VALUE-GRADIENT HYPOTHESIS OF RL FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning substantially improves pretrained language models, but it remains understudied why critic-free methods such as GRPO and critic-free PPO variants work as well as they do, and when they should provide the largest gains. We develop a value-gradient perspective of critic-free RL for LLM post-training. First, under a differentiable rollout and additive-noise parameterization, we identify when the expected actor update admits a pathwise representation: *the corresponding return costates have conditional expectation equal to the value gradient*. Second, for discrete transformer policies, we characterize empirical adjoints through the layer–position computation graph and bound their discrepancy from a specified relaxed value signal by local-signal and transition residuals. Policy entropy controls one relaxation component of this bound, not the entire discrepancy. These results motivate a hypothesis decomposing RL impact into value-gradient signal and reachable reward headroom, yielding a criterion for when RL may be most effective along a pretraining trajectory.

1 INTRODUCTION

Recently, Large Language Models (LLMs) achieve strong reasoning performance using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which discards the critic entirely, yet critic-free Reinforcement Learning (RL) can suffer from high-variance long-horizon credit assignment. What signal makes learning possible? *In this paper we argue that critic-free RL in LLMs is not value-free*. The central claim is that the actor backward pass already carries a value-gradient-like signal. In a differentiable rollout, this signal is exactly the costate propagated by Backpropagation Through Time (BPTT). In a discrete transformer, attention transports surrogate gradients between layers and positions; relating them to a relaxed return costate additionally requires controlling the local learning signal and the transition discrepancy.

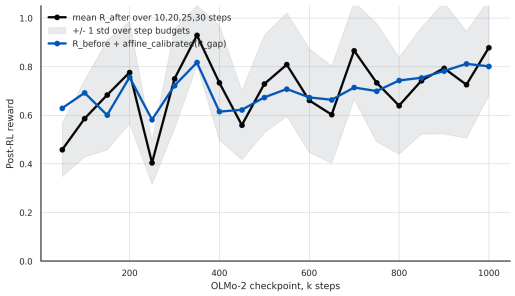

Figure 1: Real RL gain vs. predicted one using value impact formula (Section 5, Eq. 28).

First, under a continuous relaxation (Fairbank & Alonso, 2012), we identify when the Proximal Policy Optimization (PPO)/GRPO actor update is value-gradient-like in expectation, specifying the effect of group normalization. Second, for discrete transformer policies, we give a conditional approximation bound retaining both signal and transition errors. Third, we use this perspective to motivate an RL-impact decomposition into usable value-gradient signal and reachable reward headroom, which predicts when RL should be most effective along pretraining (Figure 1). This perspective gives a concrete answer to a practical question: *RL should help most at checkpoints that are simultaneously close enough to the value-gradient regime to transmit useful credit and far enough from saturation to retain reward-improving trajectories*.

Our **contributions** are: We specify the expected value-gradient interpretation of a local actor update under a differentiable rollout and shift/additive-noise policy. We characterize the transformer surrogate adjoint and bound its discrepancy from a relaxed BPTT costate under explicit comparison assumptions. We propose a predictive RL-impact decomposition into usable value-gradient signal and reachable reward headroom for checkpoint selection during pretraining.

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.

2 BACKGROUND

Notation. A prompt/question is denoted by $q \sim P(Q)$. Given q, a policy with parameters θ generates an autoregressive completion $o = (o_1, \ldots, o_T)$ with token-level factorisation $\pi_\theta(o \mid q) = \prod_{t=1}^{T} \pi_\theta(o_t \mid s_t)$, $s_t := (q, o_{<t})$, $a_t := o_t$. We write $(s_t, a_t)_{t=1}^{T}$ for the token-level trajectory. Let $r(s_t, a_t) \in \mathbb{R}$ denote a per-token reward. *Outcome-only* RL is the case where $r(s_t, a_t) = 0$ for $t < T$ and $r(s_T, a_T) = r(q, o)$ is a terminal reward. With a fixed discount $\gamma \in (0, 1]$, the discounted return and the return-to-go are $R := \sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)$, $R_t := \sum_{j=t}^{T} \gamma^{j-t} r(s_j, a_j)$.

For any policy π, define the time-indexed value function, action-value, and advantage $V_t^\pi(s) := \mathbb{E}[R_t \mid s_t = s]$, $Q_t^\pi(s, a) := \mathbb{E}[R_t \mid s_t = s,\, a_t = a]$, $A_t^\pi(s, a) := Q_t^\pi(s, a) - V_t^\pi(s)$. The *value gradient* is the state-gradient of the value function, $G_t^\pi(s) := \frac{\partial V_t^\pi(s)}{\partial s}$. Unlike the scalar V_t^π, it is a vector field on a differentiable state space. For discrete prefixes this derivative is not defined intrinsically: we differentiate a specified continuous state or activation, holding the discrete history fixed. Throughout the paper we write gradients with the partial-derivative symbol: $\frac{\partial F(\theta)}{\partial \theta}$ for a gradient in parameter space, $\frac{\partial \phi(s)}{\partial s}$ and $\frac{\partial \phi(s,a)}{\partial a}$ for gradients in state or action space, respectively.

2.1 RL FOR LLMS

GRPO (Shao et al., 2024) is commonly presented as a PPO (Schulman et al., 2017) variant that removes the learned critic/value model. For each prompt q, it samples a group of G completions $\{o^i\}_{i=1}^{G}$ from an old policy $\pi_{\theta_{\text{old}}}$, where the index i identifies the rollout (and its return, which may be terminal-only or per-token). Let $T_i := |o^i|$ and $s_{i,t} := (q, o^i_{<t})$. In the outcome-only case each completion receives a scalar reward $r_i = r(q, o^i)$. GRPO forms a group-normalised signal $\tilde{r}_i$ and the tokenwise likelihood ratio: for $t = 1, \ldots, T_i$, $\tilde{r}_i := \frac{r_i - \text{mean}(r)}{\text{std}(r) + \xi_{\text{num}}}$, $\rho_{i,t}(\theta) := \frac{\pi_\theta(o_t^i \mid s_{i,t})}{\pi_{\theta_{\text{old}}}(o_t^i \mid s_{i,t})}$, with $\xi_{\text{num}} > 0$ for numerical stability and $r = (r_1, \ldots, r_G)$. GRPO then uses a tokenwise advantage estimate constant along the completion, $\widehat{A}_{i,t} := \tilde{r}_i$ for all t (process-supervision variants can use token- or step-level reward-to-go). The objective is the GRPO clipped surrogate with a KL penalty to a fixed reference policy π_{ref}:

$$J(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\min\Big(\rho_{i,t}(\theta)\widehat{A}_{i,t},\ \text{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\widehat{A}_{i,t}\Big) - \beta\,\text{KL}\Big(\pi_\theta \,\|\, \pi_{\text{ref}}\Big)\right)\right], \tag{1}$$

where $\varepsilon > 0$ is the clipping threshold and $\beta > 0$ controls the KL regularisation. GRPO differs from PPO only in how $\widehat{A}_{i,t}$ is constructed: it uses a group-normalised return (computed from multiple rollouts for the same prompt) and typically holds the resulting scalar constant along the trajectory.

