# SHARP HORIZON SCALING OF GRADIENT VARIANCE IN ON-POLICY DISTILLATION

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

## ABSTRACT

On-policy distillation (OPD) trains a student on self-generated sequences using positionwise teacher feedback. Temporal credit rules pair feedback with scores differently, but their tight horizon dependence remains unclear. Using a score–feedback matrix and conditional score centering, we show that past links affect variance without changing the mean, while future links can carry sequence credit. Under bounded feedback and scores, arbitrary prefix dependence, and shared parameters, we establish tight worst-case total variance of $\Theta(T^2)$ for token-local and $\Theta(T^3)$ for full-return and reward-to-go (RTG) in unnormalized rollout updates. One shared-parameter OPD construction attains all three rates at a fixed nonzero teacher–student gap, tightening the previous quartic full-return bound. Our control analysis characterizes removable variance and approximation costs; even optimal time-only controls retain cubic worst-case sequence variance. Full-parameter Qwen and Llama measurements show nearly linear token-local variance and faster sequence variance growth, revealing different length sensitivities. In controlled GPT-2 training with matched objectives and rollout budgets, RTG improves optimization, reducing normalized final reverse KL from 0.73 to 0.35 at $T = 512$. This advantage persists without gradient clipping.

## 1 INTRODUCTION

Knowledge distillation transfers behavior from a teacher to a student (Hinton et al., 2015). In autoregressive large language models (LLMs), on-policy distillation (OPD) trains the student on its own generations while a frozen teacher evaluates each position (Agarwal et al., 2024; Gu et al., 2024). Each rollout produces a random update, and longer generations create more cross-position interactions. At position $t$, the student–teacher log-ratio is the feedback and the sampled token's student log-probability gradient is the score term. A temporal credit rule pairs feedback with scores across positions: token-local uses only same-position feedback, full-return uses the sequence total for every score, and reward-to-go (RTG) uses current and later feedback. Because these pairings can change both the target and the variance, we must establish what each rule estimates before comparing noise. Figure 1 summarizes the three credit patterns and their consequences.

Prior work identifies the sequence reverse-KL target shared by full-return and RTG and the local target of token-local (Fu et al., 2026). It gives quadratic token-local and quartic full-return variance upper bounds. Policy-gradient research provides score orthogonality arguments and optimal trajectory baselines using historical information (Yuan et al., 2022; Keane & Gao, 2022). These bounds and coefficients leave open the tight rates attainable within OPD and the residual rates under restricted control information. OPD feedback is a log-ratio of normalized distributions, coupled to scores generated by shared student parameters. Matching lower bounds must respect these constraints. With bounded feedback and scores, arbitrary prefix dependence, and shared parameters, we study the tight worst-case variance of unnormalized rollout updates. We ask whether one student and teacher pair attains all three rates and how available prefix information limits the variance remaining after optimal controls that preserve the mean.

Our key insight is that past links can affect variance without changing the estimand, whereas future links may carry sequence credit. This temporal asymmetry follows from conditional score centering. Past feedback is fixed by the prefix before the current token is sampled, so its pairing with the score has zero mean; future feedback can depend on that token through subsequent generation. We express



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

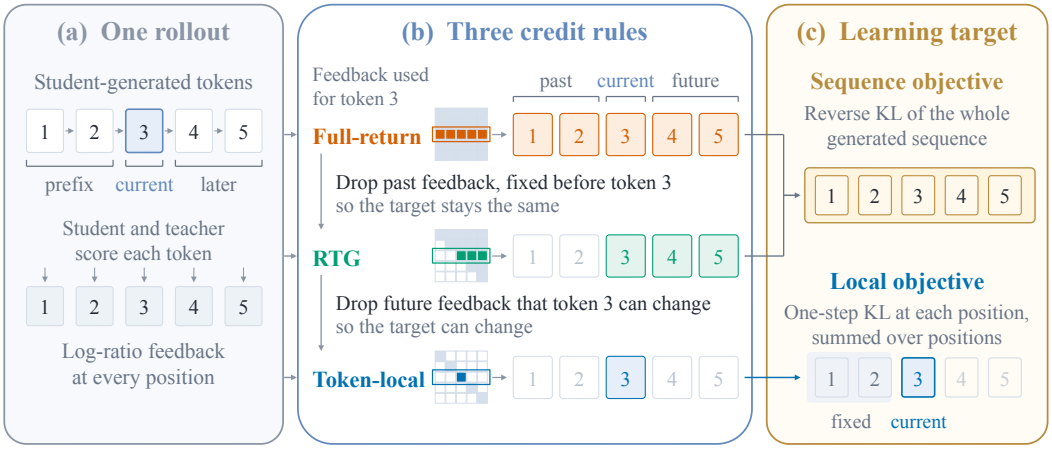


Figure 1: Temporal credit and learning targets on one rollout. **(a)** Student and teacher log probabilities give feedback at each generated position. **(b)** Mask rows index scores and columns index feedback; the highlighted row illustrates position 3. Removing past pairings preserves the mean, while removing future pairings can change it. **(c)** Full-return and RTG estimate the sequence reverse-KL gradient. Token-local estimates the sum of local reverse-KL gradients with sampled prefixes held fixed.

this structure through the score–feedback matrix in Figure 1(b). In this matrix, full-return and RTG differ only in the past block and therefore share a mean, while token-local also removes future links and generally changes the estimand. The same centering makes the cumulative score's expected squared norm grow at most linearly with length, giving cubic sequence variance upper bounds. One shared-parameter OPD construction with full support and a fixed nonzero teacher–student gap attains all three rates simultaneously, establishing tightness. Past links also form prefix controls that preserve the mean. This connects the passage from full-return to RTG with the broader analysis of optimal scalar prefix controls, their residual variance, and the limits imposed by available prefix information.

**Contributions.**

- **Past and future decomposition.** We represent token-local, full-return, and RTG as three masks of one score–feedback matrix, giving a common characterization of their estimands and variance upper bounds. Past links have zero mean under conditional score centering, while future links may carry sequence credit. This explains full-return/RTG mean equivalence and the gap between token-local and sequence targets.
- **Sharp horizon rates.** For unnormalized one-rollout updates, we prove tight worst-case total variance $\Theta(T^2)$ for token-local and $\Theta(T^3)$ for full-return and RTG. A common OPD construction attains all three rates with one shared parameter, full support, and a fixed nonzero teacher–student gap, establishing attainability within OPD and tightening the previous quartic full-return bound.
- **Prefix controls and exact approximation cost.** For square-integrable vector rollout statistics, we use a projection identity to decompose controlled variance into the optimal residual and weighted approximation error. The identity quantifies the exact cost of approximate controls and gives an error criterion for controls learned from independent auxiliary data to preserve a quadratic variance bound.
- **Prefix information limits.** Even optimal time-only controls leave cubic worst-case sequence variance. Full-prefix controls can eliminate variance in the nondegenerate scalar binary setting. A prefix-invariant local-KL schedule permits an $O(T^2)$ RTG variance bound using time-only controls.

## 2 Background and Motivation

**From fixed data to on-policy trajectories.** Traditional distillation matches a teacher on a fixed corpus, where the prefix at each position is supplied by the data rather than the current student (Hinton et al., 2015; Kim & Rush, 2016). At deployment, however, the student conditions on its own previous outputs, so an early token changes both the prefixes that follow and the teacher feedback observed there (Ross et al., 2011; Bengio et al., 2015; Ranzato et al., 2016). On-policy distillation addresses this distribution gap by training on student-generated sequences (Agarwal et al., 2024; Gu et al., 2024; Li et al., 2026), but thereby introduces a new choice. Training may treat sampled prefixes as fixed and update only the local teacher–student discrepancy, or it may retain how early decisions affect later feedback. Temporal credit rules determine how these cross-position effects enter the update. Recent variants adapt the local divergence or temporal credit (Jin et al., 2026; Liu et al., 2026; Fan et al., 2026).

**Known estimands and variance bounds.** Fu et al. (2026) clarify the mean updates of three common rules. Full-return and reward-to-go (RTG) estimate the sequence reverse-KL gradient, whereas token-local freezes the sampled-prefix distribution and generally estimates a different update. Pairings between past feedback and later token score terms can be removed from full-return without changing its mean, which produces RTG. The same work gives a quadratic variance upper bound for token-local and a quartic one for full-return, but does not establish whether either horizon rate is tight. Under an independent-noise model, Zhang et al. (2026b) analyze the error growth of a return assigned to one early decision; their statistic differs from the complete rollout gradient. Prefix-only distillation reduces rollout sampling cost (Zhang et al., 2026a). Score-function gradients and reward-to-go are classical tools (Williams, 1992; Sutton et al., 1999; Baxter & Bartlett, 2001); advantages and baselines reduce variance (Schulman et al., 2016; Greensmith et al., 2004), and Oh et al. (2026) derive a local prefix-dependent baseline for sampled-token OPD. Other work studies divergence choices, credit formulas, and detachment corrections (Wen et al., 2023; Zhang, 2026; Zhao et al., 2026). Together, these analyses identify the mean updates of the main estimators and several local corrections, but do not give a unified variance characterization of the complete rollout update.

**The remaining question.** Once the expected update of each rule is known, how quickly can its variance grow with generation length, and how much of that variation can be removed using the prefix already observed during sampling? Answering this question requires allowing later feedback to depend arbitrarily on earlier generations and all positions to share the same model parameters. It also requires matching upper and lower bounds, so that attainable horizon growth can be distinguished from slack introduced by pathwise analysis. Such a characterization would place the long-sequence cost of temporal credit and the limits of prefix-based variance reduction in one framework.

## 3 Setting

For a finite vocabulary $\mathcal{Y}$ and fixed horizon $T$, let $\pi_\theta$ and $q$ be autoregressive student and teacher distributions. The student samples $Y_t \sim \pi_{\theta_0}(\cdot \mid H_t)$ with prefix $H_t = (Y_1, \dots, Y_{t-1})$, including the fixed initial context; write $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$. The feedback is the student–teacher log-probability difference, and the score is the student's sampled log-probability gradient:

$$r_t = \log \pi_{\theta_0}(Y_t \mid H_t) - \log q(Y_t \mid H_t), \qquad s_t = \nabla_\theta \log \pi_\theta(Y_t \mid H_t)|_{\theta=\theta_0}. \tag{1}$$

The score $s_t$ is in $\mathbb{R}^d$, and parameters may be shared arbitrarily across positions.

Let $P_\theta$ and $Q$ be the student and teacher trajectory laws. The two training objectives are

$$J_{\mathrm{seq}}(\theta) = \mathrm{KL}(P_\theta \,\|\, Q), \tag{2}$$

$$J_{\mathrm{tok}}(\theta; \theta_0) = \sum_{t=1}^{T} \mathbb{E}_{H_t \sim P_{\theta_0}} [\mathrm{KL}(\pi_\theta(\cdot \mid H_t) \,\|\, q(\cdot \mid H_t))]. \tag{3}$$

At $\theta = \theta_0$ they have the same value by the KL chain rule, but not generally the same gradient: the second functional freezes the prefix law, whereas the first differentiates it.

Define the three rollout estimators by

$$g_{\mathrm{tok}} = \sum_{t=1}^{T} r_t s_t, \tag{4}$$

$$g_{\mathrm{full}} = \left(\sum_{u=1}^{T} r_u\right)\left(\sum_{t=1}^{T} s_t\right), \tag{5}$$

$$g_{\mathrm{rtg}} = \sum_{t=1}^{T}\left(\sum_{u=t}^{T} r_u\right) s_t \tag{6}$$

If $P_\theta \ll Q$ near $\theta_0$ and differentiation commutes with the trajectory and conditional KL integrals, they satisfy

$$\nabla J_{\mathrm{seq}}(\theta_0) = \mathbb{E} g_{\mathrm{full}} = \mathbb{E} g_{\mathrm{rtg}}, \qquad \nabla_\theta J_{\mathrm{tok}}(\theta;\theta_0)|_{\theta=\theta_0} = \mathbb{E} g_{\mathrm{tok}}. \tag{7}$$

Full-return and RTG estimate the sequence gradient; token-local is a sampled-token estimator of the frozen-prefix local gradient. Both sequence estimators use score-function gradients (Schulman et al., 2015; Mohamed et al., 2020). Full-vocabulary summation and sampling conventions are discussed in Appendix G.

**Definition 1** (Regular bounded process)**.** A process belongs to $\mathcal{G}_T(R,\sigma)$ if $r_t$ is $\mathcal{F}_t$-measurable,

$$|r_t| \le R, \qquad \mathbb{E}[\|s_t\|_2^2 \mid \mathcal{F}_{t-1}] \le \sigma^2, \qquad \mathbb{E}[s_t \mid \mathcal{F}_{t-1}] = 0 \quad \text{a.s.} \tag{8}$$

All displayed variables are square integrable. Dominated differentiation of a normalized conditional density implies the final identity.

Conditional score energy bounds yield dimension-free trace bounds. Appendix G covers general output spaces and fourth-moment extensions of the feedback bound. Write

$$\mathrm{Var}_{\mathrm{tr}}(G) = \mathbb{E}\|G - \mathbb{E}G\|_2^2, \qquad \mathrm{sd}_{\mathrm{tr}}(G) = \sqrt{\mathrm{Var}_{\mathrm{tr}}(G)}, \qquad M_u = \sum_{t=1}^{u} s_t, \qquad M_0 = 0. \tag{9}$$

For finite $d$, $\mathrm{Var}_{\mathrm{tr}}(G) = \mathrm{tr}\,\mathrm{Cov}(G)$.

Let a deterministic weight matrix $A \in \mathbb{R}^{T\times T}$ index scores by row and feedback by column, and define

$$G_A = \sum_{t=1}^{T}\sum_{u=1}^{T} A_{tu} r_u s_t, \qquad A_{\mathrm{tok}} = I, \quad A_{\mathrm{full}} = \mathbf{1}\mathbf{1}^\top, \quad (A_{\mathrm{rtg}})_{tu} = \mathbf{1}\{t \le u\}. \tag{10}$$

For these masks, $\|A\|_F^2$ counts $T$, $T^2$, and $T(T+1)/2$ retained links. Past feedback is fixed before the current token is sampled, so past links have zero mean but can affect variance. The next proposition connects these patterns to estimand relations and a common variance bound.

**Proposition 2** (Mean and variance bound for a credit rule)**.** *Every process in $\mathcal{G}_T(R,\sigma)$ obeys*

$$\mathbb{E}G_A = \sum_{1\le t\le u\le T} A_{tu}\mathbb{E}[r_u s_t], \qquad \mathrm{Var}_{\mathrm{tr}}(G_A) \le R^2\sigma^2 T\|A\|_F^2. \tag{11}$$

*In particular,*

$$\mathbb{E}g_{\mathrm{rtg}} = \mathbb{E}g_{\mathrm{full}} =: \mu_{\mathrm{seq}}, \qquad \mu_{\mathrm{seq}} - \mathbb{E}g_{\mathrm{tok}} = \sum_{1\le t<u\le T} \mathbb{E}[r_u s_t], \tag{12}$$

*and the token-local, full-return, and RTG masks give upper bounds $R^2\sigma^2T^2$, $R^2\sigma^2T^3$, and $\frac{1}{2}R^2\sigma^2T^2(T+1)$, respectively.*

Section 4 proves these worst-case rates sharp over the admissible process class. Dividing updates by $T$ preserves the mean relations and gives worst-case variance rates $\Theta(1)$, $\Theta(T)$, and $\Theta(T)$ (Appendix B).

For matching lower bounds, fix $c \in (0, 1/2)$ and $R, S > 0$ independently of $T$ to keep probabilities, feedback, and scores uniformly controlled.

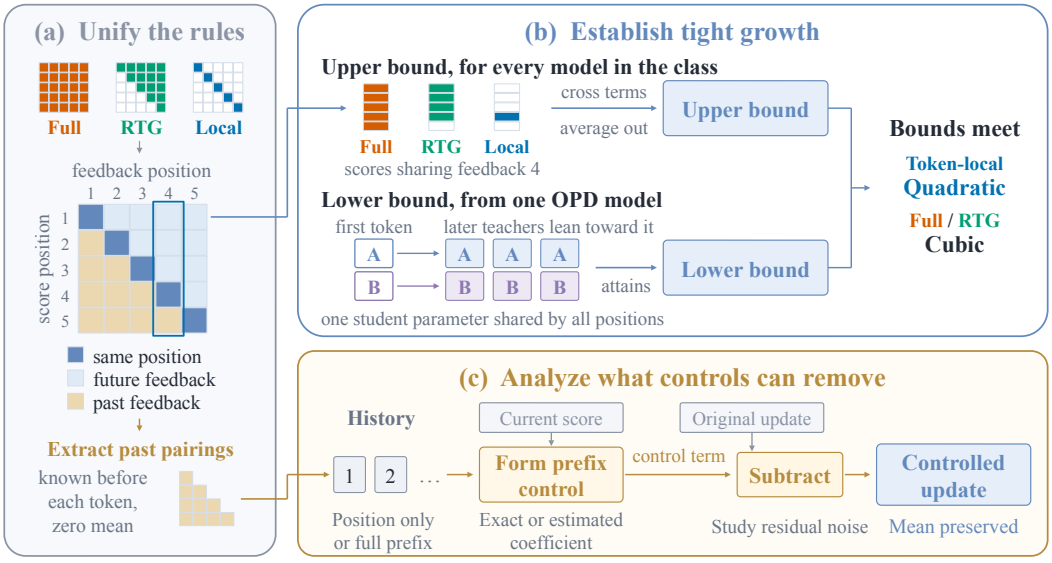


Figure 2: Establishing tight variance rates and analyzing prefix controls. **(a)** One matrix represents all three rules. Past feedback is known before the current token; its pairing with the score has zero mean. **(b)** Grouping scores by feedback yields upper bounds matched by an OPD instance with one shared parameter, full support, and a fixed gap between teacher and student. Raw worst-case variance is quadratic for token-local and cubic for both sequence rules. **(c)** Prefix controls combine history with the current score and preserve the mean. Available history and coefficient error govern the remaining variance. Appendix B gives the resulting rollout counts under both normalization conventions.

**Definition 3** (Bounded binary class)**.** A binary teacher–student pair is in $\mathcal{C}_T(c, R, S)$ if, at every student-reachable prefix and for both $x \in \{-1, 1\}$,

$$\pi_{\theta_0}(x \mid h), q(x \mid h) \in [c, 1-c], \qquad |r_t| \le R, \qquad \|s_t\|_2 \le S \quad \text{a.s.}, \tag{13}$$

and the parameterized conditional is differentiable and normalized near $\theta_0$.

This class lies inside $\mathcal{G}_T(R, S)$. Let $V_{\mathrm{tok}}(T)$, $V_{\mathrm{full}}(T)$, and $V_{\mathrm{rtg}}(T)$ be the corresponding variance suprema. The construction attaining all three rates below uses one scalar parameter shared across all positions.

## 4 Sharp Horizon Rates

Proposition 2 applies conditional score orthogonality, as in the policy-gradient moment bounds of Yuan et al. (2022). The cumulative score has $O(T)$ expected squared norm. Together with bounded feedback, this gives cubic sequence upper bounds.

Sharpness requires normalized teacher and student distributions whose log-ratios realize the feedback and whose shared parameter generates the scores. Our binary OPD constructions satisfy these constraints with constants uniform for $T \ge 4$.

$$L_c = \log \frac{1-c}{c}, \qquad \alpha = \frac{1}{2}\min\{S, 1\}, \qquad \delta = \frac{1}{4}\min\{R, L_c\}, \qquad a = \log\cosh \delta. \tag{14}$$

They are independent of $T$, and $\alpha, \delta, a$ are positive.

**Theorem 4** (Sharp worst-case variance rates)**.** *For every fixed $c \in (0, 1/2)$, $R, S > 0$, and integer $T \ge 4$,*

$$\frac{\alpha^2\delta^2}{4}T^2 \le V_{\mathrm{tok}}(T) \le R^2S^2T^2, \tag{15}$$

$$\alpha^2 a^2 T^3 \le V_{\mathrm{full}}(T) \le R^2S^2T^3, \tag{16}$$

$$\frac{\alpha^2 a^2}{3}T^3 \le V_{\mathrm{rtg}}(T) \le \frac{R^2S^2}{2}T^2(T+1) \le R^2S^2T^3. \tag{17}$$

*With $(c, R, S)$ fixed these read $V_{\mathrm{tok}}(T) = \Theta(T^2)$, $V_{\mathrm{full}}(T) = \Theta(T^3)$, and $V_{\mathrm{rtg}}(T) = \Theta(T^3)$, and the same powers are sharp for worst-coordinate variance and covariance operator norm.*

The operator statement holds because every covariance eigenvalue and every coordinate variance is at most the covariance trace, while the scalar lower-bound constructions make all three quantities identical.

