) = \nabla_\theta R_{\mathrm{CE}}(\theta, \mathcal{B}) = \frac{1}{B}\sum_{j=1}^{B} \nabla_\theta \log p_\theta(y_j \mid x_j).$$

If $g_{\mathrm{CE}}(\theta, \mathcal{B}) \neq 0$, then h is not identically zero; the zero set of a non-identically-zero real analytic function on $\mathbb{R}^d$ has Lebesgue measure zero, and since the Gaussian measure is absolutely continuous with respect to Lebesgue measure, $\mathbb{P}(h(\varepsilon) = 0) = 0$. Hence $K_N = \sum_i \mathbf{1}[A_i \neq 0] = N$ almost surely, and the variance-averaging structure used in the proof of Proposition 1 is well-defined seedwise.

Where the argument breaks for binary reward. The decomposition $\operatorname{Cov}(\hat{g}_N) = \Sigma/N$ assumed each seed contributes independently to the per-step covariance. Under binary reward, this fails because the effective sample size in the variance averaging is the random count K_N, not N. When $K_N = 0$, the single-step update is either zero (raw advantages) or normalization-dependent (normalized advantages), and no learning-rate rescaling $\eta_N \propto N$ can recover a gradient-carrying update. This is the precise sense in which CE population indifference is dense-reward-specific.

Appendix D. Binary Degeneracy: Full Proof

Setup. Let $X_j^+, X_j^- \in \{0, 1\}$ be the correctness indicators for example j under $\theta \pm \sigma\varepsilon$. Under the homogeneous-batch approximation, $\mathbb{E}[X_j^+] = \mathbb{E}[X_j^-] = p_0$ and $\operatorname{Corr}(X_j^+, X_j^-) = \rho$. Define $Y_j = X_j^+ - X_j^- \in \{-1, 0, 1\}$; then $\mathbb{E}[Y_j] = 0$ and $\operatorname{Var}(Y_j) = 2p_0(1 - p_0)(1 - \rho) \triangleq v$. The batch-level accuracy difference is proportional to $S_B = \sum_{j=1}^{B} Y_j$, and $A_{\mathrm{acc}} = 0$ iff $S_B = 0$.

Local CLT, not standard CLT. The standard CLT approximates CDFs, but we need the point probability $\mathbb{P}(S_B = 0)$. Gnedenko's local CLT works directly on integer-valued PMFs: for iid integer summands $Y_1, \ldots, Y_B$ with mean 0 and variance v,

$$\left| \mathbb{P}(S_B = k) - \frac{1}{\sigma_S\sqrt{2\pi}}\,\varphi\!\left(\frac{k}{\sigma_S}\right) \right| \leq \frac{C\,\mathbb{E}[|Y_j|^3]}{\sigma_S^3}, \tag{6}$$

where $\sigma_S^2 = Bv$, φ is the standard normal density, and $C > 0$ is a universal constant. Setting $k = 0$ and $\varphi(0) = 1/\sqrt{2\pi}$ gives the leading-order $\mathbb{P}(S_B = 0) = (2\pi Bv)^{-1/2}$ with a Berry–Esseen error of $O(B^{-1/2})$.

$O(B^{-1})$ error: vanishing third cumulant. For our $Y_j \in \{-1, 0, +1\}$ the error is one order better. Since $Y_j^3 = Y_j$ exactly, the third cumulant

$$\kappa_3 = \mathbb{E}[Y_j^3] = \mathbb{E}[Y_j] = 0.$$

The Edgeworth expansion at $k = 0$ reads

$$\mathbb{P}(S_B = 0) = \frac{1}{\sqrt{2\pi B v}} \left[1 + \frac{\kappa_3}{6v^{3/2}\sqrt{B}} H_3(0) + \frac{\kappa_4}{24 v^2 B} H_4(0) + O(B^{-3/2}) \right],$$

with $H_3(0) = 0$ and $H_4(0) = 3$. The $O(B^{-1/2})$ Berry–Esseen term proportional to $\kappa_3 H_3(0)$ vanishes identically, leaving the $O(B^{-1})$ fourth-cumulant term as leading. Substituting $\sigma_S^2 = Bv$ yields Proposition 3:

$$\mathbb{P}(A_{\text{acc}} = 0) = \frac{1}{\sqrt{4\pi B\, p_0(1-p_0)(1-\rho)}} + O(B^{-1}).$$