2.2 GRADIENT ESTIMATORS

Let x be a random variable, c a differentiable scalar cost, and $F(\theta) := \mathbb{E}_x[c(x)]$ the objective. There are two canonical ways θ can enter this expectation (Schulman et al., 2015). For a **Score-function (SF) estimator**, if $x \sim p(\,\cdot\,; \theta)$, then by the log-derivative trick

$$\frac{\partial}{\partial \theta}\,\mathbb{E}_{x\sim p(\cdot;\theta)}\big[c(x)\big] \;=\; \mathbb{E}_x\left[c(x)\,\frac{\partial}{\partial \theta}\log p(x;\theta)\right]. \tag{2}$$

This identity assumes parameter-independent support and an integrable bound permitting differentiation under the expectation. If c depends explicitly on θ, the additional term $\mathbb{E}[\partial c/\partial\theta]$ is required. Crucially, c need not be differentiable, or even continuous, in x. That is exactly why SF is the natural estimator for classical RL: discrete actions, non-differentiable rewards, and unknown dynamics are all admissible. The cost is variance: the estimator uses only the *scalar* $c(x)$, not its slope, and therefore *ignores all local geometry of c.*

Pathwise-derivative (PD) estimator. If instead $x = x(z, \theta)$ is a differentiable function of θ and an exogenous noise variable $z \sim p(z)$ whose distribution does *not* depend on θ (a *reparameterization*), then differentiation and expectation commute directly:

$$\frac{\partial}{\partial \theta}\,\mathbb{E}_z\big[c(x(z,\theta))\big] \;=\; \mathbb{E}_z\left[\frac{\partial}{\partial \theta}c(x(z,\theta))\right]. \tag{3}$$

PD exploits $\frac{\partial c}{\partial x}$ directly and is typically lower variance than SF when both apply (Rezende et al., 2014). The price is a stronger regularity requirement: $c \circ x(\cdot, \theta)$ must be (almost everywhere) differentiable, and the sampling must admit a reparameterization with locally integrably dominated derivatives; almost-everywhere differentiability alone is insufficient.

2.3 Costates and Value-Gradients

We now lift the single-variable policy gradient setup of §2.2 to trajectories and ask what object the backward pass in RL settings computes.

Definition 1 (Differentiable rollout)**.** *The horizon is finite, the functions π_θ, f_θ, r are continuously differentiable, and the return and its relevant derivatives admit locally integrable dominating bounds. The reward has no explicit parameter dependence. The initial state is independent of θ unless its derivative is included separately. The noise law $p(\xi)$ does not depend on the initial state or θ, and*

$$a_t = \pi_\theta(s_t, \xi_t), \qquad s_{t+1} = f_\theta\big(s_t, \pi_\theta(s_t, \xi_t)\big), \qquad \xi_t \sim p(\xi) \text{ i.i.d., independent of } \theta. \tag{4}$$

Let D denote the total derivative with respect to the state, accounting for both the direct state dependence and the indirect dependence through the policy's action. $Df_\theta(s_t, a_t) = \frac{\partial f_\theta}{\partial s} + \frac{\partial f_\theta}{\partial a}\frac{\partial \pi_\theta}{\partial s}$ for the vector-valued dynamics, and similarly Dr for the reward. Because all randomness is exogenous under Definition 1, the sampled return $R_1(\theta, \xi_{1:T})$ is a deterministic function of the parameters given the noise. This allows differentiation to commute with expectation, resulting in a pathwise identity formula analogous to equation 3: $\frac{\partial J(\theta)}{\partial \theta} = \mathbb{E}\big[\frac{\partial R_1}{\partial \theta}\big]$. The gradient $\frac{\partial R_1}{\partial \theta}$ is computed by differentiating the unrolled computation graph equation 4, a process known as *backpropagation through time* (BPTT) (Fairbank & Alonso, 2012). Crucially, BPTT does not propagate the parameter gradient itself. It propagates the state-sensitivity adjoint, which we call the **costate**:

$$\lambda_t \;:=\; \frac{\partial R_t}{\partial s_t}, \qquad \lambda_{T+1} := 0. \tag{5}$$

Intuitively, the effect of the current state on future return has two pieces: the immediate effect on the current reward, plus the future effect pushed backward through the transition Jacobian. Formally:

Proposition 1 (Adjoint recursion, (Fairbank & Alonso, 2012))**.** *Under Definition 1, the costates satisfy*

$$\boxed{\lambda_t \;=\; Dr(s_t, a_t) \;+\; \gamma\,\big(Df_\theta(s_t, a_t)\big)^\top \lambda_{t+1}} \tag{6}$$

for $t = T, \dots, 1$, and the exact pathwise parameter gradient is

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta} = \mathbb{E}\Bigg[\sum_{t=1}^{T} \gamma^{t-1}\Bigg(&\gamma\left(\frac{\partial f_\theta(s_t, a_t)}{\partial \theta}\right)^\top \lambda_{t+1} \\ &+ \left(\frac{\partial \pi_\theta(s_t, \xi_t)}{\partial \theta}\right)^\top \left(\frac{\partial r(s_t, a_t)}{\partial a_t} + \gamma\left(\frac{\partial f_\theta(s_t, a_t)}{\partial a_t}\right)^\top \lambda_{t+1}\right)\Bigg)\Bigg].\end{aligned} \tag{7}$$

The proof (see Appendix A) is a direct chain-rule expansion of $R_t = r(s_t, a_t) + \gamma R_{t+1}$.

Costates are value-gradient estimators. Conditioning on $s_t = s$, the value function satisfies $V_t^\pi(s) = \mathbb{E}\big[r(s_t, a_t) + \gamma V_{t+1}^\pi(s_{t+1}) \,\big|\, s_t = s\big]$. Comparing this value recursion with the sampled costate recursion in equation 6, we see that the two have the same form. More precisely, differentiating the Bellman value recursion with respect to s and interchanging differentiation and expectation (under Definition 1) gives

$$\frac{\partial V_t^\pi(s)}{\partial s} = \mathbb{E}\Big[Dr(s_t, a_t) + \gamma\big(Df_\theta(s_t, a_t)\big)^\top G_{t+1}^\pi(s_{t+1}) \,\Big|\, s_t = s\Big]. \tag{8}$$

The value gradient is the corresponding *conditional expectation* over that future noise. Therefore,

$$\mathbb{E}[\lambda_t \mid s_t = s] = \frac{\partial V_t^\pi(s)}{\partial s}. \tag{9}$$

Thus the quantity propagated backward by BPTT is, at each step, a Monte Carlo sample of the value gradient. Thus *critic-free* does not mean *value-free*: the relevant value information is present as a propagated gradient signal rather than as a separately fitted scalar critic.

Takeaway

The key object for this paper is the costate: in a differentiable rollout, the backward signal propagated by BPTT is a Monte Carlo estimator of the value gradient.

3 Continuous Lens: Why Critic-Free RL Is Value-Gradient-Like

We now identify when the local GRPO/PPO actor update admits a pathwise representation in a continuous-relaxed rollout. The standard formulation uses discrete token actions, $s_t := (q, o_{<t})$, $a_t := o_t$, $s_{t+1} = (s_t, a_t)$, for which a sampled token is not pathwise differentiable. Nevertheless, score-function gradients remain valid. Under a shift/additive-noise parameterization, integration by parts supplies a bridge to action derivatives. Applying Definition 1 to a specified differentiable surrogate then makes its return costate explicit. The bridge equates expectations for the same target; it does not equate the two samplewise backward passes.

