CRITIC ERROR AND SCORE GEOMETRY IN TOKEN-LEVEL POLICY GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Token-level credit information need not reduce policy-gradient noise when token scores share parameters. We study this dependence in a deterministic, terminal-reward Token MDP. The advantage sequence is a martingale difference sequence, yielding an exact return-variance identity and a trajectory-gradient decomposition into an oracle token term and zero-mean cross-noise. The cross term has no fixed sign: a valid four-step shared-parameter policy reverses the second-moment ordering. Pathwise orthogonal scores permit an attainable $1/H$ ratio, or $2/(H+1)$ under homogeneous advantages when both estimators access the true value function. A spectral condition extends these conditional comparisons to random score geometry, but is uninformative on our measured neural-policy Grams. For estimated critics, a conditional mean-squared-error identity separates critic bias, its interaction with oracle error, and the full critic-error covariance; perturbation bounds and an exact-value GAE(λ) spectrum complement this analysis. A pretrained GPT-2 prefix-continuation study exhibits a critic-budget-dependent estimator gap at substantial sampling cost. A separate three-seed, token-budget-matched LoRA diagnostic uses conditional-Q weights rather than oracle advantages and does not establish a training advantage. The results identify assumptions needed for an estimator separation and measurements needed before interpreting it as a practical efficiency gain.

1 INTRODUCTION

Reinforcement learning from human feedback (RLHF) is widely used to align large language models (LLMs) with human preferences (Ouyang et al., 2022; Bai et al., 2022). RLHF trains a policy (the LLM) to maximize a reward signal from a learned reward model, subject to a KL-divergence constraint against a reference policy. The standard formulation treats the entire generated response as a single action and assigns a scalar reward at the sequence level—the *bandit* or *trajectory-level* formulation.

Recent work models LLM generation as a *token-level Markov Decision Process* (Token MDP), where each token is an action, the state is the prompt concatenated with all previously generated tokens, and the transition is deterministic (appending the chosen token) (Rafailov et al., 2024). This perspective enables *token-level credit assignment*: attributing the success or failure of a response to individual token decisions, rather than assigning a single reward to the entire trajectory.

PPO with a learned value function (Schulman et al., 2017), VinePPO with Monte Carlo token-level values (Kazemnejad et al., 2025), and RTO with token-level rewards extracted from DPO (Zhong et al., 2025) motivate this comparison. Their estimator, critic, and optimization choices differ; performance reported in those settings does not establish a general ordering against trajectory-level REINFORCE or GRPO (Shao et al., 2024).

Yet the empirical evidence leaves a basic theoretical question open:

> *What estimator structure is specific to token-level credit assignment, and under what conditions does it yield a horizon-dependent reduction?*

Contributions. The analysis has three parts:

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

1. **Why an ordering can fail.** An exact trajectory–token decomposition exposes the sign-indefinite cross term (Proposition 4); a bounded-reward shared-parameter policy reverses the ordering (Section 5.4).
2. **Conditions that permit a separation.** We give attainable orthogonal-score comparisons, a spectral sufficient condition, and an exact-value GAE(λ) spectrum. These supplement the general second-moment upper bounds, which alone do not compare the actual estimators.
3. **What critic error and cost add.** A conditional error identity retains bias and covariance terms, while perturbation bounds make the required critic guarantee explicit. The empirical checks report sampling cost and the limits of the measured geometry; they distinguish oracle advantages from finite-budget conditional-Q weights.

2 Related work

Token MDP formulation. Rafailov et al. (2024) showed that DPO implicitly learns the optimal Q-function of the Token MDP, establishing that the log-ratio of the trained policy to the reference encodes token-level rewards. Zhong et al. (2025) gave a formal advantage of the MDP formulation over the bandit formulation and proposed the RTO two-phase framework. Q$^\sharp$ (Zhou et al., 2025) provided provable bounds for KL-regularized RL in deterministic MDPs.

Credit assignment methods for LLMs. Recent empirical work proposes token-level credit assignment: VinePPO (Kazemnejad et al., 2025) uses Monte Carlo rollouts for token-level value estimation; TEMPO (Tran et al., 2026) builds prefix trees; SPO (Guo et al., 2025) operates at the segment level; GRPO-λ (Parthasarathi et al., 2025) uses eligibility traces; MA-RLHF (Chai et al., 2025) groups tokens into macro-actions. These methods address practical credit assignment through different estimator and optimization choices. Our comparison concerns a single on-policy estimator under explicit score geometry, rather than a general claim that previous token-level methods lack theoretical guarantees.

Variance reduction in policy gradients. The classical analysis of REINFORCE variance (Williams, 1992) uses action-independent baselines without changing the mean; suitably chosen baselines can reduce variance. Greensmith et al. (2004) analyzed variance reduction via baselines and control variates. GAE (Schulman et al., 2016) introduced the bias-variance tradeoff through λ-returns. Baxter & Bartlett (2001) analyzed the G(PO)MDP estimator. We quantify conditional second-moment reductions from token-level credit assignment in the deterministic Token MDP.

Relation to classical PG analysis. The qualitative insight that per-step baselines reduce variance is classical (Baxter & Bartlett, 2001; Greensmith et al., 2004). Our contribution is the exact Token-MDP variance and cross-noise structure, together with conditional factor-H constructions and the GAE(λ) spectrum under explicit geometry. See Appendix G for detailed theorem-level comparisons with concurrent work (Zhou et al., 2025; Kazemnejad et al., 2025; Li et al., 2026).

RLHF theory. Recent theoretical work on RLHF includes analyses of DPO's implicit reward structure (Rafailov et al., 2024), connections between RLHF and imitation learning (Xiao et al., 2025), and the observation that GRPO implicitly performs process-reward-model-style credit assignment when trajectories share prefixes (Sullivan & Koller, 2026).

3 Preliminaries

Definition 1 (Token MDP)**.** *A Token MDP is a tuple $\mathcal{M} = (\mathcal{V}, H, R)$ where:*

- *$\mathcal{V}$ is a finite vocabulary with $|\mathcal{V}| = V$.*
- *The state at step t is $s_t = (x, a_0, \ldots, a_{t-1})$, where x is a **fixed** prompt.*
- *Transitions are **deterministic**: $s_{t+1} = (s_t, a_t)$.*
- *$R : \mathcal{V}^H \to [0, 1]$ is a terminal-only reward function.*

- $H \in \mathbb{N}$ *is the horizon.*

The policy $\pi_\theta(a_t \mid s_t)$ has parameters $\theta \in \mathbb{R}^d$.

For a trajectory $\tau = (s_0, a_0, \ldots, s_{H-1}, a_{H-1}, s_H)$, define

$$V^\pi(s_t) = \mathbb{E}_\pi[R(s_H) \mid s_t], \quad Q^\pi(s_t, a_t) = \mathbb{E}_\pi[R(s_H) \mid s_t, a_t], \tag{1}$$

$$A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t). \tag{2}$$

For an autoregressive language model, this representation loses no information: the factorization $p(a_{0:H-1} \mid x) = \prod_t \pi(a_t \mid x, a_{<t})$ makes the complete prefix a sufficient state, and appending a token is exactly the deterministic transition. Approximation enters through finite context, function approximation, finite rollouts, or rewards depending on hidden external state, rather than through the Token-MDP representation itself.

Remark 1 (Deterministic transition structure)**.** *Since $s_{t+1} = (s_t, a_t)$ is deterministic, $Q^\pi(s_t, a_t) = V^\pi(s_{t+1})$ and therefore $A^\pi(s_t, a_t) = V^\pi(s_{t+1}) - V^\pi(s_t)$.*

The policy optimization objective is $J(\theta) = \mathbb{E}_{\pi_\theta}[R(s_H) \mid s_0 = x]$. We write $\phi_t := \nabla_\theta \log \pi_\theta(a_t \mid s_t)$ for the score function.

Assumption 1 (Pointwise bounded score)**.** *At the parameter θ at which the estimators are compared, $\|\phi_t\| \le G$ for all t, s_t, a_t. A uniform statement over training would require restricting the parameter set.*

Assumption 2 (Bounded reward)**.** $R(s_H) \in [0, 1]$.

4 Policy gradient estimators

We compare two canonical estimators of $\nabla_\theta J(\theta)$.

Trajectory-level REINFORCE. With constant baseline b:

$$\hat{g}_{\mathrm{traj}}(\tau) = (R(s_H) - b) \sum_{t=0}^{H-1} \phi_t. \tag{3}$$

For the comparison below we use the centered-return baseline $b = V^\pi(s_0) = \mathbb{E}_\pi[R(s_H)]$. The second-moment-minimizing scalar baseline can instead be score-norm weighted when the score norm is trajectory dependent.

Token-level actor-critic. With the true advantage function:

$$\hat{g}_{\mathrm{token}}(\tau) = \sum_{t=0}^{H-1} A^\pi(s_t, a_t) \cdot \phi_t. \tag{4}$$

Both estimators are unbiased: $\mathbb{E}[\hat{g}_{\mathrm{traj}}] = \mathbb{E}[\hat{g}_{\mathrm{token}}] = \nabla_\theta J(\theta)$ (by the policy gradient theorem). We compare their *second moment* $\mathbb{E}[\|\hat{g}\|^2]$.