Non-IID heterogeneity. For heterogeneous batches with per-example variance $v_j = 2p_j(1 - p_j)(1 - \rho_j)$, Petrov's non-iid local CLT applies with $\text{Var}(S_{\text{het}}) = \sum_j v_j = B\bar{v}$ where $\bar{v} = B^{-1} \sum_j v_j$. At leading order, $\mathbb{P}(S_{\text{het}} = 0) \approx (2\pi B\bar{v})^{-1/2}$—the same as Proposition 3 with v replaced by $\bar{v}$; higher-order corrections depend on the empirical distribution of $\{v_j\}$ and we do not characterize them here. The worst case is bimodal accuracy (p_j near 0 or 1), where each $v_j \approx 0$ and $\mathbb{P}(A = 0) \to 1$. The empirical probe (Appendix E) measures $\hat{\mathbb{P}}(A = 0)$ directly and is preferred whenever the homogeneous-batch assumption is doubtful.

Appendix E. Pre-Training Degeneracy Probe: Protocol and Results

The degeneracy probe is a forward-pass-only diagnostic that we use to validate Proposition 3 and that practitioners can also run before committing to a population size.

Protocol.

1. Sample K perturbation pairs $\{\varepsilon_i\}_{i=1}^K$ from $\mathcal{N}(0, I_d)$ at the σ intended for training.
2. Evaluate $R(\theta_0 + \sigma\varepsilon_i, \mathcal{B})$ and $R(\theta_0 - \sigma\varepsilon_i, \mathcal{B})$ on a fixed batch of size B. Total cost: $2KB$ forward passes.
3. Estimate the zero-advantage rate $\hat{q} = K^{-1} \left|\{i : R(\theta_0 + \sigma\varepsilon_i, \mathcal{B}) = R(\theta_0 - \sigma\varepsilon_i, \mathcal{B})\}\right|$, the base accuracy p_0, and the intra-pair correctness correlation $\hat{\rho}$ from the paired indicators.
4. Compute $N_{\text{avail}}(\delta) = \lceil \log \delta / \log \hat{q} \rceil$ at the desired confidence (we use $\delta = 0.05$ throughout).

For $K = 200$ and $B = 16$, the probe costs 6,400 forward passes—on the order of a handful of training iterations—and returns a data-driven population floor that bypasses the homogeneous-batch approximation in Proposition 3. We recommend it as a cheaper alternative to sweeping N empirically: under binary reward, any choice of $N < N_{\text{avail}}$ is dominated by the availability failure mode of Section 3, irrespective of optimizer or learning-rate tuning. The same probe outputs also yield $\hat{\rho}$ at no extra forward-pass cost, giving a complementary read on whether the local reward landscape is becoming nearly frozen ($\rho \to 1$).

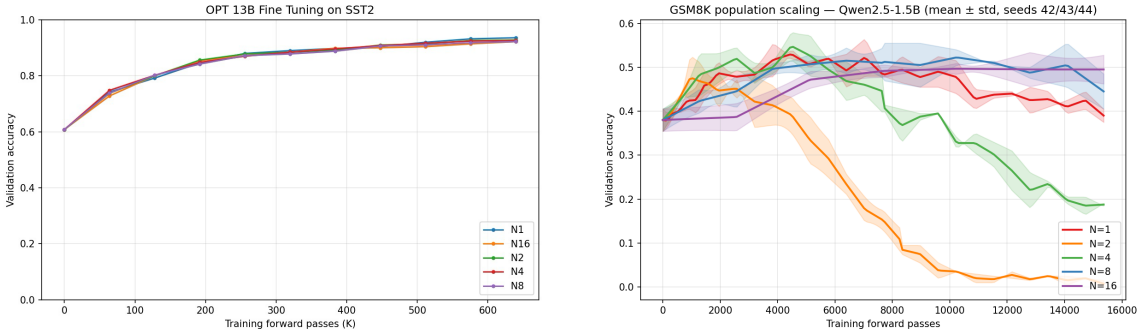
Table 1: Full pre-training degeneracy probe on GSM8K at $\sigma = 10^{-3}$ with $K = 200$ perturbation pairs. $|\Delta q| = |\text{empirical } q - \text{predicted } q|$; mean absolute error across the 12 rows is 0.020.