Lemma 1 (Fairbank's (Fairbank & Alonso, 2012) SF = PD under a shift policy)**.** *Fix a state s and let $F(\theta) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[r(s,a)]$, with r differentiable in a. If*

$$\pi_\theta(a \mid s) = \nu\big(a - \bar{a}_\theta(s)\big),$$

for a differentiable density ν with vanishing boundary terms and the required integrability, then the score-function and pathwise forms agree in expectation. In particular,

$$\underbrace{\mathbb{E}\left[r(s,a)\,\frac{\partial}{\partial\theta}\log\pi_\theta(a \mid s)\right]}_{\textit{SF form}} = \underbrace{\left(\frac{\partial\bar{a}_\theta(s)}{\partial\theta}\right)^{\top}\mathbb{E}\left[\frac{\partial r(s,a)}{\partial a}\right]}_{\textit{PD form}}. \tag{10}$$

Lemma 1 is the bridge, but the scalar in it must be the actual actor weight. Fix $\theta_0 = \theta_{\text{old}}$ and a state s, and define the frozen-reference conditional signal $\bar{A}_t(s,a) := \mathbb{E}_{\theta_0}[\widehat{A}_t \mid s_t = s, a_t = a]$. If this function satisfies the lemma's regularity conditions, then

$$\mathbb{E}\left[\widehat{A}_t\,\frac{\partial}{\partial\theta}\log\pi_\theta(a_t \mid s)\right]\Big|_{\theta=\theta_0} = \left(\frac{\partial\bar{a}_\theta(s)}{\partial\theta}\right)^{\top}\Big|_{\theta=\theta_0}\mathbb{E}_{\theta_0}\left[\frac{\partial\bar{A}_t(s,a_t)}{\partial a_t}\right]. \tag{11}$$

There is no additional reward factor. Detaching $\widehat{A}_t$ in the score-function implementation does not remove the action dependence of $\bar{A}_t$. With an unbiased return-to-go weight and an action-independent baseline, $\bar{A}_t = Q_t^{\pi_{\theta_0}} - b_t$. For parameter-independent dynamics, summing the discount-weighted actor terms then matches the expected pathwise return gradient. Its costates satisfy

$$\lambda_t := \frac{\partial R_t}{\partial s_t}, \quad \mathbb{E}[\lambda_t \mid s_t = s] = \frac{\partial V_t^{\pi_{\theta_0}}(s)}{\partial s}. \tag{12}$$

Group normalization generally changes the reward target. Locally, outcome-only GRPO instead has a frozen-reference *effective trajectory reward*, including its $1/T_i$ weighting; Appendix A.5 gives the exact identity and the additional assumptions needed for a continuous interpretation. Clipping and KL regularization are kept separate. Thus, the value-gradient interpretation is target-specific, not an identity between normalized GRPO and the original expected return.

Takeaway

Under a differentiable continuous relaxation, a local actor update can be value-gradient-like in expectation for its specified target; group normalization must be included in that target.

4 Discrete GRPO Sends Approximate BPTT Signals

Section §3 identifies an expected-update equivalence under a continuous relaxation and appropriate target assumptions. In practice, LLM rollouts use a *discrete* categorical policy. Autodifferentiation of the actor surrogate still traverses the transformer's internal computation graph. We characterize that adjoint first, then state the conditions under which it approximates a specified relaxed return costate.

Outcome-only GRPO uses $\gamma = 1$ here; discounted formulas refer to consistently discount-weighted surrogates.

For a single trajectory, the local ascent surrogate, with detached weights and clipping inactive at θ_{old}, is

$$L(\theta) = \sum_{t=1}^{T} \widehat{A}_t \cdot \ell_t(\theta), \qquad \ell_t := \log \pi_\theta(o_t \mid s_t) := \log \operatorname{softmax}\big(W_{\text{head}} h_t^{(L)}(\theta)\big)_{o_t}. \tag{13}$$

The group and length averages remain as in equation 1. A transformer layer satisfies $h_k^{(l+1)} = \text{TransformerBlock}^{(l+1)}(h_{1:k}^{(l)})$. The relevant cross-position Jacobian is $J_{k\leftarrow t}^{(l)} := \partial h_k^{(l+1)} / \partial h_t^{(l)}$, not a final-layer-to-final-layer derivative: for fixed sampled tokens, $\partial h_k^{(L)} / \partial h_t^{(L)} = 0$ when $k > t$.

Definition 2 (Empirical costate)**.** *The empirical costate at layer l and position t is the surrogate adjoint*

$$\hat{\lambda}_t^{(l)} := \frac{\partial}{\partial h_t^{(l)}} \sum_{k \geq t} \widehat{A}_k \ell_k. \tag{14}$$

The exact layer–position recursion is

$$\boxed{\begin{aligned} \hat{\lambda}_t^{(L)} &= \widehat{A}_t \frac{\partial \ell_t}{\partial h_t^{(L)}}, \\ \hat{\lambda}_t^{(l)} &= \sum_{k \geq t} \big(J_{k\leftarrow t}^{(l)}\big)^\top \hat{\lambda}_k^{(l+1)}, \qquad l < L. \end{aligned}} \tag{15}$$

To write a *temporal* recursion, h_t must instead be a complete pre-token computational state, including the relevant key/value cache, with discrete history fixed. With fixed horizon, caches can be padded to a common dimension. Let $h_{t+1} = f_\theta(h_t, e_{o_t})$ and $J_t = \partial_h f_\theta(h_t, e_{o_t})$ with the token held fixed. The corresponding suffix-surrogate adjoint obeys

$$\hat{\lambda}_t = b_t + \gamma J_t^\top \hat{\lambda}_{t+1}, \qquad b_t = \widehat{A}_t \frac{\partial \ell_t}{\partial h_t}, \tag{16}$$

$$\lambda_t = d_t + \gamma F_t^\top \lambda_{t+1}, \qquad d_t = D r_t, \quad F_t = D f_\theta^{\text{rel}}. \tag{17}$$

The second line is the relaxed return recursion. The recursions share an adjoint form, but neither $b_t = d_t$ nor $J_t = F_t$ follows from that structural similarity. Both terminal adjoints are zero.

4.1 What is missing vs. exact BPTT

The sampled-token path is

$$h_t \xrightarrow{\text{diff}} z_t \xrightarrow{\text{sample}} o_t \xrightarrow{\text{embed}} e_{o_t} \xrightarrow{\text{diff}} h_{t+1}. \tag{18}$$

A score-function estimator does not differentiate through this sample, nor does it need to do so to estimate a discrete expected-return gradient (Schulman et al., 2015). A pathwise comparison requires an explicit relaxation. Set $p_t = \operatorname{softmax}(z_t/\tau_{\text{rel}})$, $e_t = E_{\text{emb}} p_t$, with fixed $\tau_{\text{rel}} > 0$. At the *same* relaxed state and embedding, let $J_t^{\text{stop}} = \partial_h f_\theta(h_t, e_t)$ hold e_t fixed. Then

$$F_t^{\text{rel}} = J_t^{\text{stop}} + \frac{\partial f_\theta}{\partial e} E_{\text{emb}} \frac{\partial p_t}{\partial z_t} \frac{\partial z_t}{\partial h_t}. \tag{19}$$

This is a relaxed-chain-rule identity, not a derivative through a categorical draw. Its stopped Jacobian need not equal the discrete trajectory's J_t.

Proposition 2 (Sampling-gap bound for a specified relaxation)**.** *Let $H_t = -\sum_v p_{t,v} \log p_{t,v}$ and suppose $\|\partial_e f_\theta\| \, \|E_{\text{emb}}\| \, \|\partial_h z_t\| \leq C_0$. For the relaxation above, with operator norms throughout,*

$$\|F_t^{\text{rel}} - J_t^{\text{stop}}\| \leq \frac{C_0}{\tau_{\text{rel}}} \min\{1, H_t\} \leq C \sqrt{\frac{H_t}{\log |V|}}. \tag{20}$$

Here $C = C_0 \sqrt{\log |V|}/\tau_{\text{rel}}$ and $|V| \geq 2$.

Remark 1 (Why this bound is natural)**.** *At fixed temperature and bounded surrounding Jacobians, a concentrated softmax has a small Jacobian. The bound is not uniform as $\tau_{\mathrm{rel}} \to 0$ and is not an unbiasedness claim for straight-through estimators.*

Remark 2 (Why this matters for LLMs)**.** *Low entropy can suppress this particular relaxation term. It does not alone bound the reward/advantage mismatch, the discrete–relaxed trajectory mismatch, or the accumulated costate error. KL regularization does not automatically supply these bounds.*

Attention transports credit across positions. In an RNN, $h_{t+1} = f(h_t, e_{o_t})$ retains the differentiable path $\partial f/\partial h_t$ even when the sampled token is fixed. Transformers likewise transport gradients through lower-layer attention and cached keys/values, not through a dependence between distinct final-layer outputs. The attention scores

$$\alpha_{t+1,\,t'} \;\propto\; \exp\!\left(q_{t+1} \cdot k_{t'} / \sqrt{d_{\mathrm{head}}}\right) \tag{21}$$

provide multiple deterministic credit pathways. Their existence does not by itself establish agreement with a return gradient.