The lower constructions all use one scalar parameter shared across positions. Let $X_1, \dots, X_T$ be independent Rademacher variables generated by the student

$$\pi_\theta^{(\alpha)}(x) = \frac{e^{\alpha\theta x}}{2\cosh(\alpha\theta)}, \qquad x \in \{-1, 1\}, \qquad \theta_0 = 0, \tag{18}$$

so $s_t = \alpha X_t$. A prefix-independent teacher gives cubic variance for both sequence estimators and linear variance for token-local. Appendix E gives the three exact variances.

For $\beta \in \mathbb{R}$, write $q_\beta(x) = e^{\beta x}/(2\cosh\beta)$ for the binary exponential-tilt teacher. The tilt $\delta$ fixes the teacher's gap from the student independently of $T$. The preceding sharp bounds concern each rule separately. One student–teacher pair can attain all three rates simultaneously.

**Theorem 5** (One instance attaining all three worst-case rates)**.** *Use the student (18) and a teacher with conditional $q_\delta$ at $t = 1$ and $q_{\delta z}$ at $t \geq 2$ after $X_1 = z$. The pair shares one scalar parameter across positions, belongs to $\mathcal{C}_T(c, R, S)$ for every $T$, has full support, and has distinct teacher and student conditionals at every prefix. Its estimator variances satisfy*

$$\mathrm{Var}(g_{\mathrm{tok}}) = \Theta(T^2), \qquad \mathrm{Var}(g_{\mathrm{full}}) = \Theta(T^3), \qquad \mathrm{Var}(g_{\mathrm{rtg}}) = \Theta(T^3).$$

The first token aligns later conditional local updates, giving quadratic token variance; teacher normalization produces cubic sequence variance. All three means coincide at $\theta_0$ (Appendix E).

Appendix E gives the exact variances in (65)–(67), proves all calibration and boundary claims and records how the lower constants depend on the teacher–student gap. Appendix G extends the upper theory to fixed vocabularies, weaker moment assumptions, predictable stopping, and fixed discounting or clipping. The discounting and clipping bounds characterize the resulting modified update rules.

## 5 Variance Reduction with Prefix Controls

Removing past feedback is a prefix control that maps full-return to RTG. Starting from $G = g_{\mathrm{full}}$, the choice $b_t = \sum_{u<t} r_u$ below gives $G^b = g_{\mathrm{rtg}}$ on every rollout. More generally, a prefix control subtracts $\sum_t b_t s_t$ using scalar, $\mathcal{F}_{t-1}$-measurable coefficients $b_t$. For square-integrable vector $G$, define

$$G^b = G - \sum_{t=1}^{T} b_t s_t, \qquad \mathbb{E}[b_t^2 \|s_t\|_2^2] < \infty. \tag{19}$$

Conditional score centering gives $\mathbb{E}G^b = \mathbb{E}G$. Action-dependent baselines are studied and reassessed (Wu et al., 2018; Tucker et al., 2018).

### 5.1 Optimal prefix control

For rollout gradients, the coefficient below agrees with Keane & Gao (2022). Its projection identity for vector statistics lets us analyze horizon rates under restricted prefix information.

**Theorem 6** (Optimal prefix control)**.** *Let*

$$D_t = \mathbb{E}[\|s_t\|_2^2 \mid \mathcal{F}_{t-1}], \qquad N_t = \mathbb{E}[\langle G - \mathbb{E}G, s_t\rangle \mid \mathcal{F}_{t-1}], \tag{20}$$

*and define*

$$b_t^\star = \begin{cases} N_t/D_t, & D_t > 0, \\ 0, & D_t = 0. \end{cases} \tag{21}$$

*Then $N_t = 0$ almost surely on $\{D_t = 0\}$ and, for every admissible sequence of coefficients $b$,*

$$\mathrm{Var}_{\mathrm{tr}}(G^b) = \mathrm{Var}_{\mathrm{tr}}(G) - \sum_{t=1}^{T} \mathbb{E}\left[\frac{N_t^2}{D_t}\mathbf{1}\{D_t > 0\}\right] + \sum_{t=1}^{T} \mathbb{E}[D_t(b_t - b_t^\star)^2]. \tag{22}$$

*Consequently $b^\star$ is a global minimizer within this scalar prefix-control class, unique modulo $D_t\,dP$-null sets.*

The identity separates residual variance from weighted approximation error. Critic and learned control variates (Gu et al., 2017; Grathwohl et al., 2018) and across-update variance reduction (Papini et al., 2018; Xu et al., 2020) use different information. Its unknown conditional expectations require estimation; the next corollary quantifies the cost with independent auxiliary data.

**Corollary 7** (Exact cost of a learned control)**.** *Let $\widehat{b}$ be an admissible sequence of prefix control coefficients, possibly trained on auxiliary data independent of the rollout used for $G$. Define its weighted oracle error by*

$$\mathcal{E}_T(\widehat{b}) = \sum_{t=1}^{T} \mathbb{E}[D_t(\widehat{b}_t - b_t^\star)^2]. \tag{23}$$

*Then*

$$\mathrm{Var}_{\mathrm{tr}}(G^{\widehat{b}}) = \mathrm{Var}_{\mathrm{tr}}(G^{b^\star}) + \mathcal{E}_T(\widehat{b}). \tag{24}$$

*In particular, if the oracle-controlled variance and $\mathcal{E}_T(\widehat{b})$ are both $O(T^2)$, the learned control preserves quadratic variance. Under $D_t \le \sigma^2$, ordinary prediction bounds $\mathbb{E}[(\widehat{b}_t - b_t^\star)^2] \le e_t^2$ imply*

$$\mathcal{E}_T(\widehat{b}) \le \sigma^2 \sum_{t=1}^{T} e_t^2. \tag{25}$$

The factor $D_t$ weights coefficient errors by conditional score energy, so equal coefficient errors contribute more variance at prefixes with larger score energy. For quadratic oracle residual variance, total weighted error $O(T^2)$ preserves the rate.

Time-only controls make both sequence variances quadratic on the common instance (Proposition 26). Yet even optimal coefficients leave cubic worst-case variance (Theorem 22). Products of early and later tokens survive every time-only control. Full-prefix controls eliminate variance for scalar $G, s_t$, binary tokens, and $D_t > 0$ almost surely (Corollary 21). Corollary 7 quantifies the error of learned coefficients.

In $\mathcal{G}_T(R, \sigma)$, RTG equals $P_T = \sum_u m_u M_u$ plus a residual of variance $O(T^2)$, where $m_u = \mathbb{E}[r_u \mid \mathcal{F}_{u-1}]$ (Theorem 23). Thus cubic variance requires cubic $P_T$ variance. Prefix-invariant $m_u$ make $P_T$ a time-only control, yielding an $O(T^2)$ bound after subtraction. General prefix dependence can make subtraction change the mean (Lemma 25). Proofs are in Appendix H.

### 5.2 When token-local variance is nearly linear

Dependence between local updates determines the token-local variance rate. Let $z_t$ be the local update, $\mu_t$ its prefix-conditional mean, and $A_T$ the accumulated predictable drift:

$$z_t = r_t s_t, \qquad \mu_t = \mathbb{E}[z_t \mid \mathcal{F}_{t-1}], \qquad A_T = \sum_{t=1}^{T} \mu_t. \tag{26}$$

**Theorem 8** (Token drift characterization)**.** *Every process in $\mathcal{G}_T(R, \sigma)$ obeys*

$$|\mathrm{sd}_{\mathrm{tr}}(g_{\mathrm{tok}}) - \mathrm{sd}_{\mathrm{tr}}(A_T)| \le R\sigma\sqrt{T}. \tag{27}$$

*If centered drifts $\bar{\mu}_t = \mu_t - \mathbb{E}\mu_t$ satisfy*

$$|\mathbb{E}\langle \bar{\mu}_i, \bar{\mu}_j \rangle| \le \Gamma(|i-j|), \qquad \Gamma(0) + 2\sum_{k \ge 1} \Gamma(k) < \infty, \tag{28}$$

*then $\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{tok}}) = O(T)$. If instead $\mathbb{E}\langle \bar{\mu}_i, \bar{\mu}_j \rangle \ge \kappa > 0$ for every $2 \le i \ne j \le T$, then $\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{tok}}) = \Theta(T^2)$. More generally, (27) implies $\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{tok}}) = \Theta(T^2)$ if and only if $\mathrm{Var}_{\mathrm{tr}}(A_T) = \Theta(T^2)$; the two covariance conditions are explicit tests for the linear and quadratic cases.*

This criterion explains the near-linear slopes observed below. Residual score noise contributes only the linear scale; quadratic growth requires conditional local updates to remain aligned across distant positions, while summable dependence leaves linear accumulation.

Appendix H gives the proofs and additional special cases.

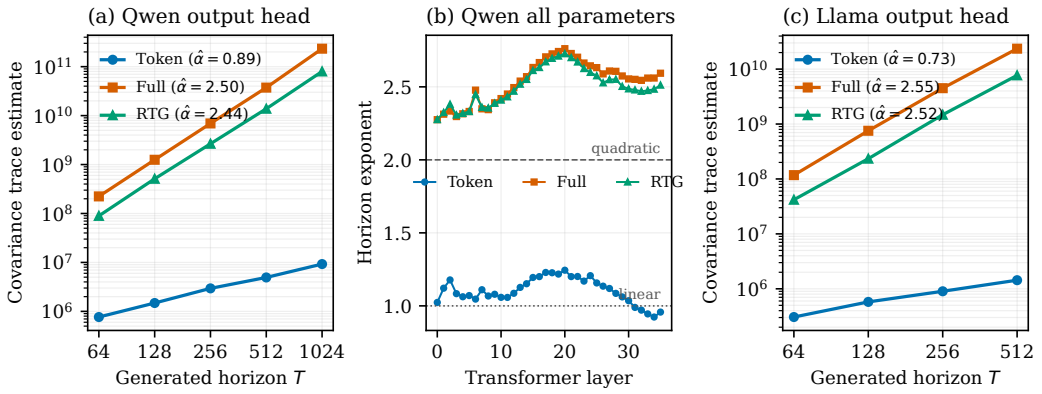


Figure 3: Gradient covariance-trace scaling: **(a)** 16 Qwen output-head probes, **(b)** exact exponents by Qwen layer, and **(c)** 16 Llama output-head probes.

Table 1: Fitted horizon exponents. The checkpoint row spans twelve students, and directional rows report quadratic forms of the covariance matrix.

| Measurement | Token | Full | RTG |
|---|---|---|---|
| Qwen, temperature direction ($T \leq 512$) | 0.72 | 2.12 | 2.18 |
| Qwen, 16 output-head probes | 0.89 | 2.50 | 2.44 |
| Qwen, all 4.02B parameters | 1.09 | 2.53 | 2.49 |
| Llama, 16 output-head probes | 0.73 | 2.55 | 2.52 |
| Llama, all 3.21B parameters | 1.16 | 2.47 | 2.55 |
| GPT-2, twelve training checkpoints | 0.99–1.38 | 2.68–2.92 | 2.63–2.89 |
| Qwen, held-out estimated signal direction | 0.61 | 1.77 | 1.72 |

## 6 Experiments

We measure variance growth with independently pretrained Qwen (Yang et al., 2025) and Llama (Grattafiori et al., 2024; Meta AI, 2024) pairs, then test mechanisms and full-parameter optimization with GPT-2 (Radford et al., 2019). We compare measured growth with the theoretical worst-case rates. Control fitting and evaluation use disjoint samples; Appendix C gives the protocols.

**Variance Scaling.** Token-local variance grows approximately linearly in the Qwen and Llama all-parameter measurements, while full-return and RTG grow substantially faster (Table 1). Figure 3 shows the same separation in output-head probes and across Qwen layers, indicating that it extends beyond a single parameter direction or layer. The Qwen3-4B/Qwen3-8B and Llama-3.2-3B/Llama-3.1-8B pairs use independently pretrained teachers. The separation also persists across saved GPT-2 checkpoints, prompt types, and teacher temperatures (Tables 9, 4, and 12), showing a recurring difference in length sensitivity across the tested settings. The separation in Qwen's all-parameter exponents persists after removing the largest rollout pair at each horizon (Appendix C.3).

**Variance Decomposition.** In the measured scalar directions, the RTG predictable local-KL component has variance close to the total and grows faster than the residual in both GPT-2 and Qwen (Figure 5). Thus the accumulation of this component is central to understanding RTG's length sensitivity. The GPT-2 token-local residual grows approximately linearly, consistent with the residual bound in Section 5. The decomposition thus locates which components carry the growth and helps identify what prefix controls would need to remove.

**Mean Gradient Comparison.** The full-return/RTG mean differences are consistent with sampling variation around a shared estimand, while full-return has roughly twice RTG's variance (Appendix C.6). Across internal probes, nonzero token-local means are detected in 12.3% to 14.8% of coordinates, compared with 4.8% to 5.7% for full-return and RTG (Table 10). We also project rollout

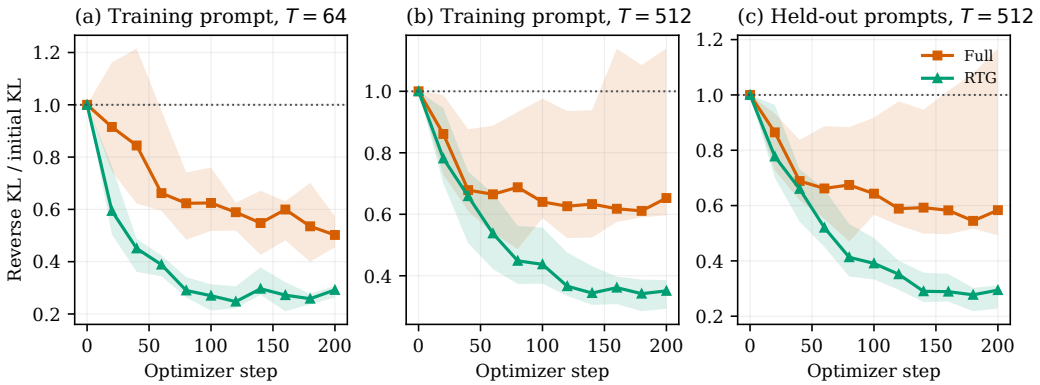


Figure 4: Full-parameter GPT-2 optimization. Reverse KL per token is normalized by its paired initial value; lines and shading show the median and range over five seeds. RTG finishes lower at both horizons and on held-out prompts under the same objective and rollout budget.

gradients onto a sequence gradient direction estimated from independent data. Along this direction, the variance exponents are 0.61 for token-local, 1.77 for full-return, and 1.72 for RTG (Table 11). The separation thus persists in a direction aligned with the mean sequence update.

**Temporal Dependence Analysis.** The measured upper quantiles of single-position feedback remain stable across horizons, while the variance of accumulated feedback grows faster than length (Appendix C.8). The combination of stable feedback magnitudes and growing accumulated variance supports temporal dependence as a source of the growth.

**Effect of Sample Size on Prefix Controls.** At fixed model parameters, we vary the sample size used to fit prefix controls. With the small auxiliary Qwen sample, both time-only and prefix-feature controls increase held-out variance (Figure 6). In GPT-2 at $T = 512$, we hold 1,024 evaluation rollouts per seed fixed. Full-return and RTG cross from variance inflation to reduction between 512 and 768 auxiliary rollouts, reaching median controlled/raw ratios of 0.32 and 0.31 at 2,048 (Table 2). This crossover illustrates Corollary 7, which separates the variance attainable by the optimal prefix control from the cost of estimating it.

**Optimization Performance.** In controlled GPT-2 training, RTG reaches lower final reverse KL than full-return under the same sequence objective and rollout budget (Figure 4). All five paired seeds show gains, supported by held-out evaluations (Tables 14 and 15), and the ordering persists without clipping (Table 18). Using the same reverse-KL metric, token-local at either tested learning rate reaches lower final KL than both sequence rules at $10^{-5}$ in every paired condition (Appendix C.12). At the common learning rate and $T = 512$ with clipping, mean final normalized KL is 0.027 for token-local, 0.355 for RTG, and 0.700 for full-return (Table 20). The normalized KL curve area also favors token-local at both tested rates in every paired comparison. The runs use AdamW (Loshchilov & Hutter, 2019) and a sharpened teacher; student entropy also falls (Appendix C.10).

## 7 Conclusion

This paper studies what the three temporal credit rules of on-policy distillation estimate and how their variance grows with generation length. Past links can affect variance without changing the mean; future links may carry sequence credit. Full-return and RTG thus share an estimand, while token-local may differ. Building on it, we prove that the worst-case variance is tightly quadratic for token-local and tightly cubic for full-return and RTG, with one shared-parameter construction attaining all three rates. Past links thereby become prefix controls, for which we derive the optimal coefficients, the exact cost of learned controls, and the limits of prefix information. Full-parameter Qwen and Llama measurements show slower variance growth for token-local than full-return or RTG. Controlled GPT-2 training favors RTG over full-return even without gradient clipping.

## AI USE STATEMENT

We used generative AI tools to assist with drafting, editing, translation, and feedback on the presentation of theoretical and experimental results. We also used these tools to prepare and revise figure and table code. During revision, we checked the resulting text against the stated assumptions and claims, and the displayed values against saved experiment summaries. The authors take responsibility for the accuracy and final content of this work.

## ETHICS STATEMENT

This work studies gradient variance in on-policy distillation through mathematical analysis and controlled experiments with existing pretrained language models. The experiments use rollouts generated by models and fixed prompts, with no human participants or new collection of personal data. The results concern estimator behavior and optimization under the tested settings. They do not establish the safety or fairness of deployed models.

## REPRODUCIBILITY STATEMENT

Section 3 states the assumptions and estimator definitions. Appendix A maps the proof dependencies, and the subsequent theoretical appendices provide detailed proofs and construction checks. Appendix C records model pairs, sampling and evaluation protocols, seed counts, gradient measurements, and computational costs. Appendices C.11 and C.12 specify the clipping interventions, learning rates, and paired comparisons. These details distinguish data used to fit controls from data used to evaluate them.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001. doi: 10.1613/JAIR.806. URL https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume15/baxter01a.pdf.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28*, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Jiabin Fan, Guoqing Luo, Michael Bowling, and Lili Mou. KETCHUP: K-step return estimation for sequential knowledge distillation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 778–796, 2026. doi: 10.18653/v1/2026.findings-eacl.39. URL https://aclanthology.org/2026.findings-eacl.39/.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026. URL https://arxiv.org/abs/2603.25562.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.00123.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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. URL https://jmlr.org/papers/v5/greensmith04a.html.

Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. Q-Prop: Sample-efficient policy gradient with an off-policy critic. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.02247.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026. URL https://arxiv.org/abs/2603.07079.

Ronan Keane and H. Oliver Gao. Variance reduction for score functions using optimal baselines. *arXiv preprint arXiv:2212.13587*, 2022. URL https://arxiv.org/abs/2212.13587.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop on Deep Reinforcement Learning Meets Structured Prediction*, 2019. URL https://openreview.net/forum?id=r1lgTGL5DE.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026. URL https://arxiv.org/abs/2604.13016.

Shiqi Liu, Zeyu He, Letian Tao, Guojian Zhan, Jiaxin Gao, Feihong Zhang, Jingliang Duan, Wei Xiong, Kehua Sheng, Bo Zhang, Yang Guan, and Shengbo Eben Li. Beyond token-local imitation: Reward-compatible temporal credit assignment for on-policy distillation. *arXiv preprint arXiv:2609.16937*, 2026. URL https://arxiv.org/abs/2609.16937.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Meta AI. Llama 3.2-3B model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020. URL https://www.jmlr.org/papers/v21/19-346.html.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. KL for a KL: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026. URL https://arxiv.org/abs/2605.07865.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4026–4035, 2018. URL https://proceedings.mlr.press/v80/papini18a.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06732.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems 28*, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/de03beffeed9da5f3639a621bcab5dd4-Abstract.html.