Remark 2 (Comparison scope)**.** *Equations* (3)–(4) *compare single on-policy gradient estimators at the same reference policy. The exact-A^π estimator is an oracle benchmark that isolates the statistical value of token-level credit information; it does not include the cost of learning V^π. The comparison also does not model PPO clipping, importance ratios, repeated optimization epochs, or critic fitting, and therefore is not a ranking of complete training algorithms. Section 6 gives an exact conditional comparison in which both estimators use V^π.*

5 Main results

5.1 Martingale decomposition of the return

The following structural lemma is the foundation of our analysis.

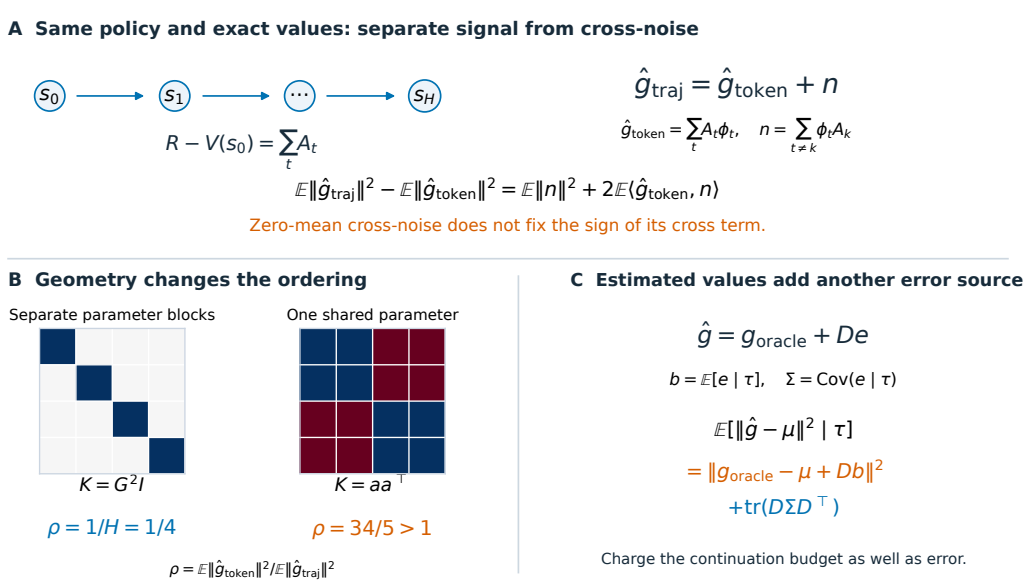


Figure 1: How estimator structure and critic error enter the comparison. (A) The exact decomposition leaves a sign-indefinite cross term. (B) At $H = 4$, pathwise orthogonal scores attain ratio $1/4$; the shared-parameter counterexample at $\eta = 1/64$ gives $34/5$. Matrices show positive (blue), negative (red) and zero (white) entries at $G = 1$. The right matrix is one trajectory; ratios average over trajectories. (C) Conditional critic bias and covariance enter gradient MSE, with $\mu = \nabla J$. These are algebraic results, not neural measurements.

Lemma 2 (Advantage martingale difference sequence)**.** *Fix $s_0 = x$, let $\mathcal{F}_{-1} = \{\varnothing, \Omega\}$ and $\mathcal{F}_t = \sigma(a_0, \ldots, a_t)$, and write $A_t^\pi := A^\pi(s_t, a_t)$. Then, for every $t \in \{0, \ldots, H-1\}$:*

(a) $\{A_t^\pi\}_{t=0}^{H-1}$ is a martingale difference sequence: $\mathbb{E}[A_t^\pi \mid \mathcal{F}_{t-1}] = 0$ almost surely.

(b) The return decomposes pathwise: $R(s_H) = V^\pi(s_0) + \sum_{t=0}^{H-1} A_t^\pi$.

(c) The variance identity holds: $\text{Var}_\pi[R(s_H) \mid s_0] = \sum_{t=0}^{H-1} \mathbb{E}[(A_t^\pi)^2]$.

Proof sketch. (a) follows from the tower property: s_t is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[A_t \mid s_t] = 0$. (b) uses the telescoping sum $A_t = V^\pi(s_{t+1}) - V^\pi(s_t)$. (c) uses the MDS property to eliminate cross-terms: $\mathbb{E}[A_t A_{t'}] = 0$ for $t \neq t'$. Full proof in Appendix I. □

5.2 Second moment comparison

Theorem 3 (General second-moment upper bounds)**.** *Under Assumptions 1–2, with fixed s_0 and $\sigma_R^2 := \text{Var}_\pi[R(s_H) \mid s_0]$:*

(a) $\mathbb{E}\left[\|\hat{g}_{\text{traj}}\|^2\right] \leq H^2 G^2 \sigma_R^2$.

(b) $\mathbb{E}\left[\|\hat{g}_{\text{token}}\|^2\right] \leq H G^2 \sigma_R^2$.

Here part (a) uses the centered-return choice $b = V^\pi(s_0)$ in (3)*.*

Proof sketch. (a) uses $\|\sum_t \phi_t\| \leq HG$ and $\mathbb{E}[(R - V^\pi(s_0))^2] = \sigma_R^2$. (b) bounds $\|\hat{g}_{\text{token}}\|^2 = \mathbf{A}^\top K \mathbf{A} \leq \lambda_{\max}(K) \sum_t A_t^2$ with $\lambda_{\max}(K) \leq HG^2$, then applies the MDS variance identity. Both are upper bounds; no ratio between the actual second moments follows without a lower bound or additional structure. Full proof in Appendix I. □

5.3 STRUCTURAL DECOMPOSITION

A decomposition identifies terms that can create an estimator gap.

Proposition 4 (Trajectory = token + cross-noise)**.** *For the centered-return baseline $b = V^\pi(s_0)$, the trajectory-level estimator decomposes as:*

$$\hat{g}_{\mathrm{traj}} = \underbrace{\sum_{t=0}^{H-1} A_t \phi_t}_{\hat{g}_{\mathrm{token}}} + \underbrace{\sum_{t=0}^{H-1} \sum_{\substack{k=0 \\ k \neq t}}^{H-1} A_k \phi_t}_{\textit{cross-noise } \mathbf{n}} .$$

The cross-noise $\mathbf{n}$ satisfies $\mathbb{E}[\mathbf{n}] = \mathbf{0}$. It is generally nonzero, but the decomposition alone does not order the two second moments because $\mathbf{n}$ may be correlated with $\hat{g}_{\mathrm{token}}$. Under the additional geometry used below, its contribution yields an exact separation.

Proof. Expanding $R - V^\pi(s_0) = \sum_k A_k$ (Lemma 2(b)): $\hat{g}_{\mathrm{traj}} = (\sum_k A_k)(\sum_t \phi_t) = \hat{g}_{\mathrm{token}} + \mathbf{n}$. The cross-noise satisfies $\mathbb{E}[\mathbf{n}] = 0$ by the MDS property and $\mathbb{E}[\phi_t \mid s_t] = 0$. See Appendix I for details. □

5.4 SHARED PARAMETERS CAN REVERSE THE ORDERING

Let $H = 4$, $a_t \in \{-1, 1\}$, and use the same scalar parameter at every step: $\pi_\theta(a_t = 1 \mid s_t) = \mathrm{sigmoid}(2\theta)$. At $\theta = 0$, all 16 trajectories are equiprobable and $\phi_t = a_t$. Define a raw contrast and its normalized terminal reward by

$$U(a) = \mathbf{1}\{a = (1, 1, -1, -1)\} - \mathbf{1}\{a = (1, -1, 1, -1)\},$$
$$R_\eta(a) = \frac{1 + U(a) + \eta \sum_t a_t}{2}, \qquad 0 \leq \eta \leq \tfrac{1}{4}.$$

The two nonzero-U trajectories have total score zero. On all other paths, $|\eta \sum_t a_t| \leq 1$, so $R_\eta \in [0, 1]$ as required by the Token MDP. Enumerating the exact prefix values gives

$$\mathbb{E}[\hat{g}_{\mathrm{traj}}] = \mathbb{E}[\hat{g}_{\mathrm{token}}] = 2\eta, \tag{5}$$
$$\mathbb{E}[\hat{g}^2_{\mathrm{traj}}] = 10\eta^2, \qquad \mathbb{E}[\hat{g}^2_{\mathrm{token}}] = \tfrac{1}{64} + 4\eta^2. \tag{6}$$

At $\eta = 0$, the trajectory estimator is identically zero while the token estimator has second moment $1/64$. At $\eta = 1/64$, their common mean is nonzero, the token/trajectory second-moment ratio is $34/5$, and the centered-variance ratio is $32/3$. Although $\mathbb{E}[K] = I$, pathwise scores are not orthogonal. Thus even mean score orthogonality does not justify a universal token-level ordering. The orthogonal-score construction below establishes a separate attainable regime, rather than a conclusion that holds across parameterizations.