Proposition 3 (Attention-pathway rank and magnitude)**.** *For one standard multi-head attention sublayer, a cross-position Jacobian has rank at most $\min\{d_{\mathrm{model}}, n_{\mathrm{heads}} d_{\mathrm{head}}\}$. For an earlier input position t', each head's norm is bounded by its attention weight times a factor depending on query, key, and value norms and projection norms. Full-stack bounds also require the intervening layer Jacobians; broad attention alone does not guarantee useful alignment.*

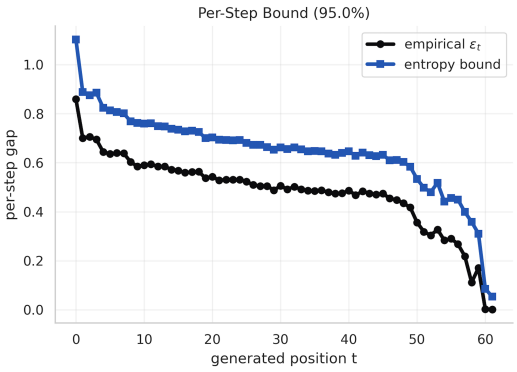

Figure 2: Bound inequality plot. Real value gradient gap vs proposed bound (Section 4, Prop. 2).

4.2 Approximation Theorem

Theorem 1 (Conditional approximation of BPTT costates)**.** *Fix a complete state h_t. Couple a discrete suffix and a specified relaxed suffix started at that same state, using common state coordinates. Suppose their adjoints satisfy equation 16–equation 17, the relaxed rollout satisfies Definition 1, and the quantities below are integrable. Define $\epsilon_k := \|b_k - d_k\| + \gamma\|J_k - F_k\|\,\|\lambda_{k+1}\|$. Then, with $G_t^{\mathrm{rel}}(h_t) = \partial V_t^{\mathrm{rel}}(h_t)/\partial h_t$,*

$$\begin{aligned}&\left\|\mathbb{E}[\hat{\lambda}_t \mid h_t] - G_t^{\mathrm{rel}}(h_t)\right\| \\ &\qquad \le \mathbb{E}\left[\sum_{k=t}^{T} \gamma^{k-t} \left(\prod_{j=t}^{k-1} \|J_j\|\right) \epsilon_k \,\middle|\, h_t\right].\end{aligned} \tag{22}$$

where an empty product equals one. The fixed discrete history is included in the conditioning. Jacobians and future costates remain inside the conditional expectation.

Theorem 1 makes the approximation conditional on *all* residuals. In particular, the terminal local-signal mismatch is retained. Proposition 2 bounds only the through-action component of $F_k - J_k$ at a shared relaxed input; differences of evaluation states and embeddings also contribute. Thus $\mathbb{E}[\hat{\lambda}_t \mid h_t] \approx G_t^{\mathrm{rel}}(h_t)$ follows when the full right-hand side is small, not from low entropy alone. Figure 2 shows the empirical diagnostic; the requirements for a direct test of this bound are detailed in Appendix A.13.

Takeaway

In discrete transformers, autodiff propagates a surrogate adjoint through the actual attention graph. Its proximity to a relaxed value gradient requires control of both local-signal and transition residuals.

5 RL Impact Law Hypotheses

The previous sections identify conditions under which the actor adjoint approximates a specified hidden-state value gradient. We now use this perspective to formulate a predictive hypothesis of *RL*

readiness. The key idea is simple: RL should be effective when the actor update contains a usable value-gradient signal. The resulting score is a diagnostic, not by itself a parameter-update guarantee. For any critic-free normalized policy-gradient method m, let $\hat{\lambda}_t^{(m)}$ denote the empirical hidden-state costate produced by the actor backward pass, and let G_t denote a specified hidden-state value gradient in the same coordinates, measurable given h_t. The quantity

$$\mathbb{E}_{q,\tau,t}\left[\left\langle \mathbb{E}[\hat{\lambda}_t^{(m)} \mid h_t],\, G_t \right\rangle\right] \tag{23}$$

measures how much useful RL signal the checkpoint actually sends. Using $\langle u, v\rangle = \|v\|_2^2 + \langle u - v, v\rangle$, we obtain the lower bound

$$\mathbb{E}_{q,\tau,t}\left[\left\langle \mathbb{E}[\hat{\lambda}_t^{(m)} \mid h_t],\, G_t \right\rangle\right] \geq \Sigma(\theta) - \Lambda_m(\theta), \tag{24}$$

where

$$\Sigma(\theta) := \mathbb{E}_{q,\tau,t}\left[\|G_t\|_2^2\right], \qquad \Lambda_m(\theta) := \mathbb{E}_{q,\tau,t}\left[\|G_t\|_2 \left\|\mathbb{E}[\hat{\lambda}_t^{(m)} \mid h_t] - G_t\right\|_2\right]. \tag{25}$$

We therefore define the *usable value-gradient signal*

$$S_m(\theta) := \left(\Sigma(\theta) - \Lambda_m(\theta)\right). \tag{26}$$

Intuitively, $\Sigma(\theta)$ measures how much value-gradient signal is available, while $\Lambda_m(\theta)$ measures how much of that signal is lost because the actual actor update is still separated from the value-gradient regime. By §4, the gap becomes small when the full propagated residual bound is small. Hidden-state alignment must additionally be mapped through parameter Jacobians and optimizer geometry to control an actual update (Appendix A.11). Signal quality alone is not enough. Even a clean RL signal will produce little gain if the current policy has no useful reward headroom left to exploit. We therefore define the *reachable headroom*

$$\mathcal{H}_\alpha(\theta) := \mathbb{E}_{q\sim P(Q)}\left[\frac{1}{\alpha}\log \mathbb{E}_{\tau\sim\pi_\theta(\cdot|q)}\left[e^{\alpha R(\tau)}\right] - \mathbb{E}_{\tau\sim\pi_\theta(\cdot|q)}\left[R(\tau)\right]\right], \tag{27}$$

where $\alpha > 0$ and $R(\tau)$ is a bounded trajectory return. This is a KL-penalized distributional reweighting headroom, not guaranteed improvement within a parameterized policy class. It vanishes for a constant reward and, at fixed α, becomes small as bounded rewards concentrate at either a low or a saturated level (Appendix A.12). These two terms motivate the central empirical hypothesis:

$$\boxed{\text{RL Impact} \propto \underbrace{S_m(\theta)}_{\text{value-gradient signal}} \times \underbrace{\mathcal{H}_\alpha(\theta).}_{\text{reachable headroom}}} \tag{28}$$

For a fixed task, reward, and RL budget B, we write this in the simplest one-constant form as

$$\Delta\text{Perf}_{m,B}(\theta) \approx \kappa_{m,B}\, S_m(\theta)\, \mathcal{H}_\alpha(\theta), \qquad \kappa_{m,B} > 0. \tag{29}$$

Eq. equation 28 is a finite-budget prediction to be tested, not a consequence of the signal bound alone. RL should work best when the checkpoint is already close enough to the value-gradient regime to provide a usable actor signal, but still has enough reward headroom for RL to exploit. Practically, if $\{\theta_N\}$ is a sequence of pretrained *checkpoints*, then the predicted best point to start RL is

$$N^\star = \arg\max_N S_m(\theta_N)\, \mathcal{H}_\alpha(\theta_N). \tag{30}$$

Takeaway

We hypothesize that RL gain depends on two factors at once: how much usable value-gradient signal the model transmits, and how much reward-improving headroom remains in its trajectory distribution.