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

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pp. 1057–1063, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5015–5024, 2018. URL https://proceedings.mlr.press/v80/tucker18a.html.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. $f$-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834, 2023. doi: 10.18653/v1/2023.acl-long.605. URL https://aclanthology.org/2023.acl-long.605/.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696.

Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M. Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1tSsb-AW.

Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pp. 541–551, 2020. URL https://proceedings.mlr.press/v115/xu20a.html.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3332–3380, 2022. URL https://proceedings.mlr.press/v151/yuan22a.html.

Bowen Zhang. A formula-driven survey and research agenda for on-policy distillation. *arXiv preprint arXiv:2606.22793*, 2026. URL https://arxiv.org/abs/2606.22793.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler, II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026a. doi: 10.18653/v1/2026.findings-acl.1276. URL https://aclanthology.org/2026.findings-acl.1276/.

Yaocheng Zhang, Jiajun Chai, Yuqian Fu, Songjun Tu, Xiaohan Wang, Wei Lin, Guojun Yin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Are full rollouts necessary for on-policy distillation? *arXiv preprint arXiv:2605.31490*, 2026b. URL https://arxiv.org/abs/2605.31490.

Hanyang Zhao, Haoxian Chen, Han Lin, Genta Indra Winata, David Yao, and Wenpin Tang. OPD+: Rethinking the advantage design for on-policy distillation. *arXiv preprint arXiv:2606.01039*, 2026. URL https://arxiv.org/abs/2606.01039.

# A Proof Architecture

## A.1 Weight-matrix isometry and the raw upper powers

**Lemma 9** (Score-martingale horizon lemma)**.** *For every regular process,*

$$\mathbb{E}[s_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}\|M_u\|_2^2 = \sum_{t=1}^{u} \mathbb{E}\|s_t\|_2^2 \le u\sigma^2. \tag{29}$$

*No independence assumption on tokens or scores is required.*

If $i < j$, then $s_i$ is $\mathcal{F}_{j-1}$-measurable and $\mathbb{E}\langle s_i, s_j\rangle = 0$, which proves the lemma. More generally, the $u$th column of the weight matrix $A$ produces $C_u = \sum_t A_{tu} s_t$ with $\mathbb{E}\|C_u\|^2 \le \sigma^2 \sum_t A_{tu}^2$. Vector Cauchy–Schwarz then gives $\mathrm{Var}_{\mathrm{tr}}(G_A) \le R^2\sigma^2 T\|A\|_F^2$. The three mask energies immediately produce the quadratic, cubic, and cubic upper powers. For later use, finite-sum reindexing of the RTG mask gives

$$g_{\mathrm{rtg}} = \sum_{u=1}^{T} r_u M_u. \tag{30}$$

Appendix D gives all centering and integrability steps.

## A.2 Walsh components and one common lower instance

Under (18), $s_t = \alpha X_t$ for independent Rademachers. The fixed teacher has $r_t = a - \delta X_t$. Its nonzero constant $a$ generates the degree-one cubic components of both sequence forms, while the token sum has only linear variance.

For the persistent-sign teacher,

$$r_1 = a - \delta X_1, \qquad r_u = a - \delta X_1 X_u \quad (u \ge 2). \tag{31}$$

The token estimator contains $-\alpha\delta(T-1)X_1$, giving quadratic variance. In the full-return and RTG forms, the constant $a$ still produces $T$ degree-one coefficients whose squared sizes sum cubically. Exact Walsh orthogonality gives (65)–(67); no covariance sign argument is used.

## A.3 Why prefix controls change the exponent

The controls $b_t s_t$ are martingale differences and are mutually orthogonal in $L_2$. Expanding the square in (19) therefore separates over $t$ and completing each conditional square gives (22). Binary martingale representation proves the oracle corollary. Subtracting the oracle identity from the identity for a learned control gives the exact weighted prediction-error penalty in (24).

For RTG, define $m_u = \mathbb{E}[r_u \mid \mathcal{F}_{u-1}]$, $\eta_u = r_u - m_u$, and $P_T = \sum_{u=1}^{T} m_u M_u$. The equality

$$g_{\mathrm{rtg}} = P_T + \sum_{u=1}^{T} \eta_u M_u$$

is pathwise. After subtracting its conditional mean, each $\eta_u M_u$ has martingale noise with total second moment $O(T^2)$; the sum of the remaining conditional means also has norm $O(T)$. Hence the second term has $O(T^2)$ variance and only $P_T$ can carry a cubic component. If $m_u$ is deterministic, $P_T$ is exactly the valid prefix control in (130).

For token gradients, $r_t s_t - \mu_t$ is itself a vector martingale difference. Its total second moment is at most $R^2\sigma^2 T$, which gives (27). Summable versus coherent covariances of $\mu_t$ then distinguish the linear and quadratic regimes.

### A.4 ESTIMAND IDENTITIES

The two raw sequence forms are different random variables, but

$$g_{\mathrm{full}} - g_{\mathrm{rtg}} = \sum_{t=1}^{T} \sum_{u<t} r_u s_t. \tag{32}$$

For $u < t$, $r_u$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[r_u s_t] = 0$. Thus $\mathbb{E}g_{\mathrm{full}} = \mathbb{E}g_{\mathrm{rtg}}$, while the separate exact polynomials determine their variances. In contrast, future rewards need not be orthogonal to earlier scores, and their cross terms give exactly the gap between the sequence and token means in (12).

## B ROLLOUT, COMPUTE, AND OPTIMIZATION CONSEQUENCES

Let $G_1, \dots, G_n$ be independent copies of a one-rollout estimator $G$, and let $\bar{G}_n = n^{-1} \sum_i G_i$.

**Proposition 10** (Exact averaging and concentration)**.** *If $\mu = \mathbb{E}G$ and $V = \mathrm{Var}_{\mathrm{tr}}(G)$, then*

$$\mathbb{E}\|\bar{G}_n - \mu\|^2 = \frac{V}{n}, \qquad \Pr\{\|\bar{G}_n - \mu\| \geq \varepsilon\} \leq \frac{V}{n\varepsilon^2}. \tag{33}$$

*For exchangeable, possibly dependent rollouts with $\mathbb{E}\langle G_i - \mu, G_j - \mu\rangle = \kappa$ for $i \neq j$,*

$$\mathrm{Var}_{\mathrm{tr}}(\bar{G}_n) = \frac{V}{n} + \frac{n-1}{n}\kappa. \tag{34}$$

*Independence implies $\kappa = 0$; more generally, zero average trace cross-covariance is exactly the condition needed for the $1/n$ reduction.*

The meaning of an absolute error target depends on whether the training objective is summed or averaged over tokens. Define

$$\widetilde{g}_e = T^{-1} g_e, \qquad \widetilde{V}_e(T) = \sup_{P \in \mathcal{C}_T(c,R,S)} \mathrm{Var}_{\mathrm{tr},P}(\widetilde{g}_e). \tag{35}$$

**Theorem 11** (Summed versus token-averaged scaling)**.** *For fixed $(c, R, S)$ and $T \geq 4$,*

$$\widetilde{V}_{\mathrm{tok}}(T) = \Theta(1), \qquad \widetilde{V}_{\mathrm{full}}(T) = \Theta(T), \qquad \widetilde{V}_{\mathrm{rtg}}(T) = \Theta(T). \tag{36}$$

*More explicitly, the lower bounds for the three quantities are $\alpha^2\delta^2/4$, $\alpha^2 a^2 T$, and $\alpha^2 a^2 T/3$, while the corresponding upper bounds are $R^2S^2$, $R^2S^2T$, and $R^2S^2T$.*

Thus, for a sample mean and fixed *absolute* squared error $\varepsilon^2$, the summed convention has sharp worst-case variance-dominated rollout requirements

$$n_{\mathrm{tok}}(T, \varepsilon) = \Theta(T^2/\varepsilon^2), \qquad n_{\mathrm{full}}(T, \varepsilon) = n_{\mathrm{rtg}}(T, \varepsilon) = \Theta(T^3/\varepsilon^2). \tag{37}$$

Under the token-averaged convention the corresponding powers are

$$\widetilde{n}_{\mathrm{tok}}(T, \varepsilon) = \Theta(1/\varepsilon^2), \qquad \widetilde{n}_{\mathrm{full}}(T, \varepsilon) = \widetilde{n}_{\mathrm{rtg}}(T, \varepsilon) = \Theta(T/\varepsilon^2). \tag{38}$$

These statements concern fixed absolute MSE for each estimator's own mean and sufficiently large $T$; integer sample counts are at least one. The targets coincide when the gap in (12) vanishes.

Once local feedback and scores have been produced, prefix and suffix scans evaluate each displayed statistic in $O(T)$ additional vector operations. In an abstract one-unit-per-token accounting, the summed convention therefore has token-evaluation requirements

$$\Theta(T^3/\varepsilon^2), \qquad \Theta(T^4/\varepsilon^2), \qquad \Theta(T^4/\varepsilon^2). \tag{39}$$

The averaged convention gives respectively $\Theta(T/\varepsilon^2)$, $\Theta(T^2/\varepsilon^2)$, and $\Theta(T^2/\varepsilon^2)$. This accounting assigns one unit to each token evaluation. Implementation costs additionally include model forward and backward passes, teacher evaluation, vocabulary normalization, memory traffic, and control training.

Absolute variance changes with normalization. Relative variance measures noise against the squared mean. For $\mathbb{E}G \neq 0$, define

$$\mathrm{RelVar}(G) = \frac{\mathrm{Var}_{\mathrm{tr}}(G)}{\|\mathbb{E}G\|^2}. \tag{40}$$

**Proposition 12** (No universal relative-error exponent). *The class $\mathcal{C}_T(c, R, S)$ does not determine a single horizon power for* RelVar. *On the fixed-teacher witness,*

$$\mathrm{RelVar}(g_{\mathrm{tok}}) = \frac{a^2}{\delta^2 T}, \tag{41}$$

$$\mathrm{RelVar}(g_{\mathrm{full}}) = \frac{a^2}{\delta^2} T + 2\frac{T-1}{T}, \tag{42}$$

$$\mathrm{RelVar}(g_{\mathrm{rtg}}) = \frac{a^2(T+1)(2T+1)}{6\delta^2 T} + \frac{T-1}{2T}. \tag{43}$$

*In contrast, the persistent-sign witness has* $\mathrm{RelVar}(g_{\mathrm{tok}}) = \Theta(T^2)$. *Relative-error guarantees therefore require a separate lower bound on signal size or an instance class that controls $\|\mathbb{E}G\|$; they cannot be inferred from the variance envelope alone. Multiplying $G$ by $1/T$ leaves* $\mathrm{RelVar}(G)$ *unchanged.*

For the summed, fixed-absolute-MSE convention on the centered common instance in Proposition 26, all three variance-dominated requirements are quadratic in rollout count and cubic in abstract token evaluations. These exponents describe estimation of each statistic's own mean.

The variance powers also enter stochastic optimization directly.

**Theorem 13** (Smooth nonconvex SGD). *Let $F$ be $L$-smooth and bounded below by $F_\star$. Suppose that, conditionally on $\theta_k$, $G_k$ is an unbiased gradient with*

$$\mathbb{E}[\|G_k - \nabla F(\theta_k)\|^2 \mid \theta_k] \le \mathcal{V}_T, \tag{44}$$

*and update $\theta_{k+1} = \theta_k - \eta G_k$ with $\eta \le 1/L$. For $\Delta = F(\theta_0) - F_\star$,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla F(\theta_k)\|^2 \le \frac{2\Delta}{\eta K} + L\eta\mathcal{V}_T. \tag{45}$$

*For a target $\varepsilon > 0$, take*

$$\eta = \min\left\{\frac{1}{L}, \frac{\varepsilon^2}{2L\mathcal{V}_T}\right\}, \qquad K \ge \frac{4L\Delta}{\varepsilon^2}\max\left\{1, \frac{2\mathcal{V}_T}{\varepsilon^2}\right\}. \tag{46}$$

*Then the left side of (45) is at most $\varepsilon^2$. For fixed $L$, $\Delta$, and $\varepsilon$, the variance contribution to this bound scales with $\mathcal{V}_T$. The raw variance bounds contribute factors $T^2$ and $T^3$; division of the estimators by $T$ changes these variance factors to $1$ and $T$. Rescaling the objective also changes its smoothness, initial gap, and the gradient norm used to define accuracy. The full iteration bound retains all these quantities.*

When $\mathcal{V}_T = 0$, the second candidate step size is interpreted as $+\infty$. The theorem concerns SGD on a fixed objective with a variance bound uniform along training. For token-local, fixing a reference prefix law gives the objective $J_{\mathrm{tok}}(\theta; \theta_0)$ in (3). The training experiments use AdamW and refresh the rollout distribution after each update.

## C Empirical Mechanism and Control Diagnostics

This appendix analyzes the saved held-out rollouts from Section 6 and the GPT-2 campaign to test the mechanisms behind the raw slopes. Model fitting, control training, and evaluation use disjoint data. Prefix-only distillation instead truncates student-generated rollouts to save sampling compute (Zhang et al., 2026a); the gradients analyzed here use full horizons.

**Pathwise mechanism decompositions.** The saved feature $m_u = \mathbb{E}[r_u \mid \mathcal{F}_{u-1}]$ is the exact conditional local KL. Consequently the RTG update on every rollout admits the pathwise decomposition

$$g_{\mathrm{rtg}} = P_T + Q_T, \qquad P_T = \sum_{u=1}^{T} m_u M_u, \qquad Q_T = \sum_{u=1}^{T} (r_u - m_u) M_u. \tag{47}$$

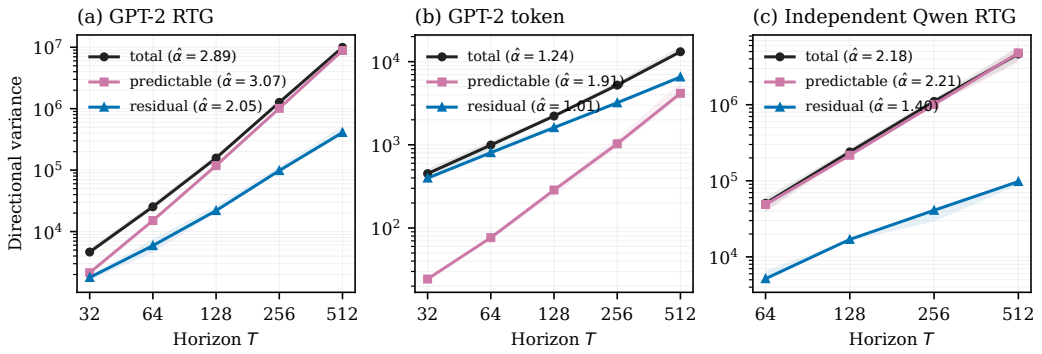


Figure 5: Predictable-component variance decomposition. "Predictable" is $P_T$ for RTG and $A_T$ for token; "residual" is $Q_T$ or $B_T$. Lines are median variances across seeds, shading is the seedwise range, and displayed slopes use $T \in \{64, 128, 256, 512\}$.

For the GPT-2 temperature teacher $q(\cdot \mid H_t) \propto \exp(\ell_t/\tau)$ with $\tau = 0.8$, the token drift is also available exactly. Since $r_t = (1 - \tau^{-1})\ell_t(Y_t) +$ a prefix-measurable constant and $s_t = \ell_t(Y_t) - \mathbb{E}[\ell_t(Y_t) \mid \mathcal{F}_{t-1}]$,

$$g_{\text{tok}} = A_T + B_T, \qquad A_T = \sum_{t=1}^{T} \mathbb{E}[r_t s_t \mid \mathcal{F}_{t-1}] = (1-\tau^{-1}) \sum_{t=1}^{T} D_t, \qquad B_T = g_{\text{tok}} - A_T. \quad (48)$$

Both pathwise identities hold to numerical precision.

**RTG variance is carried by the predictable component.** Figure 5 plots the decomposition. For GPT-2, the RTG, $P_T$, and $Q_T$ slopes are respectively 2.886, 3.068, and 2.051. At $T = 512$, the median ratios $\mathrm{Var}(P_T)/\mathrm{Var}(g_{\text{rtg}})$ and $\mathrm{Var}(Q_T)/\mathrm{Var}(g_{\text{rtg}})$ are 0.890 and 0.041. The predictable local-KL term therefore carries the near-cubic scaling of the GPT-2 RTG curve, while the theorem's residual is near quadratic.

The independent Qwen pair has lower exponents and the same mechanistic ordering. Its RTG, $P_T$, and $Q_T$ slopes are 2.177, 2.207, and 1.399. At $T = 512$, the predictable and residual variance ratios are 1.002 and 0.021; covariance makes the component ratios nonadditive. The raw curve is subcubic because $P_T$ itself is subcubic, consistent with the shrinking local feedback and score energy reported in Section 6.

For the GPT-2 token statistic, the total, drift $A_T$, and residual $B_T$ slopes are 1.242, 1.914, and 1.006. The predictable drift is the faster-growing component, while the residual remains approximately linear. This matches the structural separation in Theorem 8.

**GPT-2 control crossover.** Table 2 reports the nested auxiliary-sample intervention at $T = 512$. Each of five seeds uses one maximum-size auxiliary pool, and every row is evaluated on the same 1,024 untouched held-out rollouts. Sampling uses the row-wise CDF kernel. The table shows the learned time-only control; adding the four prefix features changes the sequence-estimator ratios by at most 0.3 percentage points, while the token ratio moves by up to 9 points at the smallest auxiliary pool.

Full and reward-to-go cross from variance inflation to reduction between 512 and 768 auxiliary rollouts and reach 68–69% median reduction at 2,048. The token control recovers more slowly, reaching a median ratio of 0.80. Together with the 16-rollout independent-Qwen result, this fixed-holdout intervention empirically separates control estimation error from the oracle variance term in Corollary 7.

### C.1 Scalar-direction campaign and measurement protocols

We sample on policy from a pretrained Qwen3-4B student and evaluate each sampled token under an independently pretrained, frozen Qwen3-8B teacher (Yang et al., 2025). Their tokenizer assets are

Table 2: Held-out controlled/raw variance ratios at $T = 512$ as the GPT-2 control-training pool grows. Entries are median [interquartile range] across five seeds; values below one are variance reduction.

| Auxiliary rollouts | Token | Full | RTG |
|---|---|---|---|
| 256 | 2.56 [2.41, 2.59] | 1.82 [1.74, 1.99] | 1.87 [1.76, 1.94] |
| 512 | 1.53 [1.44, 1.63] | 1.12 [1.03, 1.14] | 1.08 [0.96, 1.15] |
| 768 | 1.24 [1.16, 1.33] | 0.69 [0.67, 0.80] | 0.70 [0.70, 0.74] |
| 1,280 | 0.92 [0.91, 1.01] | 0.46 [0.42, 0.46] | 0.43 [0.43, 0.44] |
| 2,048 | 0.80 [0.76, 0.81] | 0.32 [0.32, 0.33] | 0.31 [0.31, 0.31] |

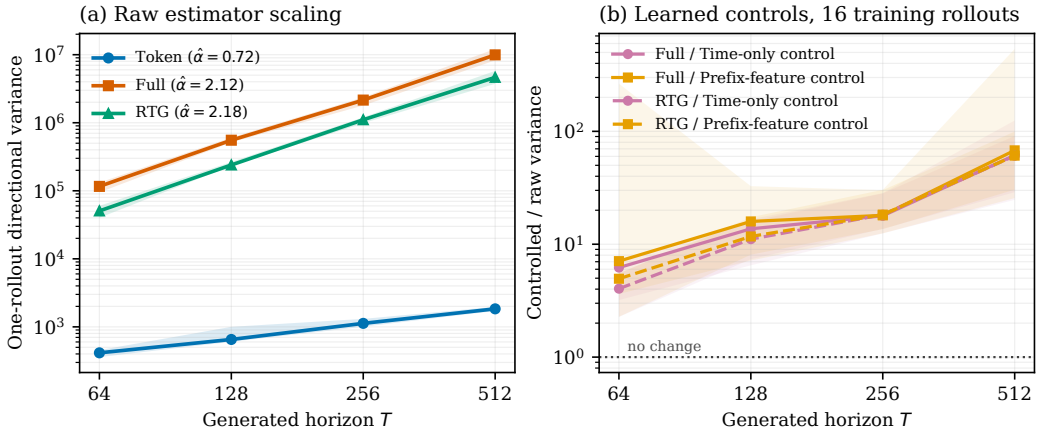


Figure 6: Independent pretrained Qwen student–teacher experiment. **Left:** raw one-rollout scalar directional variance; parenthesized values are log–log slopes fitted to the four median-variance points. **Right:** held-out controlled/raw variance for controls trained on 16 auxiliary rollouts. Lines are medians over three seeds and shading is the seedwise range.

identical. At a sampled prefix, let $\ell_t$ denote the frozen student logits and consider the shared scalar family $\pi_\beta(\cdot \mid H_t) \propto \exp(\beta \ell_t)$. At $\beta = 1$ the exact scalar score is

$$s_t = \ell_t(Y_t) - \mathbb{E}_{Y \sim \pi}[\ell_t(Y)], \qquad r_t = \log \pi(Y_t \mid H_t) - \log q(Y_t \mid H_t).$$

Horizons are $T \in \{64, 128, 256, 512\}$ with three independent seeds and 256 held-out rollouts per seed and horizon. Curves report the median seed variance and shading reports the seedwise minimum–maximum range. The fitted exponents are $0.723$, $2.122$ and $2.177$ for token, full and reward-to-go (Figure 6(a)). From $T = 64$ to $T = 512$, mean absolute local feedback decreases from $0.504$ to $0.375$ and mean conditional score energy from $3.62$ to $1.50$, which places this Qwen instance below the sharp worst-case envelope.