Model	B	p_0	$\hat{\rho}$	empirical q	predicted q	$\lvert\Delta q\rvert$	N_{avail}
Qwen2.5-0.5B	4	0.306	0.357	0.345	0.382	0.037	3
Qwen2.5-0.5B	8	0.306	0.379	0.240	0.275	0.035	3
Qwen2.5-0.5B	16	0.308	0.370	0.165	0.192	0.027	2
Qwen2.5-0.5B	32	0.308	0.362	0.140	0.135	0.005	2
Qwen2.5-1.5B	4	0.466	0.439	0.400	0.377	0.023	4
Qwen2.5-1.5B	8	0.466	0.475	0.290	0.276	0.014	3
Qwen2.5-1.5B	16	0.466	0.424	0.190	0.186	0.004	2
Qwen2.5-1.5B	32	0.466	0.449	0.125	0.135	0.010	2
Qwen2.5-7B	4	0.642	0.607	0.495	0.469	0.026	5
Qwen2.5-7B	8	0.642	0.622	0.340	0.338	0.002	3
Qwen2.5-7B	16	0.642	0.625	0.220	0.240	0.020	2
Qwen2.5-7B	32	0.642	0.612	0.130	0.167	0.037	2

Results. Table 1 reports the probe applied to Qwen2.5-Instruct (0.5B, 1.5B, 7B) on GSM8K with $K = 200$ and $\sigma = 10^{-3}$ across $B \in \{4, 8, 16, 32\}$.

Appendix F. Population Scaling Under CE vs Binary Reward

Figure 2 below shows population-scaling trajectories under CE versus binary reward. The binary runs all use the default advantage-normalized ES implementation; the normalization ablation is reported separately in Section 4.



(a) CE reward on OPT-13B/SST-2.

(b) Binary reward on Qwen2.5-1.5B (normalized ES).

Figure 2: **Reward choice changes the population-scaling regime.** Left: under CE reward with appropriate learning-rate scaling, trajectories are nearly population-indifferent, consistent with CE advantages being nonzero almost surely. Right: under binary reward, population size changes the qualitative fine-tuning dynamics; $N = 2$ and $N = 4$ fail while larger populations are more stable. All binary-reward runs use the default advantage-normalized ES implementation, so the small-N failures here motivate the normalization ablation in Figure 1 and should not be read as failures of raw-advantage ES at the same population sizes.

Appendix G. Advantage Normalization Analysis

Proof of Proposition 4. For two numbers A_1, A_2, their mean is $\bar{A} = (A_1 + A_2)/2$. Their centered values are

$$A_1 - \bar{A} = \frac{A_1 - A_2}{2}, \qquad A_2 - \bar{A} = -\frac{A_1 - A_2}{2}.$$

With the population-standard-deviation convention,

$$s_A = \sqrt{\frac{(A_1 - \bar{A})^2 + (A_2 - \bar{A})^2}{2}} = \frac{|A_1 - A_2|}{2},$$

so the normalized vector is $(\operatorname{sign}(A_1 - A_2), -\operatorname{sign}(A_1 - A_2))$. With the sample-standard-deviation convention,

$$s_A = \sqrt{(A_1 - \bar{A})^2 + (A_2 - \bar{A})^2} = \frac{|A_1 - A_2|}{\sqrt{2}},$$

so the normalized vector is $(\operatorname{sign}(A_1 - A_2)/\sqrt{2}, -\operatorname{sign}(A_1 - A_2)/\sqrt{2})$. In either case, the absolute gap $|A_1 - A_2|$ cancels.

Affine scale invariance for general N. For any vector $A \in \mathbb{R}^N$, z-score normalization is invariant to positive affine transformations $A \mapsto cA + b\mathbf{1}$. Thus normalization always removes absolute scale. At larger N, however, the centered *ratios* among coordinates still survive, which is why the pathology weakens with N.

$N = 4$ example. Suppose the four pair-level advantages are $(0, 0, \epsilon, 2\epsilon)$ with $\epsilon > 0$. After z-score normalization, the result is exactly the same as for $(0, 0, 1, 2)$. Absolute magnitude has disappeared; only the ratio pattern survives. This is less extreme than the $N = 2$ case, because the coordinates are not collapsed to a fixed sign vector, but it still removes the self-annealing of raw ES.

Why small N is unstable under sparse binary reward. Binary rewards are quantized, so many A_i are exactly zero and the surviving nonzero ones often come from one or two perturbation pairs. In that regime, the empirical standard deviation s_A is estimated from very few support points. The normalized update therefore mixes two effects: it discards absolute reward scale and simultaneously amplifies the noise of a poorly estimated scale statistic. A clipped-standard-deviation rule,

$$\tilde{A}_i = \frac{A_i - \bar{A}}{\max(s_A, \varepsilon_0)},$$

is one direct way to prevent this amplification while retaining some of normalization's scale control.