6 Experiments

We evaluate the proposed framework in two complementary settings. First, we examine the entropy-related diagnostic from §4 on OLMo-2 (Team OLMo et al., 2024) checkpoints. Second, we test whether the costate-based impact score predicts which checkpoints benefit most from RL.

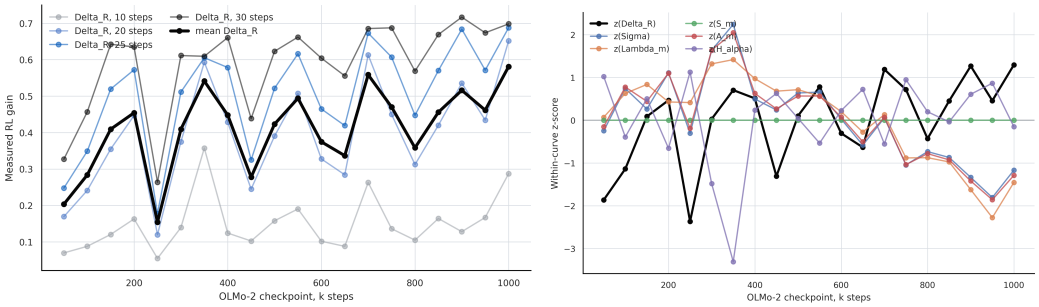

Figure 3: Results of the RL upon various OLMO-2 pretraining checkpoints. The left image shows the consistency of the achieved gains via RL using a different number of GRPO steps. The right image shows the behavior of the different components of our study.

Closed-form RL task. For the RL-impact experiment, we use OLMo-2 1B checkpoints from pretraining steps 50k to 1M in increments of 50k. The task is a controlled label-copying problem. Given a prompt containing a target label in $\{A, B, C, D\}$, the model must put probability mass on the matching answer token. The reward is $R(\theta; q) = p_\theta(y^\star \mid q)$, where $y^\star$ is the correct label. This reward is differentiable in the model logits and avoids decoded-text, classifier, or parser discontinuities. For each checkpoint θ_N, we measure the pre-RL reward $R_{\text{before}}(\theta_N) = \mathbb{E}_q\, p_{\theta_N}(y^\star \mid q)$, then run GRPO from the checkpoint and measure the final post-RL reward $R_{\text{after}}(\theta_N, K)$, where K is the number of RL updates. We use several RL budgets, $K \in \{10, 20, 25, 30\}$, and report both individual-budget results and the averaged response $\overline{\Delta R}(\theta_N) = \mathbb{E}_K\left[R_{\text{after}}(\theta_N, K) - R_{\text{before}}(\theta_N)\right]$.

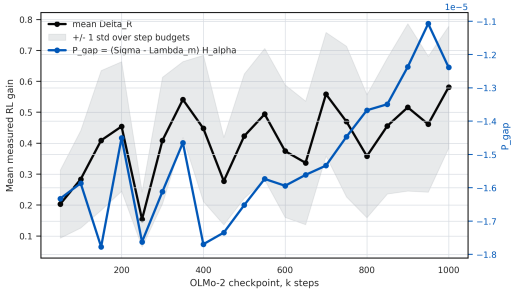

Figure 4: Real RL gain vs. predicted one using value impact formula (Section 5, Eq. 28).

Entropy-bound diagnostic. Figure 2 compares the measured costate discrepancy with the entropy-based diagnostic on OLMo-2, using the reward function described above. It is not, by itself, a verification of Theorem 1: the bound also requires the local-signal residual, the coupled transition residual, operator norms, and the specified relaxation temperature. The distinction between the plotted diagnostic and a theorem-level bound is detailed in Appendix A.13.

Costate-based predictors. For each checkpoint, we compute the differentiable-score gradient $G_t = \partial R/\partial h_t$ and the detached-reward RL costate $\widehat{\lambda}_t = \partial[\widehat{A}_t \log \pi_\theta(a_t \mid s_t)]/\partial h_t$, using the same prompts, hidden states, sampled actions, and reward evaluations. This matches the evaluation points, but does not itself identify the discrete and relaxed value gradients. We then estimate $\Sigma = \mathbb{E}\|G_t\|^2$, $\Lambda_m = \mathbb{E}[\|G_t\|\,\|\widehat{\lambda}_t - G_t\|]$, and the headroom term $H_\alpha = \frac{1}{\alpha}\log \mathbb{E}[\exp(\alpha R)] - \mathbb{E}[R]$. The impact score is $P_{\text{gap}} = (\Sigma - \Lambda_m)H_\alpha$, which preserves the ordering information.

RL-response prediction. Figure 1, 3 and 4 shows the costate components and the impact predictor. The component plot shows that the raw quantities vary systematically across checkpoints, while the predictor tracks the broad shape of the measured RL gain curve. This is consistent with the diagnostic hypothesis, subject to Appendix A.13: checkpoints differ not only in their current reward, but also in the quality of their costate signal and their remaining reward headroom.

Figure 5 evaluates the final predictive story. The left panel compares the realized mean post-RL reward against $\widehat{R}_{\text{after}} = R_{\text{before}} + \text{Affine}(P_{\text{gap}})$, where the affine map calibrates the scale of P_{gap} to the observed reward-gain scale. The right panel uses a scale-free version, $z(R_{\text{before}}) + z(P_{\text{gap}})$, and compares it to $z(\overline{R}_{\text{after}})$, where $z(\cdot)$ denotes normalization across checkpoints. This asks whether pretrained competence plus predicted RL readiness explains the final post-RL checkpoint quality. Empirically, the combined predictor correlates more strongly with averaged post-RL reward than either current reward or the impact score alone. Across the RL budgets, the score satisfies $\text{Spearman}(P_{\text{gap}}, \overline{\Delta R}) \approx 0.60$, while the combined scale-free predictor satisfies $\text{Spearman}(z(R_{\text{before}}) + z(P_{\text{gap}}), z(\overline{R}_{\text{after}})) \approx 0.73$. These reported correlations suggest a

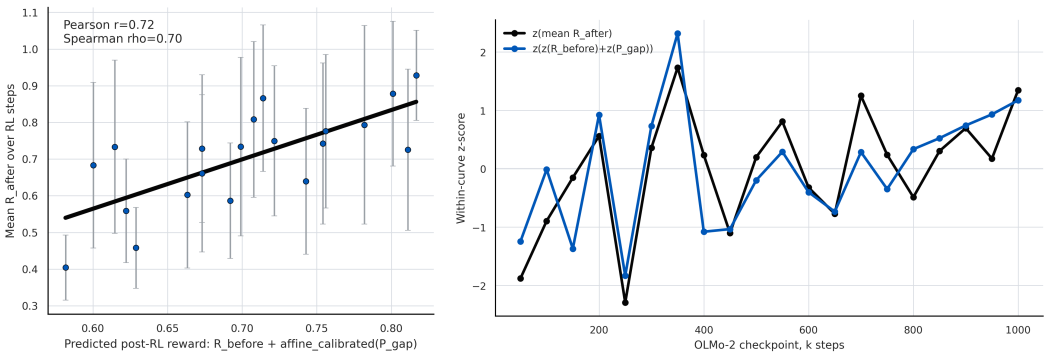

Figure 5: Z-scores of the gain after RL vs. z-scores of the predicted RL impact. Correlation (left) and curve per checkpoint (right). The std bars on left indicate RL gain variance over the various training steps.

checkpoint-dependent diagnostic signal, but do not establish the adjoint bound or out-of-sample prediction; current competence also remains relevant for predicting the final reward level.