5.5 TIGHTNESS

Theorem 5 (An exact factor-H separation is attainable)**.** *There exists a Token MDP instance and policy parameterization such that:*

$$\frac{\mathbb{E}[\|\hat{g}_{\mathrm{token}}\|^2]}{\mathbb{E}[\|\hat{g}_{\mathrm{traj}}\|^2]} = \frac{1}{H}.$$

Proof. Let the action set be $\{0, 1\}$ and give each step an independent Bernoulli head

$$\pi_{\theta_t}(1) = \mathrm{sigmoid}(2G\theta_t).$$

At $\theta = 0$, $\phi_t = G(2a_t - 1)e_t$, so the scores have conditional mean zero, norm G, and are pathwise orthogonal. Choose any bounded terminal reward with $\sigma_R^2 > 0$. Then:

- Score functions are pathwise orthogonal: $K = G^2 I_H$.
- Trajectory-level: $\|\hat{g}_{\mathrm{traj}}\|^2 = (R - V^\pi(s_0))^2 \|\sum_t G(2a_t - 1)e_t\|^2 = G^2 H (R - V^\pi(s_0))^2$.

- Token-level: $\|\hat{g}_{\text{token}}\|^2 = \|\sum_t A_t G(2a_t - 1)e_t\|^2 = G^2 \sum_t A_t^2$.

Taking expectations and using Lemma 2(c), $\mathbb{E}[\|\hat{g}_{\text{traj}}\|^2] = G^2 H \sigma_R^2$ and $\mathbb{E}[\|\hat{g}_{\text{token}}\|^2] = G^2 \sigma_R^2$. The ratio is exactly $1/H$. □

Remark 3 (MDS suppresses cross-step contributions)**.** *The MDS property (Lemma 2) forces $\mathbb{E}[A_t A_{t'}] = 0$ for $t \neq t'$. If the score Gram matrix is path-independent, this removes the token estimator's off-diagonal terms and permits the calculation in Section 10.2. A trajectory-dependent Gram matrix cannot in general be taken outside these expectations, so no analogous universal ratio follows.*

6 Fair comparison: both estimators access V^π

The bounds in Theorem 3 compare estimators with different information. The following result gives a separate, exact comparison under orthogonal scores and homogeneous advantage variances, with both estimators accessing V^π.

Causally-correct REINFORCE with state-dependent baseline. The standard "reward-to-go" formulation with per-state baseline:

$$\hat{g}_{\text{RV}}(\tau) = \sum_{t=0}^{H-1} \big(R(s_H) - V^\pi(s_t)\big)\phi_t = \sum_{t=0}^{H-1} G_t \cdot \phi_t, \tag{7}$$

where $G_t := \sum_{k=t}^{H-1} A_k^\pi$ is the "advantage-to-go" from step t. This causal trajectory-level estimator uses the same value function $V^\pi(s_t)$ as the token-level oracle.

Theorem 6 (Conditional comparison with equal access to V^π)**.** *Under Assumptions 1–2, with orthogonal score functions ($\langle \phi_t, \phi_{t'} \rangle = 0$ for $t \neq t'$, $\|\phi_t\| = G$) and homogeneous advantages $\mathbb{E}[A_t^2] = \sigma_R^2/H$ for all t:*

(a) (Exact) $\mathbb{E}[\|\hat{g}_{\text{RV}}\|^2] = \frac{H+1}{2} G^2 \sigma_R^2$.

(b) (Exact) $\mathbb{E}[\|\hat{g}_{\text{token}}\|^2] = G^2 \sigma_R^2$.

(c) Under these conditions, the ratio is exactly $\frac{2}{H+1}$.

Proof sketch. The advantage-to-go $G_t = \sum_{k \geq t} A_k$ satisfies $\mathbb{E}[G_t^2] = \sum_{k \geq t} \mathbb{E}[A_k^2]$ by MDS. With orthogonal scores and homogeneous advantages, summing yields $\mathbb{E}[\|\hat{g}_{\text{RV}}\|^2] = \frac{H+1}{2} G^2 \sigma_R^2$, while the token-level achieves $G^2 \sigma_R^2$. Full proof in Appendix I. □

Remark 4 (Source of the gap)**.** *Within the conditions of Theorem 6, the gap is not due to unequal value-function access. REINFORCE-V weights step t by $A_t + \sum_{k>t} A_k$, whereas token-level weighting isolates A_t. This conditional statement does not imply an ordering for arbitrary score geometry.*

7 Optimality within the stated score-weighted class

We show the advantage-based estimator is optimal within a natural class.

Definition 7 (Score-weighted estimator)**.** *A score-weighted estimator is any estimator of the form $\hat{g} = \sum_{t=0}^{H-1} w_t \cdot \phi_t$, where $w_t = w_t(\tau)$ is a scalar that may depend on the full trajectory τ.*

This class includes REINFORCE (all w_t equal), REINFORCE-V ($w_t = G_t$), the token-level estimator ($w_t = A_t$), and GAE ($w_t = \sum_{k=0}^{H-t-1} (\gamma\lambda)^k \delta_{t+k}$).

Definition 8 (Locally unbiased estimator)**.** *A score-weighted estimator $\hat{g} = \sum_t w_t \phi_t$ is* locally unbiased *if $\mathbb{E}[w_t \mid s_t, a_t] = A_t^\pi$ for all t. Equivalently, $w_t = A_t + \eta_t$ with $\mathbb{E}[\eta_t \mid s_t, a_t] = 0$.*

This class includes REINFORCE-V with the exact state value, GAE with the exact value in the present Token MDP, and the token-level advantage estimator. Global unbiasedness $\mathbb{E}[\hat{g}] = \nabla J$ follows from local unbiasedness but not vice versa.

Theorem 9 (Optimality of advantage weighting)**.** *Consider pathwise orthogonal score functions ($\langle \phi_t, \phi_{t'} \rangle = 0$ for $t \neq t'$, $\|\phi_t\| = G > 0$). Among all* locally unbiased *score-weighted estimators with finite second moment (Definition 8), the advantage estimator $w_t^* = A_t$ minimizes $\mathbb{E}[\|\hat{g}\|^2]$, uniquely up to almost-sure equality of its weights.*

Proof sketch. With orthogonal scores, it suffices to minimize $\mathbb{E}[w_t^2]$ per step. Writing $w_t = A_t + \eta_t$ with $\mathbb{E}[\eta_t \mid s_t, a_t] = 0$, orthogonality gives $\mathbb{E}[w_t^2] = \mathbb{E}[A_t^2] + \mathbb{E}[\eta_t^2] \geq \mathbb{E}[A_t^2]$, with equality iff $w_t = A_t$. Full proof in Appendix I. □

Remark 5. *Local unbiasedness requires the weight at step t to recover the advantage conditional on (s_t, a_t). In the exact-value setting, REINFORCE-V ($w_t = G_t$) and GAE(λ) satisfy this condition by the MDS property. Theorem 9 compares estimators only within this locally unbiased class and under pathwise orthogonal scores.*

Group-centered core estimator. Appendix B separates the $(N-1)/N$ self-centering factor from actual GRPO's normalization, clipping, and repeated optimization. The stationary-point corollary in Appendix A is likewise conditional on bounds holding uniformly along the optimization iterates; neither result ranks complete algorithms.

8 Unified analysis: GAE(λ) in the Token MDP

Generalized Advantage Estimation (Schulman et al., 2016) is the standard method for token-level credit assignment in practice. GAE(λ) uses λ-weighted TD errors:

$$\hat{A}_t^{\mathrm{GAE}(\lambda)} = \sum_{l=0}^{H-t-1} \lambda^l \delta_{t+l}, \quad \delta_k = V^\pi(s_{k+1}) - V^\pi(s_k) = A_k^\pi. \tag{8}$$

Here $V^\pi(s_H) = R(s_H)$ already includes the terminal reward; it is not added again to δ_{H-1}. Exact-value GAE interpolates between token-level weighting ($\lambda = 0$) and REINFORCE-V ($\lambda = 1$).

Theorem 10 (GAE(λ) second-moment spectrum)**.** *In the Token MDP with pathwise orthogonal scores of constant norm G, homogeneous advantages, and true V^π:*

*(a) GAE(λ) is **unbiased for all** $\lambda \in [0, 1]$: $\mathbb{E}[\hat{g}_{\mathrm{GAE}(\lambda)}] = \nabla J$.*

(b) The second moment satisfies:

$$\mathbb{E}\big[\|\hat{g}_{\mathrm{GAE}(\lambda)}\|^2\big] = G^2 \sigma_R^2 S(\lambda, H), \quad S(\lambda, H) = \frac{1}{H} \sum_{t=0}^{H-1} \sum_{l=0}^{H-t-1} \lambda^{2l}.$$

For fixed $\lambda < 1$, $S(\lambda, H) \to 1/(1-\lambda^2)$ as $H \to \infty$; the approximation is not uniform near $\lambda = 1$.

(c) Boundary values: $S(0, H) = 1$ and $S(1, H) = (H+1)/2$. The finite-H polynomial is defined throughout $[0, 1]$.