For the control intervention, 16 separate auxiliary rollouts fit either one coefficient per time step or a predictable ridge model using the current prefix's local KL, entropy, log score energy and top-one probability. Ridge strength is selected on an internal auxiliary split, and evaluation rollouts are used only for measurement. At $T = 512$ the median controlled/raw ratios are $61.36$ and $67.88$ for full with the time-only and prefix-feature controls, and $60.74$ and $60.91$ for reward-to-go.

### C.2 Multi-direction trace measurement

The independent-Qwen multi-direction campaign uses 16 fixed rank-one Rademacher perturbations of the output-head matrix. For direction $\Delta_j$, the scalar projection of a rollout gradient is evaluated without materializing the full gradient; averaging the squared projections gives a Hutchinson estimate of covariance trace in the output-head subspace. Each rollout seed contributes 128 trajectories at every horizon. Four seeds cover $T \in \{64, 128, 256, 512\}$ and two cover $T = 1024$; the same fixed probe matrix is used throughout, so every horizon enters one scaling fit.

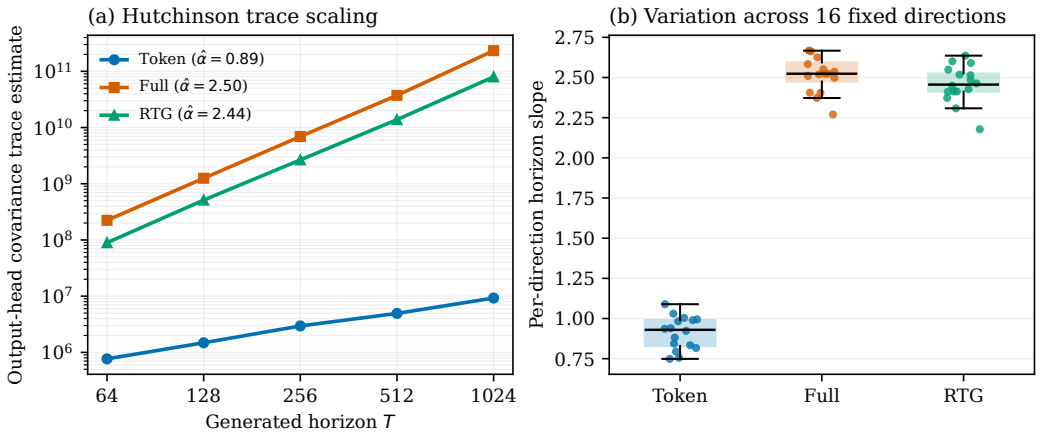


Figure 7: Multi-direction Qwen variance measurement. **Left:** pooled 16-direction output-head trace estimates over horizon. **Right:** slopes of all individual directions; boxes show the interquartile range and the horizontal line marks the median slope across directions. Full and RTG remain super-quadratic while token is near-linear.

Table 3: Independent-Qwen multi-direction slopes. Scalar slopes use the same pooled rollouts projected onto the temperature direction. The scalar fit pools $T = 64$–$1024$ and therefore differs from the $T \leq 512$ fit of Figure 6(a).

| Estimator | Trace slope | Scalar slope |
|---|---|---|
| Token | 0.893 | 0.765 |
| Full | 2.495 | 2.259 |
| RTG | 2.436 | 2.255 |

Table 3 shows that the scalar projection and output-head trace agree on the estimator ordering. Across 80 horizon-by-direction full–RTG mean comparisons, two exceed $|z| = 1.96$ and the maximum is 2.263; this is at or below the expected frequency for two unbiased estimators of the same sequence gradient.

Table 4: Independent-Qwen 16-direction slopes across prompts. Each row pools 512 rollouts per horizon over $T \in \{64, \ldots, 512\}$ for the scientific prompt and 256 for the others, with the same fixed probe matrix.

| Prompt | Token | Full | RTG |
|---|---|---|---|
| Scientific | 0.905 | 2.462 | 2.419 |
| Narrative | 0.839 | 2.374 | 2.246 |
| News | 0.848 | 2.454 | 2.355 |
| Code | 0.960 | 2.579 | 2.549 |

### C.3 Exact internal-parameter trace

Section 6 measures the covariance trace over the student's own weight blocks, extending the measurement beyond a probe subspace. The rollouts, feedback and sampled tokens are reused from the multi-direction campaign, and the student is replayed under teacher forcing to recompute per-token scores. The rollouts were generated in bfloat16 with a key-value cache, whereas replay uses a single full-sequence pass. The recomputed generation-time scale score correlates with the stored value at $0.9991$ or better, with a root-mean-square difference of at most $4.2\%$ of that score's standard deviation. Repeating the same cells in float32 changes the trace estimates by at most $0.6\%$.

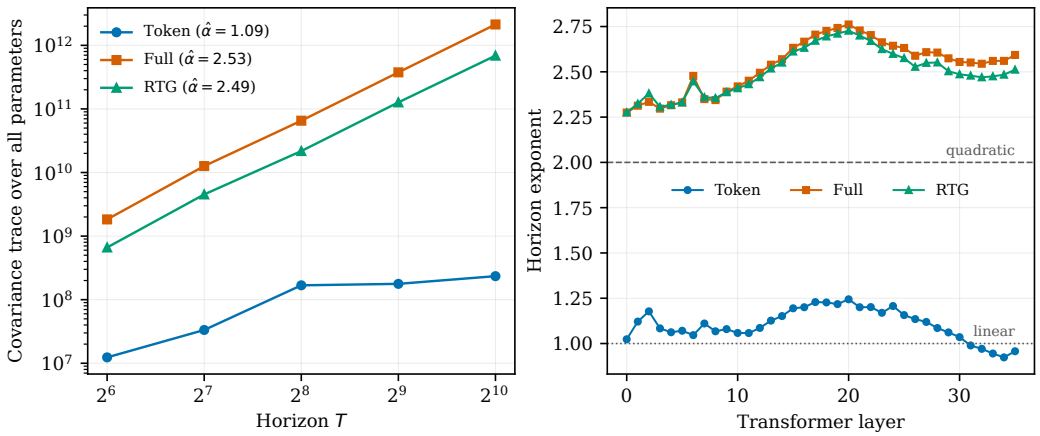


Figure 8: Exact covariance trace over the Qwen3-4B parameters. **Left:** trace over all 4.02B parameters against horizon. **Right:** the fitted horizon exponent $\hat{\alpha}$ of each transformer layer; the dashed and dotted lines mark exponents 2 and 1. Full and reward-to-go sit above the quadratic reference in every layer, while token stays near the linear reference.

For independent rollout gradients $g_i, g_j$, $\frac{1}{2}\|g_i - g_j\|^2$ is an unbiased covariance-trace sample. Each horizon uses 512 rollouts through $T = 512$ and 256 at $T = 1024$. Disjoint pairs therefore give 256 and 128 independent trace samples, respectively. Independent random re-pairing changes every reported slope by at most 0.001. The rank-one Rademacher cross-check of Table 6 restricted to the tied embedding returns 0.882, 2.515 and 2.467, reproducing the output-head result.

Table 5: Horizon exponents fitted separately to exact covariance traces. Attention + MLP combines the first two parameter groups. Adding the tied embedding trace to this combined trace gives the all-parameter trace.

| Block | Parameters | Token | Full | RTG |
|---|---|---|---|---|
| Attention projections | 0.94B | 1.184 | 2.633 | 2.599 |
| MLP projections | 2.69B | 1.062 | 2.481 | 2.445 |
| Attention + MLP | 3.63B | 1.102 | 2.534 | 2.493 |
| Tied embedding | 0.39B | 0.898 | 2.484 | 2.442 |
| All parameters | 4.02B | 1.090 | 2.527 | 2.485 |

Table 6: Two unbiased routes to the same traces. The paired estimator differentiates a contrast between two rollouts; the Hutchinson estimator projects single-rollout gradients onto rank-one Rademacher probes of each weight matrix.

| Block | Estimator | Paired exact | Rank-one Hutchinson |
|---|---|---|---|
| Attention + MLP | Token | 1.102 | 1.096 |
| Attention + MLP | Full | 2.534 | 2.496 |
| Attention + MLP | RTG | 2.493 | 2.442 |
| Tied embedding | Token | 0.898 | 0.882 |
| Tied embedding | Full | 2.484 | 2.515 |
| Tied embedding | RTG | 2.442 | 2.467 |

The token trace is sensitive to extreme rollout pairs at individual horizons. Table 7 removes the largest pair contribution separately for each rule and horizon, then refits the exponent using the remaining pairs. The exponent separation persists, although a single pair can strongly affect the absolute trace, especially for token-local. The main estimates of covariance trace use all pairs; the deletion analysis measures sensitivity to the largest contribution.

Table 7: Sensitivity of the Qwen all-parameter exponents over $T \in \{64, 128, 256, 512, 1024\}$. The two slope columns use all pairs or remove the largest contribution at each horizon and average the remaining pairs. The last columns give the largest share contributed by one pair to the original trace sum across horizons and its horizon.

| Rule | All pairs | Largest removed | Max. share (%) | $T$ at max. |
|---|---|---|---|---|
| Token-local | 1.090 | 1.064 | 58.6 | 256 |
| Full-return | 2.527 | 2.565 | 22.0 | 128 |
| RTG | 2.485 | 2.530 | 25.6 | 128 |

### C.4 Cross-family replication with Llama

The Llama campaign repeats the multi-direction and internal-parameter pipelines with a Llama-3.2-3B-Instruct student sampling on policy and an independently pretrained Llama-3.1-8B-Instruct teacher (Grattafiori et al., 2024; Meta AI, 2024). The two models share the same tokenizer. Two sampling seeds contribute 128 trajectories per horizon over $T \in \{64, \dots, 512\}$, and the same fixed probe construction is used as in the Qwen campaign. Replayed and stored scale scores have correlation at least $0.9998$. Table 8 reports both trace estimates; every parameter block is super-quadratic for the sequence estimators, and repeating the paired-trace measurement under an independent rollout pairing moves every tabulated slope by at most $0.002$.

Table 8: Llama-3.2-3B/3.1-8B horizon exponents. The output-head column is the 16-direction Hutchinson trace; the all-parameter column is the exact paired trace over 3.21B weights.

| Estimator | Output-head trace | All-parameter trace |
|---|---|---|
| Token | 0.733 | 1.158 |
| Full | 2.552 | 2.465 |
| RTG | 2.523 | 2.550 |

### C.5 Variance scaling along the training trajectory

Students saved during the 200-step end-to-end runs are re-measured with the scalar-direction pipeline, where the saved checkpoint samples on policy and feedback is evaluated using the original frozen teacher at $\tau = 0.8$, matching the training objective. Table 9 lists the raw exponents at every checkpoint; the step-0 row is the pretrained student, taken from the scientific-prompt campaign of Table 12, which used three seeds, unlike the two-seed cells listed here. The exponent ordering persists during training under both sequence rules and both training horizons. Between steps 25 and 200, the absolute variance at $T = 512$ decreases by a factor of 2.6–3.4 for the students trained at $T = 64$ and by 1.8–2.0 for the RTG run trained at $T = 512$; it remains nearly unchanged for the full run trained at $T = 512$, consistent with the optimization plateau reported in Section 6.

### C.6 Estimator means and the estimand gap

Each probe's $z$-score is its rollout mean divided by its standard error. Pooling 768 scalar rollouts gives a maximum full-minus-RTG $z$-score of $1.11$. Across the output-head probes, 2 of 80 tests exceed $|z| = 1.96$ against 4.0 expected under the null; across the internal probes, 1,018 of 20,240 tests exceed it against 1,012.0 expected. Full-return variance is $1.91$–$2.27$ times RTG variance.

Table 10 reports, for each quantity and horizon, the fraction of the 4,048 internal probe coordinates whose pooled mean differs from zero at $|z| > 1.96$; the null rate is $5\%$. The token mean exceeds this reference rate on $12.3\%$–$14.8\%$ of coordinates, whereas the full and RTG means lie in the range $4.8\%$–$5.7\%$. The full-minus-RTG calibration row, whose expectation is zero, lies between $4.7\%$ and $5.4\%$. The full-minus-token row tests the estimand gap of Proposition 2 and lies between $4.9\%$ and $5.6\%$ under these rollout budgets. In the scalar temperature direction, the paired gap test reaches $z = -2.06$ at $T = 1024$, where the sample means have opposite signs. Reported tests use the nominal two-sided threshold separately at each horizon. Table 11 reports fitted variance slopes along

Table 9: Raw log–log variance slopes measured at saved training checkpoints (GPT-2, scalar direction, two seeds, 512 held-out rollouts per horizon).

| Trained at | Rule | Step | Token | Full | RTG |
|---|---|---|---|---|---|
| – | – | 0 | 1.251 | 2.859 | 2.860 |
| $T = 64$ | Full | 25 | 1.244 | 2.913 | 2.888 |
| $T = 64$ | Full | 100 | 1.256 | 2.697 | 2.640 |
| $T = 64$ | Full | 200 | 1.350 | 2.764 | 2.680 |
| $T = 64$ | RTG | 25 | 1.235 | 2.836 | 2.786 |
| $T = 64$ | RTG | 100 | 1.375 | 2.755 | 2.626 |
| $T = 64$ | RTG | 200 | 1.331 | 2.777 | 2.700 |
| $T = 512$ | Full | 25 | 1.237 | 2.816 | 2.812 |
| $T = 512$ | Full | 100 | 1.133 | 2.852 | 2.801 |
| $T = 512$ | Full | 200 | 1.098 | 2.922 | 2.840 |
| $T = 512$ | RTG | 25 | 1.090 | 2.689 | 2.674 |
| $T = 512$ | RTG | 100 | 1.040 | 2.680 | 2.658 |
| $T = 512$ | RTG | 200 | 0.994 | 2.739 | 2.671 |

an independently estimated sequence gradient direction and the fraction of bootstrap slope estimates above one.

Table 10: Fraction of internal probe coordinates with $|z| > 1.96$ for each pooled mean (null rate $5\%$; 256 rollouts per horizon up to $T = 512$ and 128 at $T = 1024$).

| Quantity | $T = 64$ | $T = 128$ | $T = 256$ | $T = 512$ | $T = 1024$ |
|---|---|---|---|---|---|
| Token mean | 14.8% | 13.6% | 14.1% | 13.8% | 12.3% |
| Full mean | 5.0% | 4.8% | 5.5% | 5.2% | 5.1% |
| RTG mean | 5.5% | 4.9% | 5.5% | 5.7% | 4.9% |
| Estimand gap (full − token) | 5.0% | 4.9% | 5.6% | 5.2% | 5.0% |
| Calibration (full − RTG) | 5.0% | 4.7% | 5.4% | 4.9% | 5.2% |

### C.7 Variance along the estimated signal direction

The projection experiment of Section 6 uses one seed split to estimate the mean reward-to-go gradient over all $4.02$B parameters in float32, fixes its unit vector $u$, and projects per-rollout gradients from a disjoint seed split onto it. Every projection is therefore an independent scalar sample of $\langle g, u\rangle$. Table 11 gives the fitted slopes of the projection variance. Along this independently estimated direction, the one-rollout signal-to-noise ratio of the sequence estimators stays below $0.14$ at every horizon, quantifying the scale of the measured signal relative to one-rollout noise.

Table 11: Variance growth along the estimated sequence-signal direction, measured on fold-B rollouts. The probability column gives the fraction of bootstrap slope estimates above one.

| Estimator | Slope along $u$ | $P(\text{slope} > 1)$ |
|---|---|---|
| Token | 0.606 | 0.005 |
| Full | 1.767 | 1.000 |
| RTG | 1.719 | 1.000 |

### C.8 Feedback moment diagnostics

The theoretical bounds assume $|r_t| \le R$. We measure empirical feedback quantiles and moments across horizons. Over 750,000 scored positions, the $99\%$ and $99.9\%$ quantiles of $|r_t|$ stay within $[3.1, 3.5]$ and $[6.7, 7.9]$ across $T = 64$–$1024$, the observed maximum is $36.9$, and the kurtosis, though large, shows no upward trend with horizon. The variance of the total feedback grows super-linearly, with $\mathrm{Var}(\sum_t r_t)/T$ rising from $1.09$ at $T = 64$ to $3.68$ at $T = 1024$, direct evidence of the positive temporal feedback correlations that the predictable-mean decomposition of Theorem 23 ties to super-quadratic sequence variance.

### C.9 Robustness across prompts, temperatures, and models

Table 12 assembles every completed raw-scaling campaign. The GPT-2 interventions change one factor at a time and use three seeds with 512 held-out rollouts per setting. Temperature changes the magnitude of the sequence exponents, while all scientific, narrative, news, and code prompts preserve the ordering token $<$ full $\approx$ RTG. The same-model Qwen3-4B temperature teacher and the independent Qwen3-8B teacher recover the same separation.

Table 12: Raw log–log variance slopes across LLM settings. GPT-2 and Qwen3-4B temperature teachers divide the frozen student logits by $\tau$; Qwen3-4B/8B uses an independently pretrained teacher.

| Student / teacher | Setting | Token | Full | RTG |
|---|---|---|---|---|
| GPT-2 | Scientific, $\tau = 0.7$ | 1.243 | 2.937 | 2.939 |
| GPT-2 | Scientific, $\tau = 0.8$ | 1.251 | 2.859 | 2.860 |
| GPT-2 | Scientific, $\tau = 0.9$ | 1.259 | 2.626 | 2.641 |
| GPT-2 | Narrative, $\tau = 0.8$ | 1.191 | 2.847 | 2.849 |
| GPT-2 | News, $\tau = 0.8$ | 1.242 | 2.719 | 2.771 |
| GPT-2 | Code, $\tau = 0.8$ | 1.198 | 2.808 | 2.805 |
| Qwen3-4B | Temperature teacher, $\tau = 0.8$ | 1.079 | 1.365 | 1.370 |
| Qwen3-4B/8B | Independent teacher, scalar | 0.723 | 2.122 | 2.177 |
| Qwen3-4B/8B | Independent teacher, 16-dir. trace | 0.893 | 2.495 | 2.436 |

### C.10 End-to-end statistics and resource accounting

The student is GPT-2 124M (Radford et al., 2019), trained in all parameters against a frozen copy of its initial checkpoint sharpened to temperature $0.8$. Training uses AdamW (Loshchilov & Hutter, 2019) with learning rate $10^{-5}$ and zero weight decay. Each update averages four on-policy rollouts and normalizes by $T$ before unit-norm clipping; both rules use this same pipeline. We evaluate 64 fresh rollouts every 20 steps.

The common learning rate was selected in a pilot of 15 updates with candidates $\{3 \times 10^{-6}, 10^{-5}, 3 \times 10^{-5}\}$, both rules, both horizons, and seed 1103, which is also one of the five formal seeds. Evaluations used 32 rollouts every five steps. We minimized the mean area under the KL curve normalized by its initial value, averaged equally across rules and horizons. This pilot selects $10^{-5}$ for the original runs. Appendix C.11 reports a separate pilot with disjoint formal seeds and both clipping settings.