Appendix H. Spectral Decay: Resolving the Theory-Practice Gap

This appendix sits outside the main scaling-law argument: it addresses the stability concern raised by the worst-case L_σ bound of Theorem 5, which would forbid the empirically feasible learning rates used throughout this paper. Readers willing to take the empirical η regime as given may skip it.

Theorem 5 gives $L_\sigma = 2/\sigma^2 \approx 2 \times 10^6$ at $\sigma = 0.001$, implying stability $\eta < \sigma^2/2 \approx 5 \times 10^{-7}$. Yet ES fine-tuning runs successfully at $\eta \sim 10^{-3}$–10^{-4} — a gap of 3–4 orders of magnitude. We trace this to a geometric property of the ES estimator.

Average curvature, not worst-case. The ES estimator $\hat{g} = (N\sigma)^{-1} \sum_i A_i \varepsilon_i$ is a *random isotropic direction* in $\mathbb{R}^d$: the perturbations ε_i have no preferred direction. Therefore the curvature experienced per ES step is

$$\frac{\mathbb{E}[\hat{g}^\top H \hat{g}]}{\mathbb{E}[\|\hat{g}\|^2]} = \frac{\mathrm{tr}(H)}{d} = \bar{\lambda}, \tag{7}$$

where $\bar{\lambda} = \mathrm{tr}(H)/d$ is the *mean* eigenvalue, not the worst-case λ_1. The scalar bound $L_\sigma = 2/\sigma^2$ corresponds to λ_1; the effective quantity is $\bar{\lambda}$.

Spectral decay. For pretrained LLMs, Hessian eigenvalues follow approximate power-law decay $\lambda_k \propto k^{-\beta}$ with $\beta \approx 2$ [4]. With effective rank r and dimension d,

$$L_\sigma^{\mathrm{true}} \approx \frac{2\bar{\lambda}}{\sigma^2} \approx \frac{2r}{d(\beta-1)\sigma^2}. \tag{8}$$

Numerical verification. At $\sigma = 10^{-3}$, $d = 10^9$, $r = 10^2$, $\beta = 2$: $L_\sigma^{\mathrm{true}} \approx 2 \times 10^{-7} \times 10^6 = 0.2$, giving stability $\eta < 1/0.2 = 5$. This is consistent with empirical $\eta \sim 10^{-3}$–10^{-4}.

Bound	Value	Stability $\eta <$	Status
Scalar $L_\sigma = 2/\sigma^2$	2×10^6	5×10^{-7}	Too conservative
Spectral decay L_σ^{true}	0.2	5	Consistent with practice
Empirical	—	10^{-3}–10^{-4}	Matches spectral bound

Sensitivity to spectral parameters. The values $\beta \approx 2$ and $r \approx 100$ are drawn from Liang et al. [4], which studies LLM fine-tuning landscapes broadly. These have *not* been verified for QWEN2.5/GSM8K specifically; quantitative predictions ($L_\sigma^{\mathrm{true}} \approx 0.2$, stability $\eta < 5$) should be treated as order-of-magnitude estimates until directly measured.

The sensitivity is moderate: $L_\sigma^{\mathrm{true}} \propto r/(\beta-1)$, so:

β	r	L_σ^{true}	Stability $\eta <$	Status
1.5	50	0.2	5	Consistent
2.0	100	0.2	5	Nominal
2.5	200	0.27	3.7	Consistent
2.0	50	0.1	10	More permissive
1.5	200	0.8	1.25	Tighter, still $\gg 10^{-4}$

Across the plausible range $\beta \in [1.5, 2.5]$ and $r \in [50, 200]$, L_σ^{true} varies by roughly $4\times$ and the stability threshold remains many orders of magnitude above empirical $\eta \sim 10^{-3}$–10^{-4}. The qualitative conclusion — that spectral decay resolves the theory-practice gap — is robust to this uncertainty. For precise quantitative predictions on a specific model, the probe (Appendix E) provides direct empirical calibration.

The r/d factor and the stability gap. The spectral correction $r/d \approx 10^{-7}$ resolves the theory-practice stability gap by replacing worst-case curvature with average curvature experienced by isotropic perturbations (Appendix H). The isotropy of ES perturbations in $\mathbb{R}^d$ is the single geometric fact that makes the stability bound empirically meaningful.