Proof sketch. (a) Unbiasedness follows because future advantages A_{t+l} ($l \geq 1$) are mean-zero conditional on $\mathcal{F}_t$ by MDS, so only the $l = 0$ term survives. (b) With orthogonal scores and MDS, $\mathbb{E}[(\hat{A}_t)^2] = c \cdot (1 - \lambda^{2(H-t)})/(1-\lambda^2)$ with $c = \sigma_R^2/H$. Summing yields $S(\lambda, H)$. (c) Boundary evaluation. Full proof in Appendix I. □

Theorem 11 (Estimated-critic perturbation for GAE)**.** *Under the conditions of Theorem 10, compare exact-value and estimated-value GAE with $\|\hat{V} - V^\pi\|_\infty \leq \epsilon_V$. For every $\lambda \in [0, 1]$,*

$$\|\hat{g}_{\lambda, \hat{V}} - \hat{g}_{\lambda, V}\| \leq 2G\sqrt{H}\epsilon_V,$$

and therefore

$$\sqrt{\mathbb{E}\|\hat{g}_{\lambda,\hat{V}} - \nabla J\|^2} \le G\sigma_R\sqrt{S(\lambda, H)} + 2G\sqrt{H}\epsilon_V.$$

This is a worst-case perturbation bound, not a universal λ-selection rule.

Proof sketch. Writing $e_k = \hat{V}(s_k) - V^\pi(s_k)$, summation by parts bounds each weight error by $2\epsilon_V$ (Appendix D.1). Orthogonal scores give the first display; Minkowski's inequality gives the second. Without orthogonal scores, replace $\sqrt{H}$ by H. □

9 BIAS-VARIANCE TRADEOFF WITH ESTIMATED ADVANTAGE

In practice, V^π is unknown and must be estimated by a critic $\hat{V}$. The estimated advantage is $\hat{A}_t = \hat{V}(s_{t+1}) - \hat{V}(s_t)$.

Theorem 12 (Bias-variance tradeoff)**.** *Suppose $\|\hat{V} - V^\pi\|_\infty \le \epsilon_V$. Then the estimated token-level estimator $\hat{g}_{\text{est}} = \sum_t \hat{A}_t\phi_t$ satisfies:*

(a) ***Advantage error:*** *$|\hat{A}_t - A^\pi_t| \le 2\epsilon_V$ for all t.*

(b) ***Bias:*** *$\|\mathbb{E}[\hat{g}_{\text{est}}] - \nabla J\| \le 2HG\epsilon_V$.*

(c) ***MSE:*** *$\sqrt{\mathbb{E}[\|\hat{g}_{\text{est}} - \nabla J\|^2]} \le \sqrt{H}\, G\sigma_R + 2HG\epsilon_V$, and hence*

$$\mathbb{E}[\|\hat{g}_{\text{est}} - \nabla J\|^2] \le 2HG^2\sigma_R^2 + 8H^2G^2\epsilon_V^2.$$

Proof sketch. (a) follows from the triangle inequality on $\hat{V} - V^\pi$. For (b), sum the score-weighted pointwise errors. For (c), write $\hat{g}_{\text{est}} - \nabla J = (\hat{g}_{\text{token}} - \nabla J) + (\hat{g}_{\text{est}} - \hat{g}_{\text{token}})$ and apply Minkowski's inequality. Full proof is in Appendix I. Comparing this upper bound with a trajectory-level upper bound would not establish outperformance; such a claim additionally requires a lower bound for the trajectory estimator. □

Remark 6 (Oracle scope)**.** *The exact-A^π analysis measures the statistical value of token-level credit information. It does not include the samples or computation needed to learn V^π; Theorem 12 quantifies estimator error conditional on a supplied ℓ_∞ critic guarantee.*

Dense rewards. Appendix C states the KL-reward extension with consistent terminal accounting. It concerns a fixed-policy estimator, not convergence of a full KL-regularized training algorithm.

10 EMPIRICAL CHECKS AND FINITE-BUDGET TRAINING

The exact orthogonal construction and earlier GPT-2 critic-budget sweep are retained in Appendix L. A frozen short-answer GSM8K diagnostic trains Qwen2.5-1.5B-Instruct LoRA with terminal REINFORCE or one/four independent continuations after each nonterminal token. All sampled continuation tokens count toward a common 65,536-token training target.

Table 1: Three-seed finite-budget training. Accuracies are percentages on 64 fixed held-out questions; parentheses give seed standard deviations, not confidence intervals. Tokens include main trajectories and conditional continuations.

Estimator	Initial	Final (SD)	Updates	Tokens
terminal	3.12	3.65 (0.90)	3100.0	65538
mc1	3.12	5.21 (0.90)	892.0	65564
mc4	3.12	3.65 (0.90)	290.3	66092

The full batch crossing the target is retained, so realized token budgets are close but not exactly equal. Evaluation tokens, unpadded prefix-token counts, and wall time are also archived; equal generation tokens are not equal FLOPs. The accuracy changes are small relative to the 1/64 test-set

resolution. These estimators use terminal or estimated conditional Q weights, without a value baseline. They target the same truncated policy objective but do not directly test the exact-advantage or GAE theorems. A 32-token direct-answer horizon, one small model, and 64 questions do not establish long-reasoning performance or an end-to-end advantage. Appendix M reports all seeds, costs, and fixed-policy gradient diagnostics.

Geometry sensitivity. Full-adapter token Gram measurements make the sufficient spectral ratio bound practically uninformative; Appendix H gives the conditional theorem, and the final audit appendix reports the empirical limitation.

10.1 Critic error, bias, and conditional second moments

Fix a main trajectory τ and write $\widehat{g} = g_{\mathrm{oracle}} + De$, where e collects prefix value-estimation errors and D has columns $-\phi_0, \phi_0 - \phi_1, \ldots, \phi_{H-2} - \phi_{H-1}$ when the terminal value is exact. Let $b = \mathbb{E}[e \mid \tau]$, $\Sigma = \mathrm{Cov}(e \mid \tau)$, and let μ denote the true policy gradient. Direct expansion gives

$$\mathbb{E}[\|\widehat{g} - \mu\|^2 \mid \tau] = \|g_{\mathrm{oracle}} - \mu + Db\|^2 + \mathrm{tr}(D\Sigma D^\top).$$

This identity includes oracle error, its cross term with critic bias, squared bias, and all off-diagonal critic-error covariances. Only conditional unbiasedness and independent prefix estimation justify the simpler variance-weighted adjacent-score expression. A learned critic, shared rollouts, or outcome-dependent reuse of pilot samples need not meet those conditions. A sample covariance alone is not an estimate of total policy-gradient MSE. This algebraic decomposition supplies a framework for future mechanism validation; it does not turn the existing near-singular-Gram measurement into a useful bound or establish an empirical efficiency advantage.

10.2 Why an average Gram matrix is insufficient

The path-independent Gram assumption is more restrictive than a generic correlation condition. Under the score martingale-difference property and bounded scores, for $i < j$,

$$\mathbb{E}[K_{ij}] = \mathbb{E}[\langle \phi_i, \mathbb{E}[\phi_j \mid \mathcal{F}_{j-1}]\rangle] = 0.$$

Consequently, if K is deterministic across paths, all its off-diagonal entries must vanish. A path-independent Gram matrix with nonzero off-diagonal entries is therefore incompatible with valid policy scores. Conditioning on a random K does not fix this problem, because it can destroy the conditional mean-zero property. Appendix H gives a pathwise spectral condition instead.

For arbitrary random K, let $D = R - V^\pi(s_0)$ and $S = \mathbf{1}^\top K \mathbf{1}$. The exact identities are

$$\mathbb{E}\|\hat{g}_{\mathrm{token}}\|^2 = \sum_i \mathbb{E}[K_{ii}]\mathbb{E}[A_i^2] + C_T, \tag{9}$$

$$C_T = \sum_i \mathrm{Cov}(K_{ii}, A_i^2) + 2\sum_{i<j} \mathrm{Cov}(K_{ij}, A_i A_j), \tag{10}$$

$$\mathbb{E}\|\hat{g}_{\mathrm{traj}}\|^2 = \sigma_R^2 \sum_i \mathbb{E}[K_{ii}] + \mathrm{Cov}(D^2, S). \tag{11}$$

Here $\mathbb{E}[A_i A_j] = 0$ and $\mathbb{E}[K_{ij}] = 0$ for $i \neq j$, but neither fact removes the reward–geometry covariance terms. In particular, an average Gram matrix proportional to the identity does not imply an estimator ordering.

11 Discussion and conclusion

A centered trajectory estimator equals the oracle token estimator plus zero-mean cross-noise, whose covariance with the token term can improve or worsen the second moment. The shared-parameter counterexample and orthogonal construction exhibit both directions under valid bounded-reward policies. For a supplied critic, conditional bias and error covariance determine the additional MSE; continuation costs must then be charged separately. The GPT-2 and short-answer LoRA checks do not establish an end-to-end efficiency gain, and the measured neural Grams do not validate a useful near-isotropy premise. These findings make the geometry and critic assumptions explicit targets for future empirical tests. The GRPO result concerns only its group-centered core; further limits appear in Appendix F.

Reproducibility Statement

The manuscript states the conditioning, score-geometry assumptions, and estimated-critic error terms needed for each comparison; proofs and extensions are supplied in the appendix. Retained experiment code and analysis records cover the synthetic checks, critic-budget study, and fixed-policy diagnostics. Algebraic counterexamples and constructed diagrams are explicitly distinguished from measured neural-model results. These checks do not replace the additional training comparisons needed to establish a general optimization advantage.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Yekun Chai, Haoran Sun, Huang Fang, Shuohuan Wang, Yu Sun, and Hua Wu. MA-RLHF: Reinforcement learning from human feedback with macro actions. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/429d69979c22b06d6baa65caf3ab1e10-Abstract-Conference.html.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in RL for large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a6536243037d1e32c20de85137d478da-Abstract-Conference.html.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2410.01679.