Limitations. The current setup is controlled and partially toy-like. The score $R(\theta; q)$ does not specify a completion-dependent GRPO training reward, and the samplewise discrepancy and headroom estimators require the distinctions in Appendix A.13. The reward is deliberately chosen to be differentiable in the model logits, which allows exact derivatives of the chosen score. Many parameters affect the observed correspondence, including reward scale, RL learning rate, batch size, group size, and the number of RL updates. Our method may indicate that a checkpoint is ready for efficient RL, but how RL is performed remains an additional optimization question. Indeed, too few RL steps under-realize the predicted headroom, while too many steps saturate the reward and obscure checkpoint differences. For this reason, we report averages across several RL budgets.

7 Related Work and Discussion

There is a large body of work on extending the GRPO algorithm. Methods such as these modify GRPO, and applying our analysis requires retaining their specific weighting and normalization. In terms of the critic-free RL interpretation, it has been shown that GRPO implicitly computes a value-like quantity via U-statistics (Zhou et al., 2026). However, this account is specific to GRPO, while PPO (Schulman et al., 2017) and even REINFORCE also perform unexpectedly well in the LLM setting. Moreover, evidence from standard RL suggests that group-normalized episodic baselines are generally insufficient to replace a critic for bootstrapping and temporal credit assignment (de Oliveira, 2025). Therefore, we argue that the gains of RL in LLMs come not only from the GRPO baseline.

Zheng et al. (2025) propose using BPTT directly instead of RL and obtain better empirical results, motivating comparisons of pathwise and score-function signals for explicitly specified objectives. Another interesting observation is that scalar continuous rewards were used in the latest DeepSeek-V4 model (DeepSeek-AI, 2026) to achieve better results. Recently, value gradient flow-based learning was proposed (Xu et al., 2026), which achieved sustainable results in RL tasks, signaling that the value signal is useful. However, this method differs from ours because their approach is actor-free, showing how the value gradient can be used as a policy, while we focus on the critic-free LLMs.

8 Conclusion

This paper argues that critic-free RL in LLM post-training need not be value-free: its actor update can admit a value-gradient interpretation for a specified target. In a differentiable rollout, return costates are exact in expectation. In discrete transformers, attention transports surrogate adjoints, whose approximation to a relaxed value gradient requires control of local-signal and transition residuals; entropy bounds one relaxation component. This perspective motivates a simple hypothesis: RL gains should be largest when a checkpoint has both usable credit-assignment signal and reward headroom.

AI use statement

We used generative AI tools, in particular a large language model (LLM), as a writing and coding assistant. Concretely, we used them to improve the clarity and grammar of the exposition and to produce the plotting scripts. We have reviewed all AI-assisted text and code and take full responsibility for the final content of this work.

REFERENCES

et al de Oliveira, Bryan LM. Learning without critics? revisiting grpo in classical reinforcement learning environments. *Arxiv*, 12, 2025.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Michael Fairbank and Eduardo Alonso. Value-gradient learning. In *The 2012 international joint conference on neural networks (ijcnn)*, pp. 1–8. IEEE, 2012.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. *Advances in neural information processing systems*, 28, 2015.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2024. URL https://arxiv.org/abs/2501.00656.

Haoran Xu, Kaiwen Hu, Somayeh Sojoudi, and Amy Zhang. Reinforcement learning via value gradient flow. *arXiv preprint arXiv:2604.14265*, 2026.

Zhi Zheng, Yu Gu, Wei Liu, Yee Whye Teh, and Wee Sun Lee. Soft-grpo: Surpassing discrete-token llm reinforcement learning via gumbel-reparameterized soft-thinking policy optimization. *arXiv preprint arXiv:2511.06411*, 2025.

Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Shijin Gong, and Chengchun Shi. Demystifying group relative policy optimization: Its policy gradient is a u-statistic. *arXiv preprint arXiv:2603.01162*, 2026.

A PROOFS

This appendix collects the proofs and supporting derivations for the statements used in the main text. We keep the notation of the main paper: the reparameterized action is written as $a_t = \pi_\theta(s_t, \xi_t)$, the differentiable transition as $s_{t+1} = f_\theta(s_t, a_t)$, and the costate as $\lambda_t = \partial R_t / \partial s_t$.

A.1 PATHWISE DIFFERENTIATION UNDER A DIFFERENTIABLE ROLLOUT

Under Definition 1, all stochasticity enters through the exogenous noise variables $\xi_{1:T}$, whose law is independent of θ. Hence, for a fixed noise realization, the sampled return $R_1(\theta, \xi_{1:T})$ is a deterministic differentiable function of θ. Therefore,

$$J(\theta) = \mathbb{E}_{\xi_{1:T}}[R_1(\theta, \xi_{1:T})] = \int R_1(\theta, \xi_{1:T})\, p(\xi_{1:T})\, d\xi_{1:T}.$$

Under the differentiability and domination assumptions stated in Definition 1, differentiation may be interchanged with integration:

$$\frac{\partial J(\theta)}{\partial \theta} = \int \frac{\partial R_1(\theta, \xi_{1:T})}{\partial \theta} p(\xi_{1:T})\, d\xi_{1:T} = \mathbb{E}\left[\frac{\partial R_1}{\partial \theta}\right].$$

Thus, in the differentiable rollout model, the policy-gradient direction can be computed pathwise by differentiating the unrolled trajectory computation.

A.2 PROOF OF PROPOSITION 1: ADJOINT RECURSION

Proof. Recall that

$$R_t = r(s_t, a_t) + \gamma R_{t+1}, \qquad a_t = \pi_\theta(s_t, \xi_t), \qquad s_{t+1} = f_\theta(s_t, a_t), \qquad \lambda_t := \frac{\partial R_t}{\partial s_t}.$$

Throughout the proof, the noise variables $\xi_{t:T}$ are held fixed.

First differentiate R_t with respect to s_t. The immediate reward contributes the total derivative

$$Dr(s_t, a_t) = \frac{\partial r(s_t, a_t)}{\partial s} + \left(\frac{\partial \pi_\theta(s_t, \xi_t)}{\partial s}\right)^\top \frac{\partial r(s_t, a_t)}{\partial a}.$$

The future return contributes through $s_{t+1} = f_\theta(s_t, a_t)$:

$$\gamma \left(\frac{\partial f_\theta(s_t, a_t)}{\partial s} + \frac{\partial f_\theta(s_t, a_t)}{\partial a} \frac{\partial \pi_\theta(s_t, \xi_t)}{\partial s}\right)^\top \lambda_{t+1} = \gamma \big(Df_\theta(s_t, a_t)\big)^\top \lambda_{t+1}.$$

Therefore,

$$\lambda_t = Dr(s_t, a_t) + \gamma \big(Df_\theta(s_t, a_t)\big)^\top \lambda_{t+1},$$

which is the adjoint recursion equation 6. It remains to derive the parameter-gradient expression. Let

$$\dot{s}_t := \frac{ds_t}{d\theta}.$$

The total state sensitivity satisfies

$$\dot{s}_{t+1} = Df_\theta(s_t, a_t)\, \dot{s}_t + \frac{\partial f_\theta(s_t, a_t)}{\partial \theta} + \frac{\partial f_\theta(s_t, a_t)}{\partial a_t} \frac{\partial \pi_\theta(s_t, \xi_t)}{\partial \theta}.$$

The total derivative of the return is

$$\frac{dR_1}{d\theta} = \sum_{t=1}^{T} \gamma^{t-1} \left[\dot{s}_t^\top Dr(s_t, a_t) + \left(\frac{\partial \pi_\theta(s_t, \xi_t)}{\partial \theta}\right)^\top \frac{\partial r(s_t, a_t)}{\partial a_t}\right].$$