Clipping is active for both rules throughout formal training (Table 13). Every gradient is rescaled before AdamW accumulates its moments. Parameter updates therefore depend on the rescaled gradients and the optimizer's accumulated moments. The ablation in Appendix C.11 compares the same rules with and without gradient clipping. Training uses one prompt, with evaluation on that prompt and two held-out prompts.

Table 13: Clipping in the saved GPT-2 training logs. Each row pools five seeds with 200 updates each. Clipping follows rollout averaging and division by $T$, with a norm threshold of one. The final column gives the smallest pre-clipping norm across those updates.

| $T$ | Rule | Clipped (%) | Min. raw norm |
|---|---|---|---|
| 64 | Full-return | 100 | 20.46 |
| 64 | RTG | 100 | 8.04 |
| 512 | Full-return | 100 | 168.17 |
| 512 | RTG | 100 | 45.94 |

The paired end-to-end intervention starts full and RTG from identical GPT-2 parameters and uses common random-number schedules. Over 200 updates, RTG has lower final normalized KL and normalized KL area for all five paired seeds at both horizons. Table 15 adds the evaluations that lie outside the training objective, namely reverse KL on two prompts never used in training and the teacher's per-token NLL of the student's own generations. RTG has lower values on every held-out

measure in five of five seeds at $T = 512$ and in at least four of five at $T = 64$, showing the same improvement pattern beyond the training prompt.

Table 14: End-to-end paired results after 200 updates. KL/init averages each seed's final KL divided by its initial KL. The difference column averages paired differences $(\mathrm{KL}_{\mathrm{RTG}} - \mathrm{KL}_{\mathrm{Full}})/\mathrm{KL}_{\mathrm{init}}$, using the shared initial KL for each seed. "Grad." is the across-run mean of each run's median pre-clipping gradient norm.

| $T$ | Full KL/init | RTG KL/init | RTG−Full | Full grad. | RTG grad. |
|---|---|---|---|---|---|
| 64 | 0.513 | 0.291 | −0.222 | 98.8 | 32.4 |
| 512 | 0.734 | 0.346 | −0.388 | 331.1 | 110.1 |

Table 15: Step-200 outcomes beyond the training objective, averaged over five paired seeds. Columns give reverse KL per token on the training prompt and on two held-out prompts, and the teacher's per-token NLL of student generations.

| $T$ | Rule | Train KL | Narrative KL | Code KL | Teacher NLL |
|---|---|---|---|---|---|
| 64 | Full | 0.0783 | 0.0737 | 0.0741 | 2.944 |
| 64 | RTG | 0.0444 | 0.0476 | 0.0429 | 2.758 |
| 512 | Full | 0.1204 | 0.1235 | 0.1134 | 3.211 |
| 512 | RTG | 0.0568 | 0.0514 | 0.0478 | 2.646 |

The same evaluation logs also estimate the student's average conditional entropy on the training prompt. Teacher NLL and reverse KL are computed on the same $n = 64$ rollouts, giving, in nats per token,

$$\widehat{H}_{\mathrm{token}} = \widehat{\mathrm{NLL}}_{q,\mathrm{token}} - \widehat{\mathrm{KL}}_{\mathrm{token}} = -\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log \pi(Y_t^{(i)} \mid H_t^{(i)}).$$

Mean entropy decreases under both rules and is lower for RTG at the final checkpoint (Table 16). This is consistent with the sharpened teacher and shows that the KL gains accompany more concentrated student predictions.

Table 16: Student conditional entropy estimates in nats per token on the training prompt. Entries are means ± sample standard deviations across five seeds, computed from the same evaluations as Tables 14 and 15. Initial estimates are identical within each paired seed.

| $T$ | Shared initialization | Full-return, step 200 | RTG, step 200 |
|---|---|---|---|
| 64 | $3.811 \pm 0.052$ | $2.866 \pm 0.117$ | $2.714 \pm 0.073$ |
| 512 | $4.030 \pm 0.038$ | $3.091 \pm 0.246$ | $2.589 \pm 0.216$ |

Table 17: Measured computational resources for the original campaigns. GPU hours are summed sequential runtimes on one NVIDIA RTX A6000. Sampling campaigns report generation throughput; the two replay campaigns re-score stored rollouts, so their counts are positions differentiated and their rates are replay throughput.

| Campaign | GPU h | Tokens | Tokens/s | Peak GiB |
|---|---|---|---|---|
| GPT-2 scaling + control | 2.72 | 11,550,720 | 601.8–1453.3 | 0.85 |
| Independent Qwen scalar | 5.08 | 783,360 | 40.4–47.0 | 23.35 |
| Independent Qwen 16-dir. | 5.08 | 753,664 | 38.8–42.9 | 23.93 |
| Qwen prompt robustness | 5.62 | 737,280 | 33.7–41.0 | 23.36 |
| Independent Llama 16-dir. | 1.31 | 245,760 | 44.5–63.9 | 21.45 |
| Checkpoint variance | 6.73 | 17,694,720 | 520.3–841.9 | 1.66 |
| End-to-end GPT-2 (200 steps) | 9.55 | 8,663,040 | 191.5–265.9 | 5.50 |
| Internal-parameter replay | 1.40 | 2,498,560 | 117.4–837.3 | 33.18 |
| Signal-direction replay | 0.72 | 983,040 | 378.3 | 31.72 |

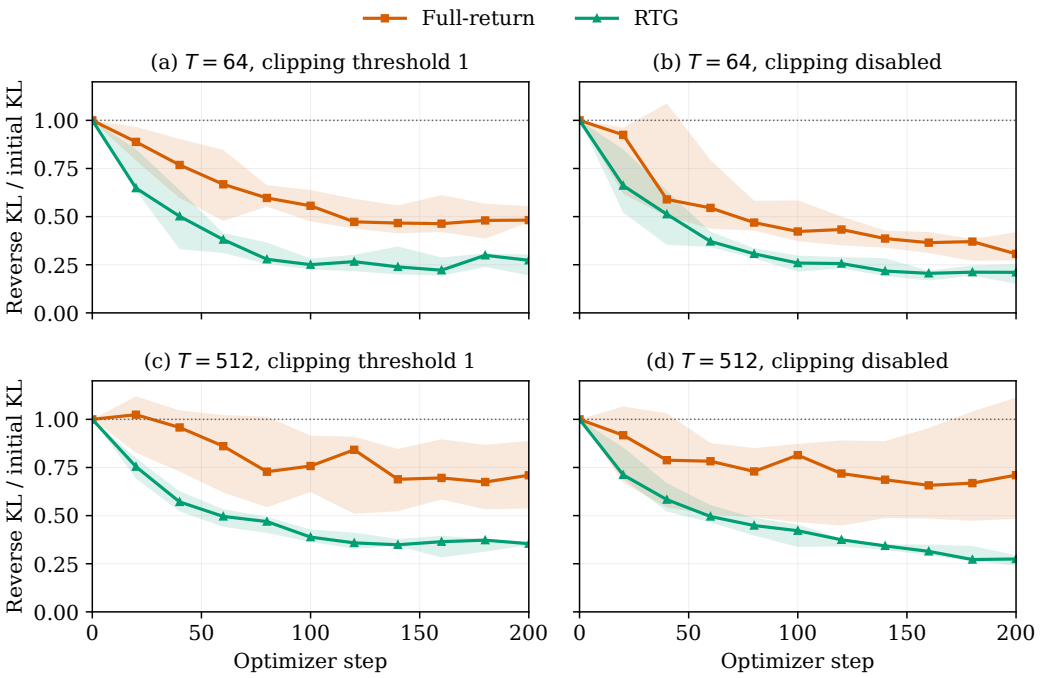


Figure 9: GPT-2 optimization with gradient clipping enabled or disabled. All conditions use learning rate $10^{-5}$. Lines show median reverse KL divided by initial KL over five seeds, and shading gives the full seed range. Evaluation uses the training prompt.

Table 17 accounts for 43.9 million token positions in 38.2 GPU-hours. Sampling dominates the billion-parameter measurements. The internal-parameter traces over the Qwen and Llama students cost 1.40 GPU-hours combined because they reuse rollouts whose generation cost 5.08 and 1.31 GPU-hours, and the signal-direction projection adds 0.72. Peak memory is 23.93 GiB while both Qwen models are resident for sampling, 33.18 GiB during the trace replay, which holds only the student but retains activations for three weighted backward passes, and 5.50 GiB for the full-parameter GPT-2 optimization.

### C.11 Gradient clipping ablation

We test whether RTG's optimization advantage persists when gradient clipping is disabled. The ablation crosses full-return and RTG with $T \in \{64, 512\}$ and either a gradient norm threshold of one or no clipping. Every condition starts from the same GPT-2 124M checkpoint and uses a frozen copy sharpened to temperature $0.8$ as teacher. All parameters are trained with AdamW, zero weight decay, and batches of four on-policy rollouts. Gradients are divided by $T$ before clipping. Dropout is disabled, and sampling uses the row-wise CDF kernel on the same scientific prompt as the original training.

The common learning rate is selected using candidates $\{3 \times 10^{-6}, 10^{-5}, 3 \times 10^{-5}\}$ and independent pilot seeds 6607 and 7703. Each of the 48 pilot runs takes 30 updates, with 32 fresh evaluation rollouts every ten updates. The selection rule minimizes the normalized KL curve area, averaged equally across rules, clipping settings, horizons, and pilot seeds, with ties favoring the smaller rate. This selects $10^{-5}$ for every formal condition. The formal comparison uses this common rate for all 200 updates.

Formal runs use seeds 1103, 2207, 3301, 4409, and 5501, with 200 updates per run and 64 fresh evaluation rollouts every 20 updates. The resulting 40 runs share initial parameters and random-number schedules within each seed and horizon, with each evolving student generating its own rollouts. Initial KL estimates are identical across all four paired conditions. Each curve is normalized by its own initial KL.

RTG reaches lower final normalized KL in every paired seed at both horizons under both clipping settings (Table 18). The normalized KL curve area has the same ordering (Figure 9). The advantage therefore persists without gradient clipping in this controlled setting.

Table 18: Final reverse KL divided by initial KL after 200 updates. Entries are means over five seeds. The final column reports the mean paired difference of the normalized values.

| $T$ | Gradient clipping | Full KL/init | RTG KL/init | RTG − Full |
|---|---|---|---|---|
| 64 | Threshold 1 | 0.493 | 0.262 | −0.231 |
| 64 | Disabled | 0.332 | 0.201 | −0.131 |
| 512 | Threshold 1 | 0.700 | 0.355 | −0.346 |
| 512 | Disabled | 0.733 | 0.272 | −0.460 |

Table 19 records both gradient norms and the actual parameter change per update. Every update triggers clipping when the threshold is one. Full-return has larger raw gradient norms and larger mean parameter update norms than RTG in every condition.

Table 19: Gradient and parameter update diagnostics. Entries average each run's 200 updates and then five seeds. All norms are Euclidean. Raw and post norms are measured before and after gradient clipping. The update norm is the actual parameter difference across an AdamW step.

| $T$ | Clipping | Rule | Clipped (%) | Raw norm | Post norm | Update norm |
|---|---|---|---|---|---|---|
| 64 | Threshold 1 | Full-return | 100.0 | 99.80 | 1.00 | 0.0274 |
| 64 | Threshold 1 | RTG | 100.0 | 40.03 | 1.00 | 0.0268 |
| 64 | Disabled | Full-return | 0.0 | 90.18 | 90.18 | 0.0232 |
| 64 | Disabled | RTG | 0.0 | 39.15 | 39.15 | 0.0211 |
| 512 | Threshold 1 | Full-return | 100.0 | 360.31 | 1.00 | 0.0292 |
| 512 | Threshold 1 | RTG | 100.0 | 115.27 | 1.00 | 0.0284 |
| 512 | Disabled | Full-return | 0.0 | 325.65 | 325.65 | 0.0275 |
| 512 | Disabled | RTG | 0.0 | 114.84 | 114.84 | 0.0221 |

All conditions use the verified `openai-community/gpt2` checkpoint and tokenizer on one MetaX C550 GPU. Their results are reported separately from the earlier NVIDIA runs. The pilot and formal campaigns use 2.65 and 11.16 GPU-hours, respectively, with peak allocated memory of 6.15 GiB; these totals exclude setup and short checks. Model and code hashes, configurations, and raw evaluations accompany the results. Pilot initial KL estimates differ by at most $0.00650$ across otherwise matched runs; these samples are retained with their original normalization. Model and evaluation identities were verified.

The ablation measures reverse KL on the training prompt against the sharpened teacher. Full-return and RTG use $10^{-5}$ throughout the formal comparison. The held-out prompt evaluations in Table 15 belong to the original campaign.

### C.12 Comparing token-local and sequence updates

We add sampled-token training to the clipping ablation to compare all three credit rules using sequence reverse KL as a common evaluation metric. Token-local multiplies each sampled token's score by its detached local log-ratio. Its expected update is the local gradient with the current prefix distribution held fixed, as defined in Section 3. Each evolving student generates its own rollouts.

The additional runs use the same GPT-2 assets, frozen teacher, optimizer, prompt, sampling procedure, and rollout budgets as Appendix C.11. Five paired seeds cover both horizons and both clipping settings, giving 20 token-local runs at the common learning rate $10^{-5}$. This rate comes from the existing full-return/RTG pilot. We also evaluate 20 token-local runs at $3 \times 10^{-5}$, the rate obtained from its separate 30-update pilot using the same candidate grid and independent pilot seeds. Every formal run takes 200 updates, with 64 fresh evaluation rollouts every 20 updates. Initial KL estimates agree exactly across all paired conditions.

Token-local reaches lower final normalized KL than both sequence rules in every paired seed at both horizons and under both clipping settings (Table 20). At the common learning rate and $T = 512$ with

clipping, its mean is $0.027$, compared with $0.355$ for RTG and $0.700$ for full-return. The additional token-local learning rate preserves this ordering against the sequence rules at $10^{-5}$. Normalized KL curve area also favors token-local in every paired comparison against the sequence rules at both tested token-local rates. The outcome ordering therefore persists across these two rates and with clipping enabled or disabled.

Table 20: Final reverse KL divided by initial KL after 200 updates. Entries are means of five seedwise ratios on the training prompt. Column headings give learning rates. The three $10^{-5}$ columns form the common learning rate comparison. The final column adds a second token-local learning rate.

| $T$ | Clipping | Full-return | RTG | Token-local | |
|---|---|---|---|---|---|
| | | $10^{-5}$ | $10^{-5}$ | $10^{-5}$ | $3 \times 10^{-5}$ |
| 64 | Threshold 1 | 0.493 | 0.262 | 0.056 | 0.155 |
| 64 | Disabled | 0.332 | 0.201 | 0.051 | 0.140 |
| 512 | Threshold 1 | 0.700 | 0.355 | 0.027 | 0.053 |
| 512 | Disabled | 0.733 | 0.272 | 0.032 | 0.042 |

Training and evaluation use one prompt and a sharpened teacher. The additional rate uses two completed runs from machine B and 18 from machine C. Both machines have C550 GPUs and use identical model assets, training code, and principal software versions; their execution provenance is recorded separately.

# D Universal Upper Bounds

All expectations in this appendix are under the student rollout at $\theta_0$.

## D.1 Conditional score centering and isometry

For completeness, suppose a conditional density $p_\theta(y \mid h)$ is differentiable at $\theta_0$ and its derivative is dominated by an integrable function in a neighborhood of $\theta_0$. Then

$$\begin{aligned}\mathbb{E}[s_t \mid H_t = h] &= \int p_{\theta_0}(y \mid h)\, \nabla_\theta \log p_\theta(y \mid h)|_{\theta_0}\, \nu_h(dy) \\ &= \nabla_\theta \int p_\theta(y \mid h)\nu_h(dy)\Big|_{\theta_0} = \nabla_\theta 1 = 0. \end{aligned} \tag{49}$$

Definition 1 takes this identity as the regularity condition, so the subsequent argument also applies when no common density is introduced.

*Proof of Lemma 9.* For $i < j$, $s_i$ is $\mathcal{F}_{j-1}$-measurable. Conditional Cauchy–Schwarz and the assumed second moments justify the tower step

$$\mathbb{E}\langle s_i, s_j\rangle = \mathbb{E}\left[\langle s_i, \mathbb{E}[s_j \mid \mathcal{F}_{j-1}]\rangle\right] = 0. \tag{50}$$

Expanding $M_u = \sum_{t=1}^{u} s_t$ therefore gives

$$\mathbb{E}\|M_u\|_2^2 = \sum_{t=1}^{u} \mathbb{E}\|s_t\|_2^2 = \sum_{t=1}^{u} \mathbb{E}\big[\mathbb{E}[\|s_t\|_2^2 \mid \mathcal{F}_{t-1}]\big] \le u\sigma^2. \tag{51}$$

□

The calculation permits arbitrary prefix dependence, shared parameters, and measurable output spaces.

### D.2 Objective and weight-matrix identities

Under the additional interchange condition stated before (7), differentiating the trajectory reverse KL at $\theta_0$ gives

$$\nabla J_{\mathrm{seq}}(\theta_0) = \mathbb{E}\left[\left(\sum_{u=1}^{T} r_u\right) M_T + M_T\right] = \mathbb{E} g_{\mathrm{full}}, \tag{52}$$

because $\mathbb{E}M_T = 0$. Conditional differentiation of each frozen-prefix reverse KL similarly gives $\mathbb{E}[(r_t+1)s_t \mid H_t] = \mathbb{E}[r_t s_t \mid H_t]$; summing proves the token identity in (7). The reward-to-go identity follows from the mean calculation below.

*Proof of Proposition 2.* If $u < t$, then $r_u$ is $\mathcal{F}_{t-1}$-measurable and

$$\mathbb{E}[r_u s_t] = \mathbb{E}[r_u \mathbb{E}[s_t \mid \mathcal{F}_{t-1}]] = 0. \tag{53}$$

Expanding $G_A$ and deleting precisely these terms proves the mean identity. The full and reward-to-go masks differ only on this strict past-feedback triangle, and comparing the full and token masks gives (12).

For the variance bound, put $C_u = \sum_t A_{tu} s_t$. Score orthogonality gives

$$\mathbb{E}\|C_u\|_2^2 = \sum_{t=1}^{T} A_{tu}^2 \mathbb{E}\|s_t\|_2^2 \le \sigma^2 \sum_{t=1}^{T} A_{tu}^2. \tag{54}$$

Since $G_A = \sum_u r_u C_u$, vector Cauchy–Schwarz yields

$$\mathrm{Var}_{\mathrm{tr}}(G_A) \le \mathbb{E}\|G_A\|_2^2 \le R^2 T \sum_{u=1}^{T} \mathbb{E}\|C_u\|_2^2 \le R^2 \sigma^2 T \|A\|_F^2. \tag{55}$$

Finally, $\|A_{\mathrm{tok}}\|_F^2 = T$, $\|A_{\mathrm{full}}\|_F^2 = T^2$, and $\|A_{\mathrm{rtg}}\|_F^2 = T(T+1)/2$, proving the three stated bounds. They establish Theorem 16 in Appendix G; taking $\sigma = S$ and the supremum over $\mathcal{C}_T(c, R, S)$ gives the upper sides of Theorem 4. □

## E Pinned Lower-Bound Constructions

We now define the two model families in Proposition 14 and Theorem 5 and verify every class constraint. The families use the constants in (14) without any dependence on $T$.

### E.1 Shared-parameter Rademacher student

At every prefix, use the same scalar conditional

$$\pi_\theta^{(\alpha)}(x) = \frac{\exp(\alpha\theta x)}{2\cosh(\alpha\theta)}, \qquad x \in \{-1, 1\}. \tag{56}$$

At $\theta_0 = 0$ this distribution is uniform, independently of the prefix. Thus $X_1, \dots, X_T$ are independent Rademacher random variables. Direct differentiation gives

$$\partial_\theta \log \pi_\theta^{(\alpha)}(x)\Big|_{\theta=0} = \partial_\theta\big(\alpha\theta x - \log(2\cosh(\alpha\theta))\big)\big|_{\theta=0} = \alpha x, \tag{57}$$

so $s_t = \alpha X_t$. The same scalar $\theta$ controls every time step.