Yingru Li, Jiawei Xu, Ziniu Li, Jiacai Liu, Wei Liu, Yuxuan Tong, Longtao Zheng, Zhenghai Xue, Yaxiang Zhang, Tianle Cai, Ge Zhang, Qian Liu, and Baoxiang Wang. The optimal token baseline: Variance reduction for long-horizon LLM-RL. *arXiv preprint arXiv:2602.07078*, 2026.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022. arXiv:2203.02155.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. GRPO-λ: Credit assignment improves LLM reasoning. *arXiv preprint arXiv:2510.00194*, 2025.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.12358.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016. arXiv:1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Michael Sullivan and Alexander Koller. GRPO is secretly a process reward model. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61734.

Hieu Tran, Zonghai Yao, and Hong Yu. Exploiting tree structure for credit assignment in reinforcement learning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10795–10810, 2026. doi: 10.18653/v1/2026.findings-acl.524. URL https://aclanthology.org/2026.findings-acl.524/.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

Teng Xiao, Yige Yuan, Mingxiao Li, Zhengyu Chen, and Vasant G Honavar. On a connection between imitation learning and RLHF. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/acf4a08f67724e9d2de34099f57a9c25-Abstract-Conference.html.

Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. DPO meets PPO: Reinforced token optimization for RLHF. In *International Conference on Machine Learning (ICML)*, 2025.

Jin Peng Zhou, Kaiwen Wang, Jonathan Chang, Zhaolin Gao, Nathan Kallus, Kilian Q. Weinberger, Kianté Brantley, and Wen Sun. $q^\sharp$: Provably optimal distributional RL for LLM post-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.20548.

A STATIONARY-POINT COMPLEXITY UPPER BOUNDS

Corollary 13 (Sufficient trajectories for a stationary point)**.** *Assume $J(\theta)$ is L-smooth and the score bound holds uniformly along the optimization iterates. Let $\bar{\sigma}_R^2$ bound the return variance at every iterate, so the corresponding second-moment bounds also hold uniformly. Define $\Delta_J = J(\pi^*) - J(\pi_0)$. To achieve $\mathbb{E}[\|\nabla J(\theta_{\mathrm{out}})\|^2] \le \epsilon$:*

(a) $N_{\mathrm{traj}} = O\Big(\frac{L\Delta_J H^2 G^2 \bar{\sigma}_R^2}{\epsilon^2}\Big)$ *trajectories are sufficient for the trajectory-level estimator.*

(b) $N_{\mathrm{token}} = O\Big(\frac{L\Delta_J H G^2 \bar{\sigma}_R^2}{\epsilon^2}\Big)$ *trajectories are sufficient for the token-level estimator with true advantages.*

Proof. Follows from standard SGD convergence for L-smooth non-convex functions (Ghadimi & Lan, 2013), substituting the two second-moment upper bounds from Theorem 3. These are sufficient bounds for finding an approximate stationary point, not a universal ratio of necessary sample sizes. See Appendix I. □

B ANALYSIS OF GRPO

GRPO (Shao et al., 2024) uses policy ratios, clipping, and group normalization. To isolate group centering, we study only the following on-policy core score estimator with N responses per prompt:

$$\hat{g}_{\mathrm{GRPO}} = \frac{1}{N}\sum_{i=1}^{N}(R^{(i)} - \bar{R})\sum_{t=0}^{H-1}\phi_t^{(i)}, \quad \bar{R} = \frac{1}{N}\sum_{j=1}^{N}R^{(j)}. \tag{12}$$

Theorem 14 (Group-centered core estimator)**.** *Under Assumptions 1–2, with N i.i.d. trajectories per prompt:*

(a) Its expectation is $\mathbb{E}[\hat{g}_{\mathrm{GRPO}}] = \frac{N-1}{N}\nabla J$ because each trajectory also enters the group mean.

(b) Its second moment obeys the conservative bound

$$\mathbb{E}[\|\hat{g}_{\mathrm{GRPO}}\|^2] \le \frac{N-1}{N}H^2G^2\sigma_R^2.$$

Proof sketch. Part (a) follows by symmetry and the zero-mean score identity. For part (b), Jensen's inequality, $\|\sum_t \phi_t^{(i)}\| \le HG$, and $\mathbb{E}[N^{-1}\sum_i (R^{(i)} - \bar{R})^2] = (N-1)\sigma_R^2/N$ give the bound. Full proof is in Appendix I. □

Remark 7 (Scope)**.** *Equation above is the group-centered score estimator underlying GRPO, not the full normalized and clipped update. Its one-step second moment cannot by itself rank complete training algorithms or account for multiple optimization epochs.*

Remark 8 (Prefix overlap)**.** *In practice, shared prefixes can provide partial token-level credit (Sullivan & Koller, 2026). The theorem describes only the i.i.d. core-estimator setting.*

C KL-REGULARIZED EXTENSION

In RLHF, $J_{\mathrm{KL}}(\theta) = \mathbb{E}_{\pi_\theta}[R(s_H) - \beta\sum_t \log(\pi_\theta(a_t|s_t)/\pi_{\mathrm{ref}}(a_t|s_t))]$. Set $r_t = -\beta\log(\pi_\theta/\pi_{\mathrm{ref}})(a_t \mid s_t) + \mathbf{1}_{t=H-1}R(s_H)$ and $V^{\mathrm{KL}}(s_H) = 0$; the terminal reward is included once.

Theorem 15 (KL-regularized extension)**.** *For a fixed policy, define $A_t^{\mathrm{KL}} = r_t + V^{\mathrm{KL}}(s_{t+1}) - V^{\mathrm{KL}}(s_t)$. The MDS identity, pathwise return decomposition, general upper bounds, and critic perturbation argument carry over with σ_R^2 replaced by $\mathrm{Var}_\pi[\sum_t r_t \mid s_0]$. Exact ratios retain their score-geometry and homogeneity conditions.*

The identity follows from conditional expectation and telescoping, with the fixed-policy dense reward treated as part of the return. Policy-dependent regularization must be differentiated consistently in any full optimization analysis; the estimator identities alone do not provide such a convergence theorem.

D COMPLETED CRITIC-BUDGET AUDIT

D.1 FINITE-HORIZON CRITIC PERTURBATION

For $m = H - t$ and $e_k = \hat{V}(s_k) - V^\pi(s_k)$, the GAE weight error is

$$\sum_{l=0}^{m-1} \lambda^l (e_{t+l+1} - e_{t+l}) = -e_t + (1-\lambda) \sum_{j=1}^{m-1} \lambda^{j-1} e_{t+j} + \lambda^{m-1} e_H.$$

For $0 \le \lambda \le 1$, the nonnegative coefficients after $-e_t$ sum to one, including the endpoint cases. Therefore the absolute weight error is at most $2\epsilon_V$, with no divergence as $\lambda \to 1$. Pathwise orthogonal scores of norm G give a squared perturbation at most $4HG^2\epsilon_V^2$. This uniform upper bound does not identify an optimal λ or guarantee that estimated-critic bias attains it.

D.2 EMPIRICAL MOMENTS AND COSTS

For unbiased estimators of the same gradient g, the population identity is $\mathbb{E}\|\hat{g}\|^2 = \operatorname{tr}\operatorname{Cov}(\hat{g}) + \|g\|^2$. A second-moment ratio is therefore not a centered-variance ratio: the common $\|g\|^2$ must first be subtracted. For estimated critics, equality of the estimator means additionally requires unbiasedness; it cannot be assumed from this identity. Neither ratio includes critic cost or proves a ratio of end-to-end sample complexities.

For each method we retain the empirical second moment $S^{-1}\sum_i \|g_i\|^2$ and squared empirical mean $\|\bar{g}\|^2$ separately. The centered sample covariance trace is $\frac{S}{S-1}(S^{-1}\sum_i \|g_i\|^2 - \|\bar{g}\|^2)$. The following ratios average five prompt-specific ratios; they are not a ratio of pooled moments. All values are conditional on a single root-value estimate per prompt/horizon. Variation from redrawing that root estimate is not measured. The reward is synthetic, the prompt set is fixed, and no policy is trained. Monte Carlo continuation draws need not be nested across M, although main trajectories and root-value estimates are held fixed. Telescoping identities and analytic token counts agree with the observed continuations. Prefill token counts and elapsed times are retained separately; timings include three gradient-statistic backward passes per main trajectory and are not deployment-throughput measurements. The earlier random-initialization proxy experiment and earlier prefix-check summary are retained in the revision archive, not pooled with the new results.

Table 2: Conditional sample-covariance-trace ratios for $M = 128$ and generated token costs. "Traj." and "RTG" are relative to the token estimator.