Using the adjoint identity

$$Dr(s_t, a_t) = \lambda_t - \gamma \big(Df_\theta(s_t, a_t)\big)^\top \lambda_{t+1},$$

we obtain

$$\frac{dR_1}{d\theta} = \sum_{t=1}^{T} \gamma^{t-1} \Bigg[\left(\frac{\partial \pi_\theta(s_t, \xi_t)}{\partial \theta} \right)^\top \frac{\partial r(s_t, a_t)}{\partial a_t} + \dot{s}_t^\top \lambda_t - \gamma \dot{s}_t^\top \big(D f_\theta(s_t, a_t) \big)^\top \lambda_{t+1} \Bigg].$$

From the state-sensitivity recursion,

$$D f_\theta(s_t, a_t)\, \dot{s}_t = \dot{s}_{t+1} - \frac{\partial f_\theta(s_t, a_t)}{\partial \theta} - \frac{\partial f_\theta(s_t, a_t)}{\partial a_t} \frac{\partial \pi_\theta(s_t, \xi_t)}{\partial \theta}.$$

Substituting this into the previous display gives

$$\begin{aligned} \frac{dR_1}{d\theta} = \sum_{t=1}^{T} \gamma^{t-1} \Bigg[& \left(\frac{\partial \pi_\theta(s_t, \xi_t)}{\partial \theta} \right)^\top \frac{\partial r(s_t, a_t)}{\partial a_t} \\ & + \gamma \left(\frac{\partial f_\theta(s_t, a_t)}{\partial \theta} + \frac{\partial f_\theta(s_t, a_t)}{\partial a_t} \frac{\partial \pi_\theta(s_t, \xi_t)}{\partial \theta} \right)^\top \lambda_{t+1} \Bigg] \\ & + \sum_{t=1}^{T} \gamma^{t-1} \left[\dot{s}_t^\top \lambda_t - \gamma \dot{s}_{t+1}^\top \lambda_{t+1} \right]. \end{aligned}$$

The second sum telescopes:

$$\sum_{t=1}^{T} \gamma^{t-1} \left[\dot{s}_t^\top \lambda_t - \gamma \dot{s}_{t+1}^\top \lambda_{t+1} \right] = \dot{s}_1^\top \lambda_1 - \gamma^T \dot{s}_{T+1}^\top \lambda_{T+1}.$$

Under the initial-state assumption of Definition 1, $\dot{s}_1 = 0$. A parameter-dependent prompt encoding instead contributes $\dot{s}_1^\top \lambda_1$ and must not be omitted. Also $\lambda_{T+1} = 0$ by definition. Hence the telescoping term vanishes. Collecting the remaining terms yields

$$\begin{aligned} \frac{dR_1}{d\theta} = \sum_{t=1}^{T} \gamma^{t-1} \Bigg[& \gamma \left(\frac{\partial f_\theta(s_t, a_t)}{\partial \theta} \right)^\top \lambda_{t+1} \\ & + \left(\frac{\partial \pi_\theta(s_t, \xi_t)}{\partial \theta} \right)^\top \left(\frac{\partial r(s_t, a_t)}{\partial a_t} + \gamma \left(\frac{\partial f_\theta(s_t, a_t)}{\partial a_t} \right)^\top \lambda_{t+1} \right) \Bigg]. \end{aligned}$$

Taking expectation over the exogenous noise gives equation 7. □

A.3 Proof of identity equation 9: costates are value-gradient estimators

Proof. By definition,

$$V_t^\pi(s) = \mathbb{E}[R_t \mid s_t = s].$$

Conditioned on $s_t = s$, the remaining randomness comes only from the future exogenous noises $\xi_{t:T}$, whose distribution is independent of s and θ. Therefore, under the regularity conditions of Definition 1,

$$\begin{aligned} G_t^\pi(s) = \frac{\partial V_t^\pi(s)}{\partial s} &= \frac{\partial}{\partial s} \mathbb{E}[R_t \mid s_t = s] \\ &= \mathbb{E}\left[\frac{\partial R_t}{\partial s_t} \,\middle|\, s_t = s \right] = \mathbb{E}[\lambda_t \mid s_t = s]. \end{aligned}$$

Thus the BPTT costate is a Monte Carlo sample whose conditional expectation is the value gradient. □

A.4 Proof of Lemma 1: SF = PD under a shift policy

Proof. Fix s. Let

$$\pi_\theta(a \mid s) = \nu(a - \bar{a}_\theta(s)), \qquad u := a - \bar{a}_\theta(s).$$

Then $a = u + \bar{a}_\theta(s)$, and the density of u is $\nu(u)$, independent of θ. By the chain rule,

$$\frac{\partial}{\partial \theta} \log \pi_\theta(a \mid s) = -\left(\frac{\partial \bar{a}_\theta(s)}{\partial \theta}\right)^\top \frac{\partial}{\partial u} \log \nu(u).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[r(s,a)\frac{\partial}{\partial \theta} \log \pi_\theta(a \mid s)\right] &= -\left(\frac{\partial \bar{a}_\theta(s)}{\partial \theta}\right)^\top \int r(s, u + \bar{a}_\theta(s)) \frac{\partial}{\partial u} \log \nu(u)\nu(u)\, du \\ &= -\left(\frac{\partial \bar{a}_\theta(s)}{\partial \theta}\right)^\top \int r(s, u + \bar{a}_\theta(s)) \frac{\partial \nu(u)}{\partial u}\, du.\end{aligned}$$

Integrating by parts and using the assumed vanishing boundary terms,

$$\int r(s, u + \bar{a}_\theta(s)) \frac{\partial \nu(u)}{\partial u}\, du = -\int \nu(u) \frac{\partial r(s, u + \bar{a}_\theta(s))}{\partial u}\, du.$$

Since

$$\frac{\partial r(s, u + \bar{a}_\theta(s))}{\partial u} = \frac{\partial r(s,a)}{\partial a},$$

we obtain

$$\mathbb{E}\left[r(s,a)\frac{\partial}{\partial \theta} \log \pi_\theta(a \mid s)\right] = \left(\frac{\partial \bar{a}_\theta(s)}{\partial \theta}\right)^\top \mathbb{E}\left[\frac{\partial r(s,a)}{\partial a}\right].$$

This proves equation 10. □

A.5 Local GRPO/PPO value-gradient form

The distinction between a detached sampled coefficient and a frozen mathematical target is important here. Fix a reference parameter $\theta_0 = \theta_{\text{old}}$. All expectations in a local actor update sample from this reference; the clipping threshold is inactive at $\rho_{i,t} = 1$. We first identify the local update exactly, before introducing a continuous relaxation.

The actual conditional score-function signal. Let $\psi_t(a) = \partial_\theta \log \pi_\theta(a \mid s)|_{\theta=\theta_0}$. Holding the state fixed and applying the tower property gives

$$\mathbb{E}_{\theta_0}[\widehat{A}_t \psi_t(a_t) \mid s_t = s] = \mathbb{E}_{a \sim \pi_{\theta_0}(\cdot|s)}[\bar{A}_t(s,a)\psi_t(a)], \qquad \bar{A}_t(s,a) = \mathbb{E}_{\theta_0}[\widehat{A}_t \mid s_t = s, a_t = a].$$

The conditional expectation includes future rollout randomness and any peer completions used to form the advantage. For a length-weighted term in equation 1, the same statement uses $\mathbb{E}_{\theta_0}[\widehat{A}_{i,t}/T_i \mid s_{i,t} = s, a_{i,t} = a]$; the length factor cannot be omitted when it is action-dependent.

The function $\bar{A}_t$ is frozen with respect to the differentiation parameter θ, but is generally not constant as a function of a. If it satisfies Lemma 1, substituting it for the lemma's scalar reward proves equation 11. No additional factor $r(s_t, a_t)$ appears. Nor is it legitimate to discard $\partial_a \bar{A}_t$ merely because the realized advantage was detached in a score-function implementation. For example, for $a \sim \mathcal{N}(\theta, 1)$ and coefficient $\widehat{A} = a$,

$$\mathbb{E}[a\, \partial_\theta \log \pi_\theta(a)] = 1 = \mathbb{E}[\partial_a a],$$

although an implementation detaches the sampled a in the score-function product.