### E.2 Fixed teacher for the sequence estimators

For both $g_{\mathrm{full}}$ and $g_{\mathrm{rtg}}$, set the teacher conditional at every prefix to

$$q_\delta(x) = \frac{e^{\delta x}}{2\cosh\delta}. \tag{58}$$

Since the student is uniform at $\theta_0$, the sampled log-ratio is

$$r_t = \log \frac{1/2}{e^{\delta X_t}/(2\cosh\delta)} = \log\cosh\delta - \delta X_t = a - \delta X_t. \tag{59}$$

The same teacher is used at every prefix.

**Proposition 14** (Fixed-teacher exact calculation)**.** *With the shared-parameter student* (18) *and the prefix-independent teacher* (58)*, all three estimators are evaluated on one instance and*

$$\mathrm{Var}(g_{\mathrm{tok}}) = \alpha^2 a^2 T, \tag{60}$$

$$\mathrm{Var}(g_{\mathrm{full}}) = \alpha^2 \left[a^2 T^3 + 2\delta^2 T(T-1)\right], \tag{61}$$

$$\mathrm{Var}(g_{\mathrm{rtg}}) = \alpha^2 \left[a^2 \frac{T(T+1)(2T+1)}{6} + \delta^2 \frac{T(T-1)}{2}\right]. \tag{62}$$

*Thus this single i.i.d. teacher–student pair separates linear token variance from cubic variance in both raw sequence estimators.*

### E.3 Persistent-sign teacher attaining all three rates

For $g_{\mathrm{tok}}$, retain $q_\delta$ at time 1. At every time $t \geq 2$, a prefix begins with some $X_1 = z \in \{-1, 1\}$; set its teacher conditional to

$$q_\delta(x \mid z) = \frac{e^{\delta z x}}{2\cosh\delta} = q_{\delta z}(x). \tag{63}$$

This is a valid autoregressive teacher because $z$ is known from the prefix. The log-ratios are

$$r_1 = a - \delta X_1, \qquad r_t = a - \delta X_1 X_t \quad (t \geq 2). \tag{64}$$

The student remains i.i.d., and every later $X_t$ is a fresh Rademacher sample.

**Exact variances on the same instance.** For the pair in Theorem 5, all three estimators satisfy

$$\mathrm{Var}(g_{\mathrm{tok}}) = \alpha^2 \left[(a - \delta(T-1))^2 + a^2(T-1)\right], \tag{65}$$

$$\mathrm{Var}(g_{\mathrm{full}}) = \alpha^2 \left\{ [aT - \delta(T-1)]^2 + (T-1)(aT-\delta)^2 + \delta^2(T-1) + 4\delta^2 \binom{T-1}{2} \right\}, \tag{66}$$

$$\mathrm{Var}(g_{\mathrm{rtg}}) = \alpha^2 \left\{ [aT - \delta(T-1)]^2 + \sum_{k=1}^{T-1} (ak - \delta)^2 + \delta^2 \binom{T-1}{2} \right\}. \tag{67}$$

Consequently their variances on this same pair are respectively $\Theta(T^2)$, $\Theta(T^3)$, and $\Theta(T^3)$. At $\theta_0$, $\mathbb{E}g_{\mathrm{tok}} = \mathbb{E}g_{\mathrm{full}} = \mathbb{E}g_{\mathrm{rtg}} = -\alpha\delta$ because the future-link terms cancel in expectation. This variance separation thus occurs even when the three mean updates coincide at the evaluation point. Appendix F gives the calculations; the next lemma verifies that the pair satisfies every class constraint.

### E.4 Uniform calibration

**Lemma 15** (Admissibility and nondegeneracy)**.** *The student* (56)*, the sequence teacher* (58)*, and the token teacher* (63) *satisfy Definition 3 for every $T$. Both student and teacher have full support, the student parameter is shared across time, and their next-token distributions differ at every prefix used by either construction.*

*Proof.* First consider probabilities. For either teacher logit $\beta \in \{\delta, -\delta\}$,

$$\min_{x \in \{-1,1\}} q_\beta(x) = \frac{1}{1 + e^{2\delta}}. \tag{68}$$

The right-hand side is at least $c$ whenever $2\delta \leq \log((1-c)/c) = L_c$. Our definition has $\delta \leq L_c/4$, so every teacher probability belongs to $[c, 1-c]$. Every student conditional at $\theta_0$ equals $1/2$ and obeys the same constraint. All these probabilities are strictly positive.

Next consider the sampled quantities. Since $\delta > 0$,

$$0 < a = \log\cosh\delta < \delta, \tag{69}$$

where the strict inequalities follow from $1 < \cosh\delta < e^\delta$. Equations (59) and (64) show that every possible log-ratio is either $a - \delta$ or $a + \delta$. Therefore

$$|r_t| \leq a + \delta < 2\delta \leq \frac{R}{2} < R. \tag{70}$$

Also, $|s_t| = \alpha$ and $\alpha = \frac{1}{2}\min\{S, 1\} \le S$.

Finally, a teacher logit is always either $\delta$ or $-\delta$, whereas the student logit at $\theta_0$ is zero. Because $\delta > 0$, the two conditional distributions differ at every prefix. Equation (69) also shows that neither possible sampled log-ratio is zero. The constants are fixed functions of $(c, R, S)$, completing the uniform-in-$T$ verification. □

Lemma 15 is the quantifier bridge from a single explicit calculation to a lower bound on each supremum, since for every $T$ the appropriate pinned pair is one member of the same admissible class.

### E.5 Gap sensitivity of the lower constants

If $\delta \le 1$, then

$$\frac{\delta^2}{4} \le a = \log\cosh\delta \le \frac{\delta^2}{2}. \tag{71}$$

Hence the displayed token lower coefficient is proportional to $\alpha^2\delta^2$, while the fixed-teacher cubic coefficients are between constant multiples of $\alpha^2\delta^4$. The teacher's total-variation gap from the uniform student is exactly $\frac{1}{2}\tanh\delta$, and $\delta$ itself is an explicit function of $(c, R)$ in (14).

## F Exact Walsh Variance Calculations

For a subset $A \subseteq [T]$, define the Rademacher monomial $X_A = \prod_{i\in A} X_i$, with $X_\varnothing = 1$. Independence and centering give the Walsh orthogonality relation

$$\mathbb{E}[X_A X_B] = \mathbf{1}\{A = B\}. \tag{72}$$

Consequently, if a scalar random variable is written as a finite sum of distinct Walsh monomials, its mean is the coefficient of $X_\varnothing$ and its variance is the sum of squares of all remaining coefficients. This turns the three lower bounds into coefficient counts.

### F.1 Token estimator under the fixed teacher

For $r_t = a - \delta X_t$ and $s_t = \alpha X_t$,

$$g_{\mathrm{tok}} = \alpha a \sum_{t=1}^{T} X_t - \alpha\delta T. \tag{73}$$

The nonconstant Walsh coefficients are $T$ copies of $\alpha a$, hence $\mathrm{Var}(g_{\mathrm{tok}}) = \alpha^2 a^2 T$, proving (60).

### F.2 Full-sequence estimator

Let $W = \sum_{t=1}^{T} X_t$. Under the fixed sequence teacher, $\sum_t r_t = Ta - \delta W$ and $\sum_t s_t = \alpha W$. Since

$$W^2 = T + 2\sum_{1\le i<j\le T} X_i X_j, \tag{74}$$

we obtain

$$\begin{aligned} g_{\mathrm{full}} &= \alpha(Ta - \delta W)W \\ &= -\alpha\delta T + \alpha a T\sum_{i=1}^{T} X_i - 2\alpha\delta \sum_{1\le i<j\le T} X_i X_j. \end{aligned} \tag{75}$$

The constant term gives $\mathbb{E}g_{\mathrm{full}} = -\alpha\delta T$. There are $T$ linear coefficients equal to $\alpha a T$ and $\binom{T}{2}$ quadratic coefficients equal to $-2\alpha\delta$. Applying (72),

$$\begin{aligned} \mathrm{Var}(g_{\mathrm{full}}) &= T(\alpha a T)^2 + \binom{T}{2}(2\alpha\delta)^2 \\ &= \alpha^2\left[a^2T^3 + 2\delta^2 T(T-1)\right] \ge \alpha^2 a^2 T^3. \end{aligned} \tag{76}$$

This proves (61) and its lower bound.

### F.3 REWARD-TO-GO ESTIMATOR

Using $s_t = \alpha X_t$ and $r_u = a - \delta X_u$,

$$\begin{aligned} g_{\mathrm{rtg}} &= \alpha \sum_{t=1}^{T} X_t \sum_{u=t}^{T} (a - \delta X_u) \\ &= -\alpha\delta T + \alpha a \sum_{t=1}^{T} (T - t + 1) X_t - \alpha\delta \sum_{1 \le t < u \le T} X_t X_u. \end{aligned} \tag{77}$$

The diagonal terms $X_t^2 = 1$ produce the constant $-\alpha\delta T$. Each unordered pair occurs once, because the RTG sum contains the product $X_t X_u$ exactly when $t < u$. Hence $\mathbb{E} g_{\mathrm{rtg}} = -\alpha\delta T$, agreeing with the full-sequence mean, and

$$\begin{aligned} \mathrm{Var}(g_{\mathrm{rtg}}) &= \alpha^2 a^2 \sum_{t=1}^{T} (T - t + 1)^2 + \alpha^2 \delta^2 \binom{T}{2} \\ &= \alpha^2 \left[ a^2 \frac{T(T+1)(2T+1)}{6} + \delta^2 \frac{T(T-1)}{2} \right]. \end{aligned} \tag{78}$$

Since $T(T+1)(2T+1)/6 \ge T^3/3$, this is at least $\alpha^2 a^2 T^3/3$. Equation (62) follows.

The linear and quadratic pieces in (77) are orthogonal, so the quadratic term cannot reduce the cubic contribution. The same observation applies to (75); no sign comparison between polynomial pieces is needed.

### F.4 SAMPLED-TOKEN ESTIMATOR

For the persistent-sign teacher, multiply (64) by $s_t = \alpha X_t$. The first position gives

$$r_1 s_1 = \alpha a X_1 - \alpha\delta, \tag{79}$$

and every $t \ge 2$ gives

$$r_t s_t = \alpha(a - \delta X_1 X_t) X_t = \alpha a X_t - \alpha\delta X_1. \tag{80}$$

Summing the $T$ terms yields the degree-one expansion

$$g_{\mathrm{tok}} = -\alpha\delta + \alpha\big(a - \delta(T-1)\big) X_1 + \alpha a \sum_{t=2}^{T} X_t. \tag{81}$$

Thus $\mathbb{E} g_{\mathrm{tok}} = -\alpha\delta$ and Walsh orthogonality gives

$$\mathrm{Var}(g_{\mathrm{tok}}) = \alpha^2 \left[ (a - \delta(T-1))^2 + a^2 (T-1) \right], \tag{82}$$

which is (65).

It remains to extract a horizon-uniform quadratic coefficient. From $0 < a < \delta$ and $T \ge 4$,

$$|a - \delta(T-1)| = \delta(T-1) - a > \delta(T-2) \ge \frac{\delta T}{2}. \tag{83}$$

Keeping only the squared $X_1$ coefficient in (82) proves

$$\mathrm{Var}(g_{\mathrm{tok}}) \ge \frac{\alpha^2 \delta^2}{4} T^2. \tag{84}$$

Together with Lemma 15, these three calculations prove all lower bounds in Theorem 4.

### F.5 ALL THREE ESTIMATORS ON THE PERSISTENT-SIGN INSTANCE

The preceding token calculation already proves (65). For the full form, put $V = \sum_{i=2}^{T} X_i$. The total feedback and score are $Ta - \delta X_1(1 + V)$ and $\alpha(X_1 + V)$, respectively. Expanding $V^2 =$

$(T-1)+2\sum_{2\le i<j\le T} X_iX_j$ gives

$$\frac{g_{\text{full}}}{\alpha} = -\delta + [aT-\delta(T-1)]X_1 + (aT-\delta)\sum_{i=2}^{T} X_i - \delta\sum_{i=2}^{T} X_1X_i - 2\delta\sum_{2\le i<j\le T} X_1X_iX_j. \tag{85}$$

All displayed nonconstant monomials are distinct. Squaring their coefficients proves (66).

For the RTG estimator, reindex by feedback time. At $u=1$ the contribution divided by $\alpha$ is $aX_1-\delta$. For $u\ge 2$,

$$(a-\delta X_1X_u)\sum_{t=1}^{u} X_t = aX_1 + a\sum_{t=2}^{u} X_t - \delta X_u - \delta X_1 - \delta\sum_{t=2}^{u-1} X_1X_tX_u. \tag{86}$$

Summing in $u$ yields

$$\frac{g_{\text{rtg}}}{\alpha} = -\delta + [aT-\delta(T-1)]X_1 + \sum_{i=2}^{T}[a(T-i+1)-\delta]X_i - \delta\sum_{2\le i<j\le T} X_1X_iX_j. \tag{87}$$

Walsh orthogonality gives (67). Since $a,\delta>0$, the leading coefficients of (65)–(67) are positive. The token polynomial has degree two in $T$, the term $(T-1)(aT-\delta)^2$ in (66) is cubic, and $\sum_{k=1}^{T-1}(ak-\delta)^2$ in (67) is a cubic polynomial with leading coefficient $a^2/3$. This proves every assertion of Theorem 5.

## G Proofs for General Alphabets, Moments, and Horizons

Related sequence-level distillation objectives and divergences are studied by Kim & Rush (2016); Wen et al. (2023); Jin et al. (2026). The rollout gradients here use the score-function identity (Schulman et al., 2015; Mohamed et al., 2020); discounted single-position and $K$-step credit rules use different statistics (Liu et al., 2026; Fan et al., 2026).

### G.1 Upper bounds on measurable alphabets

For a measurable output space $(\mathcal{Y},\mathcal{A})$, let $\pi_\theta$ and $q$ be autoregressive probability kernels. At each prefix $h$, let $\nu_h$ dominate the student locally in $\theta$, and let $q(\cdot\mid h)$ dominate $\pi_{\theta_0}(\cdot\mid h)$. The general forms of (1) are

$$r_t = \log\frac{d\pi_{\theta_0}(\cdot\mid H_t)}{dq(\cdot\mid H_t)}(Y_t), \qquad s_t = \nabla_\theta \log\frac{d\pi_\theta(\cdot\mid H_t)}{d\nu_{H_t}}(Y_t)\Big|_{\theta=\theta_0}.$$

These reduce to the finite-vocabulary definitions when $\nu_h$ is counting measure. Definition 1 and its conditional score centering and integrability assumptions remain unchanged. The gradient identities (7) additionally require $P_\theta \ll Q$ near $\theta_0$ and interchange of differentiation with the trajectory and conditional KL integrals, as in Section 3.

The score must match the distribution that generates each token. Sampling from a top-$k$ or top-$p$ truncation while retaining the untruncated score generally violates conditional centering. The score of a normalized truncated distribution can remain centered where its support is locally fixed, but defines a different objective. Detaching the reverse-KL feedback alone preserves the identities above because its direct score term has zero expectation.

For finite vocabularies, exact summation of the local gradient gives $\sum_t \mathbb{E}[r_ts_t\mid\mathcal{F}_{t-1}]=A_T$ from (26). Prefix sampling still makes this estimator random. In the persistent-sign construction, its variance is $\alpha^2\delta^2(T-1)^2$, so quadratic growth remains possible.

**Theorem 16** (Measurable-alphabet upper bounds). *Every process in $\mathcal{G}_T(R,\sigma)$ satisfies*

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{tok}}) \le R^2\sigma^2T^2, \tag{88}$$

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{full}}) \le R^2\sigma^2T^3, \tag{89}$$

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{rtg}}) \le \frac{R^2\sigma^2}{2}T^2(T+1). \tag{90}$$

*No probability floor, finite vocabulary, token independence, or position-specific parameterization is used.*

The proof is in Appendix D.

### G.2 Finite-vocabulary embedding

Theorem 16 and finite-alphabet embeddings give a direct vocabulary result.

**Corollary 17** (Every fixed vocabulary size). *Fix $K \ge 2$, $c \in (0, 1/K)$, and $R, S > 0$. Over $K$-symbol autoregressive pairs for which every conditional probability is at least $c$, $|r_t| \le R$, and $\|s_t\| \le S$, the three raw variance suprema are*

$$\Theta(T^2), \qquad \Theta(T^3), \qquad \Theta(T^3). \tag{91}$$

*All constants depend only on $(K, c, R, S)$. The lower family uses a uniform student and one shared scalar exponential tilt of a feature taking values $-1, 0, 1$.*

*Proof of Corollary 17.* The upper bounds are Theorem 16. We construct lower families. Set $\alpha = \frac{1}{2}\min\{S, 1\}$. Let the $K$ symbols carry a feature $\phi$ with $\phi(1) = 1$, $\phi(2) = -1$, and $\phi(j) = 0$ for $j \ge 3$. Let $\nu$ be uniform, put $p = \mathbb{E}_\nu[\phi^2] = 2/K$, and define

$$\pi_\theta(y) = \frac{\nu(y)e^{\alpha\theta\phi(y)}}{\mathbb{E}_\nu e^{\alpha\theta\phi(Y)}}, \qquad q_\beta(y) = \frac{\nu(y)e^{\beta\phi(y)}}{Z_\beta}, \qquad Z_\beta = \mathbb{E}_\nu e^{\beta\phi(Y)}. \tag{92}$$

At $\theta_0 = 0$, the variables $X_t = \phi(Y_t)$ are i.i.d., symmetric, $\mathbb{E}X_t = 0$, $\mathbb{E}X_t^2 = p$, and $s_t = \alpha X_t$.

Choose

$$\delta_K = \frac{1}{8}\min\left\{R, \log\frac{1}{Kc}\right\} > 0. \tag{93}$$

For every $|\beta| \le 2\delta_K$, $Z_\beta \le e^{|\beta|}$ and

$$q_\beta(y) \ge K^{-1}e^{-2|\beta|} \ge K^{-1}e^{-4\delta_K} > c.$$

Since every component is at least $c$, it is also at most $1 - (K-1)c$. Moreover $|\log Z_\beta - \beta\phi(y)| \le 2|\beta| \le 4\delta_K \le R/2$, and $|s_t| \le \alpha \le S$.

For either sequence estimator, use the fixed teacher $q_{\delta_K}$ and set $a_K = \log Z_{\delta_K} > 0$. Then $r_t = a_K - \delta_K X_t$. In the full form, the odd component $\alpha T a_K \sum_t X_t$ is orthogonal to the even component $-\alpha\delta_K(\sum_t X_t)^2$ by global sign symmetry. Hence

$$\mathrm{Var}(g_{\mathrm{full}}) \ge \alpha^2 a_K^2 T^2 \,\mathrm{Var}\left(\sum_t X_t\right) = \alpha^2 a_K^2 p T^3. \tag{94}$$

In the RTG estimator, the odd component is $\alpha a_K \sum_t (T-t+1)X_t$ and the remaining component is even, so

$$\mathrm{Var}(g_{\mathrm{rtg}}) \ge \alpha^2 a_K^2 p \sum_{k=1}^{T} k^2 \ge \frac{\alpha^2 a_K^2 p}{3}T^3. \tag{95}$$

For the token estimator, use $q_{\delta_K}$ at time one and $q_{\delta_K X_1}$ later. If $a(z) = \log Z_{\delta_K z}$, then for $t \ge 2$ $r_t = a(X_1) - \delta_K X_1 X_t$. Conditioning the token sum on $X_1$ gives

$$\mathbb{E}[g_{\mathrm{tok}} \mid X_1] = \alpha[a_K - \delta_K p(T-1)]X_1 - \alpha\delta_K X_1^2. \tag{96}$$

The odd and centered even terms on the right are orthogonal. Therefore

$$\mathrm{Var}(g_{\mathrm{tok}}) \ge \alpha^2 p[a_K - \delta_K p(T-1)]^2 = \Omega(T^2).$$

All construction constants are fixed in $T$, which completes the three lower bounds. □

### G.3 Fourth-moment envelope

Almost-sure feedback and score bounds can be replaced by joint moment control. For a scalar or vector random variable $Z$, write $\|Z\|_{L_p} = (\mathbb{E}\|Z\|_2^p)^{1/p}$.