H	Traj./token	RTG/token	Generated/main tokens
4	3.30	2.14	209
8	7.32	3.81	465
16	11.48	6.33	977
32	22.42	10.73	2001

E PROOFS OF AUXILIARY RESULTS

E.1 PROOF THAT $\mathbb{E}[\phi_t \mid s_t] = 0$

For any state s_t:

$$\mathbb{E}_{a_t\sim\pi}[\nabla_\theta \log \pi(a_t|s_t) \mid s_t] = \sum_a \pi(a|s_t) \frac{\nabla_\theta \pi(a|s_t)}{\pi(a|s_t)} = \nabla_\theta \sum_a \pi(a|s_t) = \nabla_\theta 1 = 0.$$

This identity is used in the proof of Proposition 4 to show $\mathbb{E}[\mathbf{n}] = 0$.

F EXTENDED DISCUSSION

Group centering and GRPO. The estimator in Section B isolates group centering. Actual GRPO also uses normalization, policy ratios, clipping, and repeated optimization, so the core-estimator

bound should not be read as an end-to-end ranking. When sampled trajectories share prefixes, the update may also contain partial token-level information (Sullivan & Koller, 2026).

Gap between theory and practice. Our analysis compares single on-policy gradient estimators at the same reference policy. Practical PPO-style methods add clipped surrogate objectives (Schulman et al., 2017), importance ratios, multiple optimization epochs, and entropy bonuses. These mechanisms and the cost of learning a critic lie outside the comparison. The results therefore isolate an estimator-level mechanism rather than rank complete training algorithms.

G Extended related work

Theorem-level comparison with concurrent work. Our results are complementary to several recent lines of work. Q$^\sharp$ (Zhou et al., 2025) proves convergence bounds for KL-regularized RL in deterministic MDPs using a value-based approach; here we study policy-gradient estimator structure. VinePPO (Kazemnejad et al., 2025) empirically studies Monte Carlo token-level value estimation, whereas Theorem 12 gives an estimator-error bound conditional on an ℓ_∞ value guarantee. The Optimal Token Baseline (Li et al., 2026) derives variance-minimizing baseline weights; Theorem 9 instead concerns locally unbiased score weights under orthogonal scores. Proposition 4 gives an exact algebraic relation between the trajectory and token estimators, but does not by itself order their second moments.

H Perturbation analysis for correlated scores

Theorem 16 (Pathwise spectral robustness for random Gram matrices)**.** *Assume $\sigma_R^2 > 0$ and that, on every trajectory, $\|K - G^2 I\|_{\mathrm{op}} \le \epsilon G^2$ for fixed $G > 0$ and $0 \le \epsilon < 1$. Then*

$$\frac{1-\epsilon}{H(1+\epsilon)} \le \frac{\mathbb{E}\|\hat{g}_{\mathrm{token}}\|^2}{\mathbb{E}\|\hat{g}_{\mathrm{traj}}\|^2} \le \frac{1+\epsilon}{H(1-\epsilon)}.$$

The condition permits path-dependent off-diagonal entries and does not require homogeneous advantage variances. A sufficient condition is $K_{ii} = G^2$ and $|K_{ij}| \le \rho G^2$ on every path, with $\epsilon = (H-1)\rho < 1$.

Proof. The spectral condition gives $(1-\epsilon)G^2\|x\|^2 \le x^\top K x \le (1+\epsilon)G^2\|x\|^2$ pathwise. Apply it first to $x = A$ and use $\sum_i \mathbb{E}[A_i^2] = \sigma_R^2$, and then to $x = \mathbf{1}$, multiplying by D^2 before taking expectation. The positive denominator lies between $H(1-\epsilon)G^2\sigma_R^2$ and $H(1+\epsilon)G^2\sigma_R^2$. Dividing the corresponding bounds proves the result. The entrywise sufficient condition follows from the maximum absolute row sum of $K - G^2 I$. □

Exact verification and scope. The artifact enumerates the shared-parameter construction with rational arithmetic, including all prefix values. A separate 16-trajectory check verifies the $[0,1]$-normalized reward in Section 5.4, both estimator means and second moments, and the stated variance ratio. It also checks 128 finite-policy instances with $H \in \{2,3,4,6\}$ under the spectral bound. These use valid policies with logits $2\theta^\top v_t$, where $v_i^\top v_j = \rho$ for $i \ne j$ and $\|v_i\| = 1$, so $K_{ij} = a_i a_j \rho$ is random. These are algebraic checks, not neural-training results or evidence that a transformer satisfies a uniform spectral bound. The bound itself follows from a standard spectral sandwich; our claim is the explicit condition and counterexample for this estimator comparison, not a new general spectral inequality.

I Deferred proofs

I.1 Proof of Lemma 2 (Advantage MDS)

Proof. **(a)** Since s_t is a deterministic function of $(s_0, a_0, \ldots, a_{t-1})$, it is $\mathcal{F}_{t-1}$-measurable. By the tower property:

$$\mathbb{E}[A_t^\pi \mid \mathcal{F}_{t-1}] = \mathbb{E}\big[\mathbb{E}[A_t^\pi \mid s_t] \,\big|\, \mathcal{F}_{t-1}\big] = \mathbb{E}[0 \mid \mathcal{F}_{t-1}] = 0,$$

where $\mathbb{E}[A_t^\pi \mid s_t] = \sum_a \pi(a \mid s_t)[Q^\pi(s_t, a) - V^\pi(s_t)] = 0$.

(b) By the deterministic transition property $A_t^\pi = V^\pi(s_{t+1}) - V^\pi(s_t)$. Summing:

$$\sum_{t=0}^{H-1} A_t^\pi = V^\pi(s_H) - V^\pi(s_0) = R(s_H) - V^\pi(s_0),$$

using $V^\pi(s_H) = R(s_H)$ at the terminal state.

(c) Since $V^\pi(s_0)$ is constant (fixed s_0), $\mathrm{Var}[R(s_H)] = \mathrm{Var}[\sum_t A_t^\pi]$. For the cross terms, fix $t < t'$. Then A_t^π is $\mathcal{F}_t$-measurable, and by the MDS property:

$$\mathbb{E}[A_t^\pi A_{t'}^\pi] = \mathbb{E}\big[A_t^\pi \mathbb{E}[A_{t'}^\pi \mid \mathcal{F}_{t'-1}]\big] = 0.$$

Therefore $\mathrm{Var}[\sum_t A_t^\pi] = \sum_t \mathbb{E}[(A_t^\pi)^2]$. □

I.2 Proof of Theorem 3 (Second moment comparison)

Proof. **(a)** With the centered-return baseline $b = V^\pi(s_0)$:

$$\|\hat{g}_{\mathrm{traj}}\|^2 = (R - V^\pi(s_0))^2 \left\| \sum_t \phi_t \right\|^2 \le (R - V^\pi(s_0))^2 \cdot H^2 G^2,$$

where we used $\|\sum_t \phi_t\| \le \sum_t \|\phi_t\| \le HG$. Taking expectation yields the bound.

(b) Define the Gram matrix $K \in \mathbb{R}^{H \times H}$ with $K_{ij} = \langle \phi_i, \phi_j \rangle$. Then:

$$\|\hat{g}_{\mathrm{token}}\|^2 = \left\| \sum_t A_t \phi_t \right\|^2 = \mathbf{A}^\top K \mathbf{A} \le \lambda_{\max}(K) \sum_t A_t^2,$$

where $\mathbf{A} = (A_0, \ldots, A_{H-1})^\top$. Since $\lambda_{\max}(K) \le \mathrm{tr}(K) = \sum_t \|\phi_t\|^2 \le HG^2$:

$$\mathbb{E}[\|\hat{g}_{\mathrm{token}}\|^2] \le HG^2 \sum_t \mathbb{E}[A_t^2] = HG^2 \sigma_R^2,$$

where the last equality uses Lemma 2(c).

□

I.3 Proof of Proposition 4 (Trajectory = token + cross-noise)

Proof. Using $R - V^\pi(s_0) = \sum_k A_k$ (Lemma 2(b)):

$$\hat{g}_{\mathrm{traj}} = \left(\sum_k A_k \right) \left(\sum_t \phi_t \right) = \sum_t A_t \phi_t + \sum_{t \ne k} A_k \phi_t = \hat{g}_{\mathrm{token}} + \mathbf{n}.$$

For unbiasedness of the cross-noise: $\mathbb{E}[A_k \phi_t]$ for $k > t$ vanishes by the MDS property (tower property with $k > t$ giving $\mathbb{E}[A_k \mid \mathcal{F}_{k-1}] = 0$). For $k < t$, $\mathbb{E}[A_k \phi_t] = \mathbb{E}[A_k \mathbb{E}[\phi_t \mid \mathcal{F}_{t-1}, s_t]]$; note that $\mathbb{E}[\phi_t \mid s_t] = \nabla_\theta \sum_a \pi(a|s_t) = 0$, so these terms also vanish. Thus $\mathbb{E}[\mathbf{n}] = 0$. □

I.4 Proof of Corollary 13 (Sample complexity)

Proof. Standard SGD convergence for L-smooth non-convex functions (Ghadimi & Lan, 2013): with an unbiased gradient estimator $\hat{g}$ satisfying $\mathbb{E}[\|\hat{g}\|^2] \le M$, step size $\eta = \sqrt{\Delta_J / (LMN)}$, after N steps:

$$\min_{n \le N} \mathbb{E}[\|\nabla J(\theta_n)\|^2] \le 2\sqrt{\frac{L \Delta_J M}{N}}.$$