**Theorem 18** ($L_4$ envelope)**.** *Suppose for constants $\rho, \kappa$ independent of $t, u, T$ that*

$$\|r_t\|_{L_4} \le \rho, \qquad \|s_t\|_{L_4} \le \kappa, \qquad \|M_u\|_{L_4} \le \kappa\sqrt{u}. \tag{97}$$

*Then*

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{tok}}) \le \rho^2\kappa^2 T^2, \tag{98}$$

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{full}}) \le \rho^2\kappa^2 T^3, \tag{99}$$

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{rtg}}) \le \frac{4\rho^2\kappa^2}{9}(T+1)^3. \tag{100}$$

*In the scalar case, the last two conditions in (97) hold with $\kappa = 2\sigma$ whenever*

$$\mathbb{E}[e^{\lambda s_t} \mid \mathcal{F}_{t-1}] \le e^{\sigma^2\lambda^2/2} \quad \textit{for every } \lambda \in \mathbb{R}. \tag{101}$$

*Scalar scores satisfying this condition and log-ratios with a uniform fourth moment retain the same raw upper powers. Vector scores require the Euclidean-norm conditions in (97) directly.*

*Proof of Theorem 18.* Minkowski's and Hölder's inequalities give

$$\|g_{\mathrm{tok}}\|_{L_2} \le \sum_{t=1}^{T}\|r_t s_t\|_{L_2} \le \sum_{t=1}^{T}\|r_t\|_{L_4}\|s_t\|_{L_4} \le \rho\kappa T, \tag{102}$$

$$\|g_{\mathrm{full}}\|_{L_2} \le \left\|\sum_{u=1}^{T} r_u\right\|_{L_4} \|M_T\|_{L_4} \le \rho T\,\kappa\sqrt{T}. \tag{103}$$

Using (30),

$$\begin{aligned}\|g_{\mathrm{rtg}}\|_{L_2} &\le \sum_{u=1}^{T}\|r_u\|_{L_4}\|M_u\|_{L_4} \le \rho\kappa\sum_{u=1}^{T}\sqrt{u} \\ &\le \frac{2\rho\kappa}{3}(T+1)^{3/2},\end{aligned} \tag{104}$$

where the last inequality compares the increasing summand with $\int_1^{T+1}\sqrt{x}\,dx$. Squaring and using $\mathrm{Var}_{\mathrm{tr}}(G) \le \|G\|_{L_2}^2$ proves the theorem's bounds.

Under (101), iterated conditioning yields $\mathbb{E}e^{\lambda M_u} \le e^{u\sigma^2\lambda^2/2}$. Chernoff bounds on both tails give $\Pr\{|M_u| \ge x\} \le 2e^{-x^2/(2u\sigma^2)}$. Tail integration then gives

$$\mathbb{E}|M_u|^4 = \int_0^\infty 4x^3\,\Pr\{|M_u| \ge x\}\,dx \le 16u^2\sigma^4. \tag{105}$$

The same calculation with $u = 1$, conditionally and then in expectation, bounds $\mathbb{E}|s_t|^4$ by $16\sigma^4$. Thus $\kappa = 2\sigma$ is valid. □

### G.4 Stopping times

For later stopping-time results, a discrete stopping time $\tau \le T$ satisfies $\{\tau \ge t\} \in \mathcal{F}_{t-1}$. Thus its alive indicator is predictable and preserves conditional score centering.

Variable generation length enters through predictable survival.

**Theorem 19** (Capped stopping time)**.** *Let $\tau \le T$ be a stopping time and let the three estimators be defined by retaining exactly the indices $t, u \le \tau$. Under (8),*

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{tok}}^{\tau}) \le R^2\sigma^2 T\,\mathbb{E}\tau, \tag{106}$$

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{full}}^{\tau}) \le R^2\sigma^2 T^2\,\mathbb{E}\tau, \tag{107}$$

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{rtg}}^{\tau}) \le R^2\sigma^2 T\,\mathbb{E}\left[T\tau - \frac{\tau(\tau-1)}{2}\right]. \tag{108}$$

*If additionally $\|s_t\| \le S$ almost surely, then $\mathrm{Var}_{\mathrm{tr}}(g^\tau_{\mathrm{tok}}) \le R^2 S^2 \mathbb{E}[\tau^2]$. The bounds depend on the cap $T$. For deterministic $\tau = T$, they recover the powers in Theorem 16, including its RTG constant.*

*Proof of Theorem 19.* Put $A_t = \mathbf{1}\{\tau \ge t\}$. Since $A_t$ is $\mathcal{F}_{t-1}$-measurable, $A_t s_t$ is a martingale difference and

$$\mathbb{E}\left\|\sum_{t=1}^{T} A_t s_t\right\|^2 = \sum_{t=1}^{T} \mathbb{E}[A_t \|s_t\|^2] \le \sigma^2 \mathbb{E}\tau. \tag{109}$$

For tokens,

$$\mathbb{E}\|g^\tau_{\mathrm{tok}}\|^2 \le TR^2 \sum_{t=1}^{T} \mathbb{E}[A_t \|s_t\|^2] \le R^2\sigma^2 T\, \mathbb{E}\tau. \tag{110}$$

If scores are bounded, the pathwise inequality $\|g^\tau_{\mathrm{tok}}\| \le RS\tau$ gives the stated alternative.

For the full form, $|\sum_t A_t r_t| \le RT$ and (109) imply

$$\mathbb{E}\|g^\tau_{\mathrm{full}}\|^2 \le R^2 T^2 \sigma^2 \mathbb{E}\tau. \tag{111}$$

Let $M^\tau_u = \sum_{t=1}^{u} A_t s_t = M_{u\wedge\tau}$. Reindexing the RTG sum and Cauchy–Schwarz across the at most $T$ feedback times give

$$\begin{aligned}
\mathbb{E}\|g^\tau_{\mathrm{rtg}}\|^2 &\le TR^2 \sum_{u=1}^{T} \mathbb{E}[A_u \|M^\tau_u\|^2] \\
&\le TR^2 \sum_{u=1}^{T} \mathbb{E}\|M_{u\wedge\tau}\|^2 \le TR^2\sigma^2 \sum_{u=1}^{T} \mathbb{E}[\min\{u,\tau\}] \\
&\le R^2\sigma^2 T\, \mathbb{E}\left[T\tau - \frac{\tau(\tau-1)}{2}\right].
\end{aligned} \tag{112}$$

Here the stopped isometry applies at each deterministic cap $u$, and $\sum_{u=1}^{T} \min\{u,\tau\} = T\tau - \tau(\tau - 1)/2$. Trace variance is no larger than each displayed second moment. □

### G.5 Discounting and clipping

Discounting and clipping admit sharp statements. For $0 \le \gamma < 1$, set

$$g_\gamma = \sum_{t=1}^{T} s_t \sum_{u=t}^{T} \gamma^{u-t} r_u. \tag{113}$$

**Theorem 20** (Fixed discount and clipping)**.** *For every process in $\mathcal{G}_T(R,\sigma)$,*

$$\mathrm{Var}_{\mathrm{tr}}(g_\gamma) \le \frac{R^2\sigma^2}{1-\gamma^2} T^2. \tag{114}$$

*For each fixed $\gamma < 1$, bounded binary full-support pairs attain $\Omega(T^2)$, so the worst-case discounted RTG variance is exactly quadratic in the horizon. If an arbitrary adapted feedback process is clipped as $r^{(C)}_t = \mathrm{clip}(r_t, -C, C)$ with $C > 0$, the three estimators formed from $r^{(C)}_t$ have upper powers $(2,3,3)$ with $R = C$, and bounded witnesses strictly inside the clipping interval attain the same powers.*

*Proof of Theorem 20.* Reindex (113) as

$$g_\gamma = \sum_{u=1}^{T} r_u N_u, \qquad N_u = \sum_{t=1}^{u} \gamma^{u-t} s_t. \tag{115}$$

The deterministic weights preserve score orthogonality, so

$$\mathbb{E}\|N_u\|^2 \le \sigma^2 \sum_{k=0}^{u-1} \gamma^{2k} \le \frac{\sigma^2}{1-\gamma^2}.$$

Vector Cauchy–Schwarz across $u$ proves (114).

For the lower bound use the persistent-sign binary teacher. In its Walsh expansion, the coefficient of $X_1$ in $g_\gamma$ is

$$\alpha \left[ a \sum_{k=0}^{T-1} \gamma^k - \delta(T-1) \right]. \tag{116}$$

The geometric sum is bounded uniformly in $T$, whereas the second term is linear. The squared coefficient is therefore $\Omega(T^2)$ and cannot be cancelled by other, orthogonal Walsh monomials.

For clipping, $|r_t^{(C)}| \le C$ gives all upper bounds by Theorem 16. Choose the binary witness with $\delta_C = \frac{1}{4}\min\{C, L_c\}$. Its possible ratios satisfy $|r_t| < 2\delta_C < C$, so clipping is inactive and the lower calculations in Appendix F apply unchanged. □

The quadratic discounting rate holds for fixed $\gamma < 1$. For $\gamma_T = 1 - c/T$ with fixed $c > 0$ and sufficiently large $T$, $(1-\gamma_T^2)^{-1} = \Theta(T)$, so (114) gives an $O(T^3)$ upper bound for this discount schedule that depends on the horizon.

### G.6 Gap constants

For $0 \le x \le 1$, Hoeffding's lemma for a Rademacher variable gives $\log\cosh x \le x^2/2$. Also $\cosh x \ge 1 + x^2/2$ and $\log(1+z) \ge z/2$ for $0 \le z \le 1/2$, so $\log\cosh x \ge x^2/4$. This proves (71). Finally, $q_\delta(1) = (1+\tanh\delta)/2$, hence its total-variation distance from the uniform binary distribution is $\frac{1}{2}\tanh\delta$.

## H Proofs for Prefix Controls and Dependence

**Related variance reduction methods.** Action-dependent baselines have been proposed and critically reassessed (Wu et al., 2018; Tucker et al., 2018). Other approaches use critic control variates (Gu et al., 2017), learned black-box controls (Grathwohl et al., 2018), or corrections across successive policy updates (Papini et al., 2018; Xu et al., 2020). Our theorem instead restricts each control to a scalar coefficient measurable from the current prefix.

### H.1 Predictable projection and the binary oracle

*Proof of Theorem 6.* Conditional Cauchy–Schwarz gives

$$N_t^2 \le \mathbb{E}[\|G - \mathbb{E}G\|^2 \mid \mathcal{F}_{t-1}] D_t. \tag{117}$$

Conditional score centering also gives $N_t = \mathbb{E}[\langle G, s_t\rangle \mid \mathcal{F}_{t-1}]$, so computing this numerator does not require $\mathbb{E}G$. The bound implies $N_t = 0$ on $\{D_t = 0\}$, and the convention in (21) is well defined.

For $i < j$, $b_i s_i$ is $\mathcal{F}_{j-1}$-measurable and

$$\mathbb{E}\langle b_i s_i, b_j s_j\rangle = \mathbb{E}\,\langle b_i s_i, b_j \mathbb{E}[s_j \mid \mathcal{F}_{j-1}]\rangle = 0. \tag{118}$$

Also $\mathbb{E}[b_t s_t] = 0$ and

$$\mathbb{E}\langle G - \mathbb{E}G, b_t s_t\rangle = \mathbb{E}[b_t N_t].$$

Expanding the centered square therefore yields

$$\begin{aligned} \mathrm{Var}_{\mathrm{tr}}(G^b) &= \mathrm{Var}_{\mathrm{tr}}(G) + \sum_t \mathbb{E}[D_t b_t^2] - 2\sum_t \mathbb{E}[b_t N_t] \\ &= \mathrm{Var}_{\mathrm{tr}}(G) - \sum_t \mathbb{E}\left[\frac{N_t^2}{D_t}\mathbf{1}\{D_t > 0\}\right] + \sum_t \mathbb{E}[D_t (b_t - b_t^\star)^2]. \end{aligned} \tag{119}$$

All terms are finite by (117), finite $T$, and truncation if needed; monotone convergence removes the truncation. The last sum proves optimality and the stated uniqueness. □

**Approximation and estimation against a reference class.** More generally, for any admissible reference $b^{\mathcal{B}}$ in a chosen function class, let

$$\mathcal{A}_T = \sum_t \mathbb{E}[D_t(b_t^{\mathcal{B}} - b_t^\star)^2], \qquad \mathcal{S}_T = \sum_t \mathbb{E}[D_t(\widehat{b}_t - b_t^{\mathcal{B}})^2]. \tag{120}$$

Then

$$\mathrm{Var}_{\mathrm{tr}}(G^{\widehat{b}}) \le \mathrm{Var}_{\mathrm{tr}}(G^{b^\star}) + 2\mathcal{A}_T + 2\mathcal{S}_T. \tag{121}$$

Thus a finite-capacity prefix control is governed by a transparent oracle, approximation, and estimation decomposition.

*Proof of Corollary 7.* If auxiliary training data are used, enlarge every filtration by their sigma-field. Independence from the current rollout leaves $D_t$, $N_t$, and $b_t^\star$ unchanged, while $\widehat{b}_t$ is predictable in the enlarged filtration. Applying (22) once with $b = \widehat{b}$ and once with $b = b^\star$, and subtracting, gives (24).

A baseline formed from other independent rollouts at the same prompt can also be included in this enlarged information set (Kool et al., 2019). Reusing rollouts across controls can correlate the resulting updates, so the batch variance requires the covariance terms in (34).

The bound (25) follows from $D_t \le \sigma^2$ and summation. Finally,

$$(\widehat{b}_t - b_t^\star)^2 \le 2(\widehat{b}_t - b_t^{\mathcal{B}})^2 + 2(b_t^{\mathcal{B}} - b_t^\star)^2.$$

Multiplying by $D_t$, taking expectations, summing in $t$, and using (24) proves (121). □

**Corollary 21** (Complete scalar-binary oracle)**.** *Suppose $\mathcal{F}_T$ is generated by binary tokens, $G$ is scalar and $\mathcal{F}_T$-measurable, the initial context is fixed, and $D_t > 0$ almost surely. Then prefix controls exist for which*

$$G - \sum_{t=1}^{T} b_t s_t = \mathbb{E}G \quad a.s., \qquad \inf_b \mathrm{Var}(G^b) = 0. \tag{122}$$

*In particular, the minimax variance over unrestricted oracle prefix controls has no positive horizon exponent in this scalar binary class.*

The corollary follows because a conditionally centered function of one binary draw is one-dimensional. Each Doob increment $\mathbb{E}[G \mid \mathcal{F}_t] - \mathbb{E}[G \mid \mathcal{F}_{t-1}]$ is therefore a predictable multiple of $s_t$. The oracle coefficients require the same conditional information that defines the Doob martingale, which places this result at the level of representation. Corollary 7 supplies the criterion that a realizable learned control has to meet. For larger vocabularies or vector statistics, the residual consists of the components of the Doob increments outside the scalar control class. Statistics whose increments lie in that class admit zero variance.

*Proof of Corollary 21.* Let $\Delta_t = \mathbb{E}[G \mid \mathcal{F}_t] - \mathbb{E}[G \mid \mathcal{F}_{t-1}]$. Conditional on any prefix, $\Delta_t$ and the nonzero scalar $s_t$ are both centered functions of one binary random variable. The vector space of such functions is one-dimensional, so there is an $\mathcal{F}_{t-1}$-measurable scalar $b_t$ with $\Delta_t = b_t s_t$. The Doob telescoping identity and fixed initial context give

$$\sum_{t=1}^{T} b_t s_t = \sum_{t=1}^{T} \Delta_t = G - \mathbb{E}G.$$

Square integrability follows from orthogonality of the Doob increments. Thus the controlled statistic is the constant $\mathbb{E}G$. □

### H.2 Time-only control lower bound

Let $\mathcal{B}_{\mathrm{det}} = \{(b_1, \dots, b_T) : b_t \in \mathbb{R}\}$ and define

$$W_e^{\mathrm{det}}(T) = \sup_{P \in \mathcal{C}_T(c,R,S)} \inf_{b \in \mathcal{B}_{\mathrm{det}}} \mathrm{Var}_{\mathrm{tr}}\left(g_e - \sum_{t=1}^{T} b_t s_t\right). \tag{123}$$

**Theorem 22** (Sharp time-only control barrier). *For $e \in \{\text{full}, \text{rtg}\}$ and fixed $(c, R, S)$,*

$$W_e^{\det}(T) = \Theta(T^3). \tag{124}$$

*The lower family is a two-logit binary pair whose protected Walsh coefficients survive every time-only control.*

Thus time-only centering cannot uniformly remove the cubic power, while an unrestricted prefix control removes all of it in the scalar binary case (Appendix H). The next result locates the intermediate, commonly used case.

*Proof of Theorem 22.* The upper bound follows by choosing $b_t = 0$ and applying Theorem 4. For the converse, use the Rademacher student and let the teacher at time one have logit $\delta$. For $u \geq 2$, use logit

$$\beta_{X_1} = \begin{cases} \delta, & X_1 = 1, \\ 2\delta, & X_1 = -1. \end{cases} = \beta_0 + \beta_1 X_1, \quad \beta_0 = \frac{3\delta}{2}, \quad \beta_1 = -\frac{\delta}{2}.$$

Let

$$a_{X_1} = \log\cosh(\beta_{X_1}) = a_0 + a_1 X_1,$$
$$a_0 = \frac{\log\cosh\delta + \log\cosh(2\delta)}{2}, \qquad a_1 = \frac{\log\cosh\delta - \log\cosh(2\delta)}{2} < 0.$$

The maximum teacher-logit magnitude is $2\delta$. Because $4\delta \leq L_c$ and $\log\cosh(2\delta) + 2\delta < 4\delta \leq R$, the calibration proof of Lemma 15 applies; the family is in $\mathcal{C}_T(c, R, S)$.

Consider a monomial $X_1 X_j$, $j \geq 2$. In the RTG estimator, the predictable feedback component $a_{X_1}$ contributes $\alpha a_1(T - j + 1)$ to its coefficient. The action-dependent feedback at time $j$, when multiplied by the score at time one, contributes $-\alpha\beta_0$. No other term produces this monomial. Hence its coefficient is

$$\alpha[a_1(T - j + 1) - \beta_0]. \tag{125}$$

Since $a_1 < 0 < \beta_0$, the sum of squared protected coefficients is at least

$$\alpha^2 a_1^2 \sum_{k=1}^{T-1} k^2 = \Omega(T^3). \tag{126}$$

For the full form, the coefficient on every $X_1 X_j$, $j \geq 2$, is

$$\alpha[a_1(T - 1) - \beta_0 - \delta]. \tag{127}$$

The three terms arise respectively from the $T - 1$ later normalization constants, the later action feedback paired with $s_1$, and the first action feedback paired with $s_j$. Its magnitude is at least $\alpha|a_1|(T - 1)$, so the $T - 1$ protected coefficients contribute at least $\alpha^2 a_1^2 (T - 1)^3$.

For deterministic $b_t$, the control $\sum_t b_t s_t = \alpha \sum_t b_t X_t$ contains only degree-one Walsh monomials. It changes none of (125) or (127). Walsh orthogonality therefore proves the lower bounds after the infimum over all deterministic controls. □

### H.3 Predictable feedback means and anchored centering

Let

$$m_u = \mathbb{E}[r_u \mid \mathcal{F}_{u-1}], \qquad \eta_u = r_u - m_u, \qquad P_T = \sum_{u=1}^{T} m_u M_u. \tag{128}$$

For an OPD log-ratio, $m_u$ is the local conditional reverse KL.