Setting the RHS to ϵ gives $N \ge 4L\Delta_J M / \epsilon^2$. Substituting $M_{\mathrm{traj}} = H^2 G^2 \bar{\sigma}_R^2$ and $M_{\mathrm{token}} = HG^2 \bar{\sigma}_R^2$ from Theorem 3 yields the result. □

I.5 Proof of Theorem 6 (Fair comparison)

Proof. The advantage-to-go satisfies $G_t = \sum_{k=t}^{H-1} A_k$. By the MDS property (Lemma 2), for $k < k'$, $\mathbb{E}[A_k A_{k'}] = 0$, so:

$$\mathbb{E}[G_t^2] = \sum_{k=t}^{H-1} \mathbb{E}[A_k^2].$$

With orthogonal scores ($K = G^2 I$):

$$\mathbb{E}[\|\hat{g}_{\mathrm{RV}}\|^2] = G^2 \sum_{t=0}^{H-1} \mathbb{E}[G_t^2] = G^2 \sum_{k=0}^{H-1} (k+1)\mathbb{E}[A_k^2].$$

With homogeneous advantages $\mathbb{E}[A_k^2] = c = \sigma_R^2/H$:

$$\mathbb{E}[\|\hat{g}_{\mathrm{RV}}\|^2] = G^2 c \cdot \frac{H(H+1)}{2} = \frac{H+1}{2} G^2 \sigma_R^2.$$

The token-level estimator achieves $\mathbb{E}[\|\hat{g}_{\mathrm{token}}\|^2] = G^2 \sum_t \mathbb{E}[A_t^2] = G^2 \sigma_R^2$. The ratio is $2/(H+1)$. □

I.6 Proof of Theorem 9 (Optimality of advantage weighting)

Proof. With orthogonal scores, $\mathbb{E}[\|\hat{g}\|^2] = G^2 \sum_t \mathbb{E}[w_t^2]$, so it suffices to minimize $\mathbb{E}[w_t^2]$ for each t separately.

Write $w_t = A_t + \eta_t$ with $\mathbb{E}[\eta_t \mid s_t, a_t] = 0$ (local unbiasedness). Since A_t is $\sigma(s_t, a_t)$-measurable:

$$\mathbb{E}[A_t \eta_t] = \mathbb{E}\big[A_t \, \mathbb{E}[\eta_t \mid s_t, a_t]\big] = \mathbb{E}[A_t \cdot 0] = 0.$$

Therefore:

$$\mathbb{E}[w_t^2] = \mathbb{E}[A_t^2] + 2\underbrace{\mathbb{E}[A_t \eta_t]}_{=0} + \mathbb{E}[\eta_t^2] = \mathbb{E}[A_t^2] + \mathbb{E}[\eta_t^2] \geq \mathbb{E}[A_t^2],$$

with equality iff $\eta_t = 0$ a.s., i.e., $w_t = A_t$. □

I.7 Proof of Theorem 14 (GRPO second moment)

Proof. Write $Z_i = R^{(i)} - \mathbb{E}[R]$, $\Phi_i = \sum_t \phi_t^{(i)}$, and let bars denote sample means. Then

$$\hat{g}_{\mathrm{GRPO}} = \frac{1}{N} \sum_i Z_i \Phi_i - \bar{Z}\,\bar{\Phi}.$$

The first term has expectation ∇J. Independence and $\mathbb{E}[Z_i] = \mathbb{E}[\Phi_i] = 0$ give $\mathbb{E}[\bar{Z}\,\bar{\Phi}] = \nabla J/N$, proving part (a).

For part (b), Jensen's inequality and $\|\Phi_i\| \leq HG$ give

$$\|\hat{g}_{\mathrm{GRPO}}\|^2 \leq \frac{1}{N} \sum_i (Z_i - \bar{Z})^2 \|\Phi_i\|^2 \leq H^2 G^2 \frac{1}{N} \sum_i (Z_i - \bar{Z})^2.$$

Taking expectations and using the expected sample variance $(N-1)\sigma_R^2/N$ yields the claim. □

I.8 Proof of Theorem 10 (GAE(λ) spectrum)

Proof. **(a)** We need $\mathbb{E}[\hat{A}_t^{\mathrm{GAE}(\lambda)} \phi_t] = \mathbb{E}[A_t \phi_t]$ for all t. Write $\hat{A}_t^{\mathrm{GAE}(\lambda)} = A_t + \sum_{l=1}^{H-t-1} \lambda^l A_{t+l}$. For $l \geq 1$:

$$\mathbb{E}[A_{t+l}\phi_t] = \mathbb{E}\big[\phi_t \, \mathbb{E}[A_{t+l} \mid \mathcal{F}_t]\big] = \mathbb{E}[\phi_t \cdot 0] = 0,$$

using the tower property and the MDS property $\mathbb{E}[A_{t+l} \mid \mathcal{F}_{t+l-1}] = 0$. So only the $l = 0$ term survives: $\mathbb{E}[\hat{A}_t^{\mathrm{GAE}(\lambda)} \phi_t] = \mathbb{E}[A_t \phi_t]$.

(b) With orthogonal scores: $\mathbb{E}[\|\hat{g}\|^2] = G^2 \sum_t \mathbb{E}[(\hat{A}_t^{\mathrm{GAE}(\lambda)})^2]$. Since $\hat{A}_t = \sum_l \lambda^l A_{t+l}$ and $\mathbb{E}[A_{t+l}A_{t+l'}] = \delta_{ll'}\mathbb{E}[A_{t+l}^2]$ (by MDS):

$$\mathbb{E}[(\hat{A}_t)^2] = \sum_{l=0}^{H-t-1} \lambda^{2l}\mathbb{E}[A_{t+l}^2] = c \cdot \frac{1-\lambda^{2(H-t)}}{1-\lambda^2},$$

with $\mathbb{E}[A_k^2] = c = \sigma_R^2/H$. Summing over t: $\sum_t \mathbb{E}[(\hat{A}_t)^2] = c\sum_t \frac{1-\lambda^{2(H-t)}}{1-\lambda^2} = \sigma_R^2 \cdot S(\lambda, H)$.

(c) At $\lambda = 0$: $S = 1$. At $\lambda = 1$: $S = \frac{1}{H}\sum_t (H-t) = (H+1)/2$. □

I.9 Proof of Theorem 12 (Bias-variance tradeoff)

Proof. **(a)** $|\hat{A}_t - A_t| = |(\hat{V}(s_{t+1}) - V(s_{t+1})) - (\hat{V}(s_t) - V(s_t))| \le 2\epsilon_V$.

(b) $\mathbb{E}[\hat{g}_{\mathrm{est}}] - \nabla J = \sum_t \mathbb{E}[(\hat{A}_t - A_t)\phi_t]$. By triangle inequality, $\|\mathbb{E}[\hat{g}_{\mathrm{est}}] - \nabla J\| \le \sum_t \mathbb{E}[|\hat{A}_t - A_t|] \cdot G \le 2HG\epsilon_V$.

(c) The pointwise error satisfies

$$\|\hat{g}_{\mathrm{est}} - \hat{g}_{\mathrm{token}}\| \le \sum_t |\hat{A}_t - A_t| \, \|\phi_t\| \le 2HG\epsilon_V.$$

Also, since $\hat{g}_{\mathrm{token}}$ is unbiased, Theorem 3(b) gives

$$\sqrt{\mathbb{E}\|\hat{g}_{\mathrm{token}} - \nabla J\|^2} \le \sqrt{\mathbb{E}\|\hat{g}_{\mathrm{token}}\|^2} \le \sqrt{H}\, G\sigma_R.$$

Minkowski's inequality proves the root-MSE bound. Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$ gives the displayed MSE bound. □

J Detailed score correlation analysis

As a diagnostic, we measured the pairwise score correlation $\rho = |\langle \phi_t, \phi_{t'} \rangle| / \|\phi_t\|^2$ for one GPT-2 parameter block (`transformer.ln_f.weight`, $d = 768$). We find $\hat{\rho} \approx 0.037$–0.040 across horizons, yielding $H\hat{\rho} \in \{0.65, 1.20, 2.39\}$ for $H \in \{16, 32, 64\}$.

For this block, $H\hat{\rho}$ is not small, while the observed proxy-based ratios are 15.8, 30.9, and 57.6 at the three horizons. Theorem 16, however, requires a uniform pathwise spectral bound on the full-gradient Gram matrix, so this single-block measurement neither verifies nor falsifies its premise.

The increasing deviation from the orthogonal reference as H grows is descriptive evidence only. The independent-prefix experiment in Section 10 more directly tests the estimator mechanism using full-model gradients and estimated conditional values.

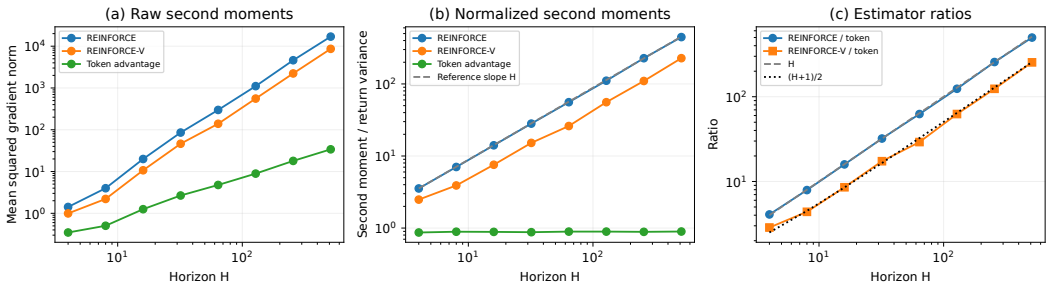


Figure 2: Fresh synthetic Token MDP check. Panel (b) divides by the actual population return variance; the trajectory reference has slope H, not H^2. The random per-step policies need not satisfy exact homogeneity, so the lines H and $(H+1)/2$ in panel (c) are conditional-geometry references.

K Score Correlation Diagnostic

Score correlation measurement. For one GPT-2 parameter block (`ln_f.weight`, $d = 768$), the measured pairwise score correlation was approximately 0.037–0.040. A single-block measurement cannot establish the full-gradient condition in Theorem 16; we report it only as a diagnostic. See Appendix J.

The original group-size plot is omitted from the revised claims because it concerns the idealized core estimator and does not validate the complete GRPO update.

L Earlier empirical checks

Synthetic Token MDP. We run a fresh reproducible check with $V = 10$, independent softmax policies in disjoint per-step parameter blocks, and unnormalized additive terminal rewards. Each horizon uses 50,000 sampled trajectories and seed $42000 + H$. We compute exact population second moments from the finite per-step action distributions and compare them with the sampled estimates. This is a new numerical check, not recovery of historical samples.

At $H = 512$, the sampled second-moment ratios are 498.96 and 254.37. The exact-moment comparison checks this implementation and its orthogonal construction; it is not evidence of universal ordering in shared-parameter LLMs.

Shared-parameter critic-budget check. We ran pretrained GPT-2 124M on five fixed text prompts with $H \in \{4, 8, 16, 32\}$, 128 main trajectories, a shared 2,048-rollout root-value estimate, and $M \in \{8, 32, 128\}$ independent continuations per visited nonterminal prefix. The terminal reward is the mean of fixed seeded token utilities in $[0, 1]$. Across all 60 configurations, the main trajectories and root values are identical across M, so the sweep holds the main samples fixed. We compute full-model gradients, not a selected parameter block.

At $M = 128$, the mean trajectory/token second-moment ratios are 3.23 ± 0.51, 7.13 ± 0.53, 11.03 ± 0.87, and 21.09 ± 1.08 for increasing H (sample standard deviations over the five fixed prompts). The corresponding reward-to-go/token ratios are 2.11, 3.72, 6.11, and 10.13. The descriptive log–log trajectory/token slopes are 0.45, 0.73, and 0.87 for $M = 8, 32, 128$, respectively. Thus the observed separation depends materially on critic budget; it is not a universal factor-H or end-to-end efficiency result.

For $S = 128$ main trajectories, prefix continuations generate $SMH(H-1)/2$ tokens, while root estimation generates $2048H$ and the main trajectories generate SH. At $H = 32, M = 128$, this is 2,001 generated tokens per main-trajectory token, versus 17 for the trajectory estimator including its shared root estimate. Token counts are not FLOPs, but they show that the critic is not free. Centered covariance traces are reported separately in Appendix D.2; the second moments above are not labeled variances. The previous centered-log-probability diagnostic is not an advantage estimator and is not used as confirmation of this experiment.

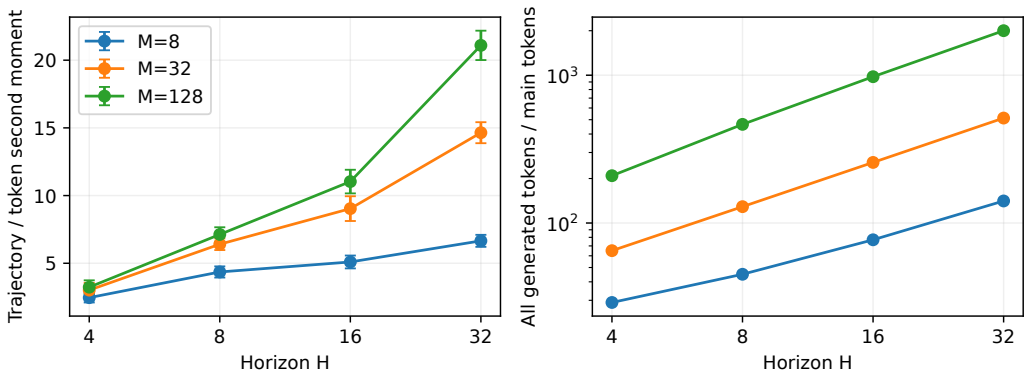


Figure 3: Completed GPT-2 critic-budget sweep. Left: second-moment ratios with standard deviations across five fixed prompts, not population confidence intervals. Right: actual generated-token accounting, including root and prefix continuations. Prefill work and gradient instrumentation are additional.

Score correlation diagnostic. The measured single-block correlation and its limitations are reported in Appendix K.

M Real finite-budget credit-assignment audit

Frozen protocol. The public model is Qwen2.5-1.5B-Instruct. LoRA modifies query/value projections with rank and alpha eight, dropout zero, and a shared seed-specific start. SGD uses learning rate 10^{-4}, batch four, and no gradient clipping. The prompt requests only a final numeric answer; generation has a strict 32-token horizon, temperature one, and no top-k/p truncation. GSM8K train/test records and engineering exclusions are frozen by content hash; the public split is reused from the preceding verifier audit, not a new benchmark. The first 64 fixed test records are evaluated greedily before and after training.

For each active nonterminal token, MC1/MC4 independently samples one/four continuations from the observed prefix to the same remaining horizon and averages their binary numeric rewards. EOS and horizon-end weights use the determined terminal reward. All conditional-generation costs are included. No critic is trained and no state-value baseline is subtracted. This conditional-Q audit therefore does not validate the oracle-advantage factor-of-horizon comparison.

Table 3: Every formal run. Time includes evaluation and checkpoint work within the runner, but not model loading; it is not isolated kernel throughput.

Seed	Arm	Final acc.	Updates	Main tokens	MC tokens	Minutes
41604	terminal	4.69	3099	65538	0	51.20
41604	mc1	6.25	905	18492	47044	32.10
41604	mc4	3.12	292	5846	60357	11.96
41605	terminal	3.12	3101	65537	0	51.66
41605	mc1	4.69	857	17695	47893	33.40
41605	mc4	3.12	269	5544	60985	11.43
41606	terminal	3.12	3100	65540	0	51.33
41606	mc1	4.69	914	18534	47033	32.95
41606	mc4	4.69	310	6133	59410	11.93

Fixed-policy gradient audit. Two additional initializations use 32 minibatches each without updating model parameters. For every arm we save the full 1,089,536-dimensional LoRA gradient. Main sampled trajectories match exactly across arms for each seed. The reported sample covariance trace subtracts the squared empirical mean gradient and uses the $n - 1$ denominator; it is not an uncentered second moment.

Table 4: All fixed-policy diagnostics. The finite-sample covariance estimates are descriptive, not population variance guarantees or evidence of training speed.

Seed	Arm	Sample variance trace	Main tokens	MC tokens
41997	terminal	3.12106	630	0
41997	mc1	2.48486	630	1741
41997	mc4	4.77727	630	6850
41998	terminal	7.28280	628	0
41998	mc1	3.32019	628	1370
41998	mc4	5.02867	628	5432

Available measurements. The analysis uses retained rewards, conditional averages, token costs and full fixed-policy gradients. Recomputing their moments does not independently repeat every neural forward and backward pass. The supplementary material provides gradient vectors, Gram matrices and analysis code; model weights must be obtained from the stated public checkpoints.

N FULL-ADAPTER TOKEN GRAM AND PAIRED GRADIENT AUDIT

At each of the two frozen initial policies (seeds 41997 and 41998), we recover the first 32 saved trajectories and differentiate each observed token log probability with respect to all trainable LoRA parameters. The resulting 64 Gram matrices have smallest and largest eigenvalues 1.82×10^{-11} and 5286.05. A single common midpoint scale $G^2 = 2643.03$ gives empirical $\epsilon = 0.9999999999999931$. Although this finite-sample value is technically below one, the upper ratio bound in Theorem 16 is uninformative at the observed horizons. Observed trajectory lengths are 2, 3, 4, 5, 6, 7, 9, and 32 tokens, so this variable-length diagnostic also does not directly instantiate the fixed-horizon comparison. Finite trajectories cannot establish its uniform pathwise premise. Thus these real-model measurements do not validate a useful near-isotropy condition or an oracle ordering for the learned MC-Q estimators.

We also bootstrap the 32 matched initial-policy minibatches jointly across terminal, MC1, and MC4 (10,000 resamples per seed). The MC1-minus-terminal variance-trace intervals are $[-3.586, 1.320]$ and $[-9.350, 0.083]$; both include zero. For MC4, the token-cost-times-variance differences have intervals $[297.3, 2185.9]$ and $[295.2, 1312.3]$. These diagnostics include continuation tokens but measure neither unbiased gradient MSE nor end-to-end optimizer efficiency. They reinforce the need to charge critic estimation and refrain from inferring a universal training speedup from the conditional theorem.