**Theorem 23** (RTG predictable-mean decomposition). *For every process in $\mathcal{G}_T(R, \sigma)$,*

$$|\mathrm{sd}_{\mathrm{tr}}(g_{\mathrm{rtg}}) - \mathrm{sd}_{\mathrm{tr}}(P_T)| \leq 4R\sigma T. \tag{129}$$

*Hence $\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{rtg}})$ is cubic exactly when $\mathrm{Var}_{\mathrm{tr}}(P_T)$ is cubic; if $\mathrm{Var}_{\mathrm{tr}}(P_T) = O(T^2)$, then $\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{rtg}}) = O(T^2)$.*

*If $m_1, \dots, m_T$ are deterministic numbers, set*

$$b_t^{\mathrm{mean}} = \sum_{u=t}^{T} m_u. \tag{130}$$

*These coefficients define a valid time-only control and*

$$\mathrm{Var}_{\mathrm{tr}}\left(g_{\mathrm{rtg}} - \sum_{t=1}^{T} b_t^{\mathrm{mean}} s_t\right) \le 16R^2\sigma^2T^2. \tag{131}$$

*Thus a prefix-invariant local-KL schedule admits a valid time-only control with quadratic RTG variance.*

*Proof of Theorem 23.* The reindexed RTG estimator has the pathwise decomposition

$$g_{\mathrm{rtg}} = P_T + Q_T, \qquad Q_T = \sum_{u=1}^{T} \eta_u M_u. \tag{132}$$

Define

$$d_u = \mathbb{E}[\eta_u s_u \mid \mathcal{F}_{u-1}], \qquad q_u = \eta_u M_u - d_u. \tag{133}$$

Because $\mathbb{E}[\eta_u \mid \mathcal{F}_{u-1}] = 0$ and $M_{u-1}$ is predictable, $\mathbb{E}[q_u \mid \mathcal{F}_{u-1}] = 0$. Also $|\eta_u| \le 2R$, and conditional Cauchy–Schwarz gives

$$\|d_u\| \le \sqrt{\mathbb{E}[\eta_u^2 \mid \mathcal{F}_{u-1}]\mathbb{E}[\|s_u\|^2 \mid \mathcal{F}_{u-1}]} \le 2R\sigma. \tag{134}$$

Conditional variance is bounded by conditional second moment, hence

$$\mathbb{E}\|q_u\|^2 \le \mathbb{E}[\eta_u^2\|M_u\|^2] \le 4R^2\mathbb{E}\|M_u\|^2 \le 4R^2\sigma^2 u. \tag{135}$$

The $q_u$ are orthogonal martingale differences. Thus, using $Q_T = \sum_u d_u + \sum_u q_u$,

$$\begin{aligned}\mathrm{sd}_{\mathrm{tr}}(Q_T) &\le \left(\mathbb{E}\left\|\sum_u d_u\right\|^2\right)^{1/2} + \left(\mathbb{E}\left\|\sum_u q_u\right\|^2\right)^{1/2} \\ &\le 2R\sigma T + \left(4R^2\sigma^2\sum_{u=1}^{T} u\right)^{1/2} \le 4R\sigma T.\end{aligned} \tag{136}$$

The reverse triangle inequality for centered $L_2$ vectors applied to (132) proves (129).

If the $m_u$ are deterministic, finite-sum reindexing gives

$$P_T = \sum_{u=1}^{T} m_u M_u = \sum_{t=1}^{T}\left(\sum_{u=t}^{T} m_u\right) s_t. \tag{137}$$

This is exactly the control in (130). The adjusted estimator equals $Q_T$, and squaring (136) proves (131). □

**Corollary 24** (A learned mean control preserves the quadratic rate)**.** *Under the deterministic-$(m_u)$ conditions of Theorem 23, let $\widehat{b}$ be trained on auxiliary data independent of the evaluated rollout and suppose*

$$\mathcal{S}_T^{\mathrm{mean}} = \sum_{t=1}^{T} \mathbb{E}[D_t(\widehat{b}_t - b_t^{\mathrm{mean}})^2] = O(T^2). \tag{138}$$

*Then*

$$\mathrm{Var}_{\mathrm{tr}}\left(g_{\mathrm{rtg}} - \sum_{t=1}^{T} \widehat{b}_t s_t\right) \le 32R^2\sigma^2T^2 + 2\mathcal{S}_T^{\mathrm{mean}} = O(T^2). \tag{139}$$

*This preserves the estimand because both choices of coefficients yield conditionally mean-zero prefix controls for the RTG estimator.*

*Proof of Corollary 24.* Use $G = g_{\mathrm{rtg}}$ and $b^{\mathcal{B}} = b^{\mathrm{mean}}$ in (121) and Corollary 7. The exact projection identity gives

$$\mathcal{A}_T = \mathrm{Var}_{\mathrm{tr}}(G^{b^{\mathrm{mean}}}) - \mathrm{Var}_{\mathrm{tr}}(G^{b^\star}).$$

Consequently,

$$\begin{aligned}\mathrm{Var}_{\mathrm{tr}}(G^{\widehat{b}}) &\le \mathrm{Var}_{\mathrm{tr}}(G^{b^\star}) + 2\mathcal{A}_T + 2\mathcal{S}_T^{\mathrm{mean}} \\ &\le 2\,\mathrm{Var}_{\mathrm{tr}}(G^{b^{\mathrm{mean}}}) + 2\mathcal{S}_T^{\mathrm{mean}} \\ &\le 32R^2\sigma^2T^2 + 2\mathcal{S}_T^{\mathrm{mean}},\end{aligned}$$

where the final step is (131). □

Centering each local feedback value requires an anchored term when $m_u$ depends on the prefix.

**Lemma 25** (Anchored local centering)**.** *With $M_0 = 0$,*

$$\mathbb{E}g_{\mathrm{rtg}} = \mathbb{E}\sum_{u=1}^{T}\left[\eta_u M_u + m_u M_{u-1}\right] = \mathbb{E}\sum_{u=1}^{T}\left[\eta_u s_u + m_u M_{u-1}\right]. \tag{140}$$

*Therefore subtracting $m_u$ from $r_u$ without the basepoint term $m_u M_{u-1}$ changes the target by $\sum_u \mathbb{E}[m_u M_{u-1}]$. Equation (140) is the unbiased locally centered form for arbitrary prefix-dependent $m_u$. Pathwise, its first sum equals $g_{\mathrm{rtg}} - \sum_u m_u s_u$, the prefix control with coefficients $b_u = m_u$.*

*Proof of Lemma 25.* For each $u$,

$$\begin{aligned}\mathbb{E}[r_u M_u] &= \mathbb{E}[(m_u + \eta_u)(M_{u-1} + s_u)] \\ &= \mathbb{E}[m_u M_{u-1}] + \mathbb{E}[\eta_u s_u],\end{aligned} \tag{141}$$

because $\mathbb{E}[m_u s_u \mid \mathcal{F}_{u-1}] = 0$ and $\mathbb{E}[\eta_u M_{u-1} \mid \mathcal{F}_{u-1}] = 0$. Similarly $\mathbb{E}[\eta_u M_u] = \mathbb{E}[\eta_u s_u]$. Summing (141) and using $g_{\mathrm{rtg}} = \sum_u r_u M_u$ proves both identities. □

### H.4 Explicit control effect on the common witnesses

On the persistent-sign instance of Theorem 5, use

$$g^c_{\mathrm{tok}} = g_{\mathrm{tok}} - aM_T, \quad g^c_{\mathrm{full}} = g_{\mathrm{full}} - TaM_T, \quad g^c_{\mathrm{rtg}} = g_{\mathrm{rtg}} - \sum_{t=1}^{T} a(T-t+1)s_t. \tag{142}$$

All subtracted terms have zero mean.

**Proposition 26** (Exact centered same-instance variances)**.** *For the common persistent-sign pair,*

$$\mathrm{Var}(g^c_{\mathrm{tok}}) = \alpha^2\delta^2(T-1)^2, \tag{143}$$

$$\mathrm{Var}(g^c_{\mathrm{full}}) = 3\alpha^2\delta^2(T-1)^2, \tag{144}$$

$$\mathrm{Var}(g^c_{\mathrm{rtg}}) = \frac{\alpha^2\delta^2}{2}(T-1)(3T-2). \tag{145}$$

*Thus the same elementary centering that preserves every mean changes both raw sequence powers from cubic to quadratic while the adversarial token power remains quadratic. For the fixed teacher of Proposition 14, the analogous variances are $0$, $2\alpha^2\delta^2T(T-1)$, and $\alpha^2\delta^2T(T-1)/2$.*

*Proof of Proposition 26.* Subtracting the controls in (142) deletes every $a$-linear term from the persistent-sign Walsh expansions. Equation (81) becomes

$$g^c_{\mathrm{tok}} = -\alpha\delta - \alpha\delta(T-1)X_1,$$

giving the first variance. From (85), the remaining nonconstant coefficients are

$$-\alpha\delta(T-1) \text{ on } X_1, \quad -\alpha\delta \text{ on each } X_i \text{ and } X_1X_i, \quad -2\alpha\delta \text{ on each } X_1X_iX_j$$

for $2 \le i < j \le T$. Their squared sum is

$$\alpha^2\delta^2\left[(T-1)^2 + 2(T-1) + 4\binom{T-1}{2}\right] = 3\alpha^2\delta^2(T-1)^2.$$

From (87), the remaining coefficients are $-\alpha\delta(T-1)$ on $X_1$, $-\alpha\delta$ on each $X_i$ for $i \geq 2$, and $-\alpha\delta$ on each $X_1X_iX_j$. Their squared sum is

$$\alpha^2\delta^2\left[(T-1)^2 + (T-1) + \binom{T-1}{2}\right] = \frac{\alpha^2\delta^2}{2}(T-1)(3T-2).$$

For the fixed teacher, deleting the degree-one $a$ components from (73), (75), and (77) leaves respectively no random coefficient, $\binom{T}{2}$ coefficients of magnitude $2\alpha\delta$, and $\binom{T}{2}$ coefficients of magnitude $\alpha\delta$. This gives the final three formulas. □

### H.5 Token dependence

*Proof of Theorem 8.* Let $\xi_t = z_t - \mu_t$. It is a vector martingale difference and

$$\mathbb{E}\left\|\sum_{t=1}^{T}\xi_t\right\|^2 = \sum_{t=1}^{T}\mathbb{E}\|\xi_t\|^2 \leq \sum_{t=1}^{T}\mathbb{E}\|z_t\|^2 \leq R^2\sigma^2 T. \tag{146}$$

Since $g_{\mathrm{tok}} = A_T + \sum_t \xi_t$, the reverse triangle inequality for centered $L_2$ vectors proves (27).

Under (28),

$$\mathrm{Var}_{\mathrm{tr}}(A_T) = \sum_{i,j}\mathbb{E}\langle\bar{\mu}_i, \bar{\mu}_j\rangle \leq T\left[\Gamma(0) + 2\sum_{k\geq 1}\Gamma(k)\right]. \tag{147}$$

Combining with (27) gives $O(T)$ token variance. The fixed initial context gives $\bar{\mu}_1 = 0$. Under the positive covariance condition for $2 \leq i \neq j \leq T$, $\mathrm{Var}_{\mathrm{tr}}(A_T) \geq \kappa(T-1)(T-2)$. The lower side of (27) then gives $\Omega(T^2)$ token variance for all sufficiently large $T$, while Theorem 16 gives $O(T^2)$. Finally, if $\mathrm{Var}_{\mathrm{tr}}(A_T) = \Theta(T^2)$, the $O(\sqrt{T})$ standard-deviation gap in (27) gives $\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{tok}}) = \Theta(T^2)$. Conversely, if $\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{tok}}) = \Theta(T^2)$, the same inequality gives $\mathrm{sd}_{\mathrm{tr}}(A_T) = \mathrm{sd}_{\mathrm{tr}}(g_{\mathrm{tok}}) + O(\sqrt{T}) = \Theta(T)$, hence $\mathrm{Var}_{\mathrm{tr}}(A_T) = \Theta(T^2)$. □

The same fourth-moment argument gives a companion guarantee for the full form under a deterministic trajectory control, recorded here because it shares the drift objects just introduced.

**Corollary 27** (Full-form mean centering)**.** *Suppose $m_1, \ldots, m_T$ are deterministic and, for constants $\rho_0, \kappa_0$ independent of $T$,*

$$\left\|\sum_{u=1}^{T}\eta_u\right\|_{L_4} \leq \rho_0\sqrt{T}, \qquad \|M_T\|_{L_4} \leq \kappa_0\sqrt{T}. \tag{148}$$

*With the common control coefficient $B = \sum_u m_u$,*

$$\mathrm{Var}_{\mathrm{tr}}(g_{\mathrm{full}} - BM_T) \leq \rho_0^2\kappa_0^2 T^2. \tag{149}$$

*Proof of Corollary 27.* Since $B = \sum_u m_u$ is deterministic, $BM_T = \sum_t Bs_t$ is a valid control. Moreover,

$$g_{\mathrm{full}} - BM_T = \left(\sum_{u=1}^{T}\eta_u\right)M_T. \tag{150}$$

Hölder's inequality and (148) give $\|g_{\mathrm{full}} - BM_T\|_{L_2} \leq \rho_0\kappa_0 T$. Squaring and using trace variance no larger than the uncentered second moment proves (149). □

## I Proofs of Statistical and Optimization Consequences

*Proof of Proposition 10.* Let $Z_i = G_i - \mu$. Independence and centering eliminate all cross inner products:

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}Z_i\right\|^2 = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\|Z_i\|^2 = \frac{V}{n}.$$

Markov's inequality applied to the squared norm proves the probability bound. Under the exchangeable covariance assumption, the same expansion has $n$ diagonal terms equal to $V$ and $n(n-1)$ ordered off-diagonal terms equal to $\kappa$, which gives (34). □

*Proof of Theorem 11.* For every random vector $G$ and deterministic scalar $c$, $\mathrm{Var}_{\mathrm{tr}}(cG) = c^2 \mathrm{Var}_{\mathrm{tr}}(G)$. Therefore $\widetilde{V}_e(T) = T^{-2} V_e(T)$. Dividing each lower and upper bound in Theorem 4 by $T^2$ gives the stated constants and powers; for the RTG upper bound, $(T+1)/2 \le T$. □

Equations (37) and (38) follow because the exact MSE of the sample mean is $V_e(T)/n$ or $\widetilde{V}_e(T)/n$. The universal upper bounds are sufficient, and the positive lower constants in Theorems 4 and 11 are necessary on the displayed witnesses. Prefix and suffix scans are linear in $T$, proving the two token-oracle work statements.

*Proof of Proposition 12.* On the fixed-teacher witness, the three means are all $-\alpha\delta T$. Dividing the exact variances in Proposition 14 by $\alpha^2\delta^2 T^2$ gives (41)–(43). On the persistent-sign witness, Appendix F gives $\mathbb{E} g_{\mathrm{tok}} = -\alpha\delta$ and

$$\mathrm{Var}(g_{\mathrm{tok}}) = \alpha^2\big[(a-\delta(T-1))^2 + a^2(T-1)\big] = \Theta(T^2).$$

Its relative variance is therefore $\Theta(T^2)$. Finally, both numerator and denominator in (40) are multiplied by $c^2$ when $G$ is multiplied by a nonzero deterministic scalar $c$. □

*Proof of Theorem 13.* Smoothness gives, conditionally on $\theta_k$,

$$\begin{aligned}\mathbb{E}[F(\theta_{k+1}) \mid \theta_k] &\le F(\theta_k) - \eta\langle \nabla F(\theta_k), \mathbb{E}[G_k \mid \theta_k]\rangle + \frac{L\eta^2}{2}\mathbb{E}[\|G_k\|^2 \mid \theta_k] \\ &\le F(\theta_k) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla F(\theta_k)\|^2 + \frac{L\eta^2}{2}\mathcal{V}_T. \end{aligned} \tag{151}$$

Here unbiasedness and $\mathbb{E}\|G_k\|^2 = \|\nabla F(\theta_k)\|^2 + \mathbb{E}\|G_k - \nabla F(\theta_k)\|^2$ were used. Since $\eta \le 1/L$, the gradient coefficient is at least $\eta/2$. Take full expectation, sum from 0 to $K-1$, telescope, and use $\mathbb{E}F(\theta_K) \ge F_\star$:

$$\frac{\eta}{2}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla F(\theta_k)\|^2 \le \Delta + \frac{L\eta^2 K}{2}\mathcal{V}_T.$$

Division proves (45). The choice (46) makes each term on its right at most $\varepsilon^2/2$. □

## J Finite-Horizon and Construction Details

This appendix records the quantifiers, finite-horizon constants, and variance-producing components used by the lower bounds.

### J.1 Order of quantifiers

The constants $c, R, S$ are chosen first and remain fixed. Equation (14) then selects fixed positive $\alpha, \delta, a$. For every integer $T \ge 4$, the student and relevant teacher from Appendix E satisfy the same bounds with these constants. The distributional rule is unchanged as $T$ grows; only the number of sampled conditionals changes.

The lower-bound inference has the following form. If $\mathcal{C}_T(c, R, S)$ denotes the admissible class at horizon $T$, let $P_{e,T}$ denote the corresponding pinned pair, which uses the persistent-sign teacher for $e = \mathrm{tok}$ and the fixed teacher for either sequence label. Lemma 15 gives $P_{e,T} \in \mathcal{C}_T(c, R, S)$. Therefore, for each estimator label $e \in \{\mathrm{tok}, \mathrm{full}, \mathrm{rtg}\}$,

$$V_e(T) = \sup_{P \in \mathcal{C}_T(c,R,S)} \mathrm{Var}_{\mathrm{tr},P}(g_e) \ge \mathrm{Var}_{\mathrm{tr},P_{e,T}}(g_e). \tag{152}$$

The exact expressions in Appendix F evaluate the right-hand side, and membership alone yields the required supremum lower bound.

### J.2 Centering constants versus variance-producing components

For the sequence witness, both forms have mean $-\alpha\delta T$. This agrees with the general equality proved in Appendix D, although their variance polynomials differ:

$$\mathbb{E}g_{\text{full}} = \mathbb{E}g_{\text{rtg}} = -\alpha\delta T, \tag{153}$$

$$\text{Var}(g_{\text{full}}) \geq \alpha^2 a^2 T^3, \qquad \text{Var}(g_{\text{rtg}}) \geq \frac{\alpha^2 a^2}{3} T^3. \tag{154}$$

The mean arises from diagonal products $X_t^2$. The cubic variance instead comes from the constant $a$ in each log-ratio multiplying a random score sum. Keeping these components separate prevents an uncentered second moment from being mistaken for the lower-bound variance.

For the token witness, $\mathbb{E}g_{\text{tok}} = -\alpha\delta$. The variance-producing $X_1$ coefficient in (81) is $\alpha(a-\delta(T-1))$. Equivalently, for every $t \geq 2$,

$$\mathbb{E}[r_t s_t \mid X_1] = \mathbb{E}[\alpha a X_t - \alpha\delta X_1 \mid X_1] = -\alpha\delta X_1. \tag{155}$$

All $T-1$ later positions contribute to the same random conditional mean. The first term is therefore not the sole source of the quadratic component.

### J.3 The boundary at horizon four

The exact variance formulas hold whenever the corresponding sums are defined. The restriction $T \geq 4$ enters the stated token lower constant through the elementary comparison $T - 2 \geq T/2$. At $T = 4$,

$$\text{Var}(g_{\text{tok}}) = \alpha^2\left[(a-3\delta)^2 + 3a^2\right] \geq 4\alpha^2\delta^2, \tag{156}$$

$$\text{Var}(g_{\text{full}}) = \alpha^2(64a^2 + 24\delta^2), \tag{157}$$

$$\text{Var}(g_{\text{rtg}}) = \alpha^2(30a^2 + 6\delta^2). \tag{158}$$

The first inequality is the specialization of (84); the other two are direct substitutions into (61) and (62). Thus the stated constants hold from $T = 4$ onward.

### J.4 Construction properties

The two constructions satisfy four properties simultaneously:

1. every student token is uniform and nondegenerate on $\{-1, 1\}$ at $\theta_0$;
2. the scalar parameter in (56) is shared at every position;
3. every teacher conditional has logit magnitude $\delta > 0$, so its gap from the uniform student is nonzero at every prefix; and
4. the logits and all constraint margins are fixed independently of the horizon.

For the sequence witness the teacher is identical across all prefixes. For the token witness it uses only the already-observed first sign, which is a valid prefix-dependent autoregressive rule. Both are interior, full-support witnesses for the supremum rates.

### J.5 Origin of the horizon factor

The pathwise estimate $\|M_T\|_2 \leq TS$ allows $\mathbb{E}\|M_T\|_2^2$ to be as large as $T^2S^2$. Conditional score centering replaces this by $TS^2$ in expectation. The feedback sum in the full estimator still contributes $T^2R^2$, leaving a cubic product. In the RTG form, the same replacement applies at each partial horizon $u$, and $T\sum_{u\leq T} u$ is cubic. The explicit fixed-teacher construction retains exactly enough common structure between log-ratios and scores to realize that cubic order without violating score orthogonality.