Unnormalized return-to-go. If $\widehat{A}_t = R_t - b_t(s_t)$, with an action-independent baseline, then $\bar{A}_t(s,a) = Q_t^{\pi_{\theta_0}}(s,a) - b_t(s)$ and $\partial_a \bar{A}_t = \partial_a Q_t^{\pi_{\theta_0}}$. For a continuous rollout with parameter-independent dynamics and initial state, the policy-gradient identity is

$$\left.\frac{\partial J(\theta)}{\partial \theta}\right|_{\theta_0} = \mathbb{E}_{\theta_0}\left[\sum_{t=1}^{T} \gamma^{t-1}(R_t - b_t(s_t)) \left.\frac{\partial \log \pi_\theta(a_t \mid s_t)}{\partial \theta}\right|_{\theta_0}\right].$$

Under the shift-policy and differentiation assumptions, this expectation also equals the expected pathwise derivative of the same return. One may prove the policy-gradient formula by differentiating the trajectory density and removing past rewards and the baseline using the zero conditional mean of the score. The pathwise formula follows by differentiating the same expectation under the parameter-independent noise law. Thus the two estimators agree in expectation, while their sample values and internal adjoints can differ.

Group normalization and the effective reward. For the outcome-only GRPO objective, let $G \geq 2$ and let $o^1, \ldots, o^G$ be conditionally independent completions from $\pi_{\theta_0}(\cdot \mid q)$. Assume a finite maximum length, a fixed support including the termination convention, bounded rewards, and differentiable positive trajectory probabilities. Use the same convention for std as in the implemented normalization. Define

$$w_i(q, o^{1:G}) := \frac{1}{T_i} \frac{r_i - \bar{r}}{\mathrm{std}(r_{1:G}) + \xi_{\mathrm{num}}}, \qquad \widetilde{R}_0(q, o) := \mathbb{E}_{o^{2:G} \sim \pi_{\theta_0}(\cdot|q)}[w_1(q, o, o^{2:G})].$$

The peers, normalization rule, and numerical stabilizer are frozen in $\widetilde{R}_0$. It is a fixed function of the focal completion, not a constant over completions. Let

$$\widetilde{J}_0(\theta) := \mathbb{E}_{q \sim P(Q),\, o \sim \pi_\theta(\cdot|q)}[\widetilde{R}_0(q, o)].$$

This definition retains the completion-length factor exactly. Variable-length completions can equivalently be represented at a finite common horizon using absorbing padding with zero subsequent log-probability contributions.

Corollary 1 (Local GRPO target and conditional value-gradient interpretation)**.** *Under the preceding assumptions, the expected reward-driven actor gradient of equation 1 at $\theta = \theta_0$, excluding its KL term, equals $\partial_\theta \widetilde{J}_0(\theta)\big|_{\theta=\theta_0}$. For a continuous version of the same group-sampling construction, suppose additionally that its effective reward and rollout satisfy the pathwise regularity conditions, and that its state transition and initial state have no explicit parameter dependence. Then this local gradient has a pathwise representation for the effective reward. Its return costates obey*

$$\mathbb{E}[\widetilde{\lambda}_t \mid s_t = s] = \frac{\partial \widetilde{V}_t(s)}{\partial s},$$

where $\widetilde{V}_t$ is the value function of the frozen-reference effective reward at the reference policy, with $\gamma = 1$. The continuous statement concerns that specified continuous model; it is not an identity between a categorical model and an arbitrary relaxation.

Proof. At the reference parameter, $\rho_{i,t}(\theta_0) = 1$. Since $\varepsilon > 0$, differentiation of the actor term at this point gives

$$g_0 = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} w_i(q, o^{1:G}) \sum_{t=1}^{T_i} \frac{\partial \log \pi_\theta(o_t^i \mid s_{i,t})}{\partial \theta}\bigg|_{\theta_0}\right].$$

The tokenwise scores sum to the trajectory score. By exchangeability of the conditionally independent group and then conditioning on its first completion,

$$\begin{aligned} g_0 &= \mathbb{E}[w_1(q, o^{1:G})\, \partial_\theta \log \pi_{\theta_0}(o^1 \mid q)] \\ &= \mathbb{E}[\widetilde{R}_0(q, o)\, \partial_\theta \log \pi_{\theta_0}(o \mid q)] \\ &= \frac{\partial}{\partial \theta} \mathbb{E}_{q,\, o \sim \pi_\theta(\cdot|q)}[\widetilde{R}_0(q, o)]\bigg|_{\theta_0}. \end{aligned}$$

Here $\partial_\theta \log \pi_{\theta_0}$ denotes the derivative of $\log \pi_\theta$ evaluated at θ_0. The peers' sampling law is not differentiated. This establishes the discrete local identity without any pathwise differentiation through sampled tokens.

For the continuous version in the corollary, $\widetilde{R}_0$ is a fixed differentiable trajectory functional. Augmenting the state with the requisite past makes it a terminal reward if necessary. Differentiation under the fixed noise law therefore represents $\partial_\theta \widetilde{J}_0$ pathwise. The proof of equation 9, applied to this reward, gives the displayed conditional costate identity. The mean score-function update and the mean pathwise update agree for this same target; their intermediate samplewise adjoints need not agree. □

Scope of the corollary. The statement is local at the reference parameter. Away from this point, clipped tokenwise ratios are not generally the trajectory likelihood ratio for $\widetilde{J}_0$. Updating the reference policy or the peer distribution also changes $\widetilde{R}_0$. The KL penalty contributes its own gradient and is not

silently absorbed into the return identity. Process-supervised, token-dependent weights require their own conditional targets; they need not admit this particular single terminal-reward representation. Independence and symmetry of the sampled peers are used in the proof and are not automatic for other grouping rules.

For parameter-dependent deterministic dynamics, the full pathwise gradient includes the explicit transition terms in equation 7; for a learned initial encoding it additionally includes $(\partial_\theta s_1)^\top \lambda_1$. A score-only actor term is not the entire derivative of a reward functional that also depends directly on these quantities. More generally, for a parameter-dependent reward on a trajectory space,

$$\frac{\partial}{\partial \theta}\mathbb{E}_{o\sim\pi_\theta}[r_\theta(o)] = \mathbb{E}[r_\theta(o)\,\partial_\theta \log \pi_\theta(o)] + \mathbb{E}[\partial_\theta r_\theta(o)],$$

where the second derivative holds the trajectory fixed. These qualifications also explain why treating a differentiable evaluation score as a detached reward can change the objective being optimized.

Normalization is not generally a scalar rescaling. Consider a one-step three-action softmax policy at the uniform distribution with rewards $(0, 1, 3)$ and group size two. The expected-reward gradient with respect to logits is

$$\frac{\partial \mathbb{E}[r]}{\partial z} = \left(-\frac{4}{9}, -\frac{1}{9}, \frac{5}{9}\right).$$

With population standard deviation and a positive stabilizer ξ, direct enumeration gives the expected group-normalized gradient

$$\frac{1}{9}\left(-\frac{1}{1+2\xi} - \frac{3}{3+2\xi}, \quad \frac{1}{1+2\xi} - \frac{2}{2+2\xi}, \quad \frac{3}{3+2\xi} + \frac{2}{2+2\xi}\right).$$

As $\xi \downarrow 0$ this tends to $(-2/9, 0, 2/9)$, which is not parallel to the original reward gradient. By continuity, the nonalignment also holds for sufficiently small strictly positive stabilizers. Equal-reward pairs have zero coefficients for every positive stabilizer. This is why the reward target in the corollary cannot generally be replaced by the original reward.

A.6 Empirical costate recursion in a transformer

Derivation of equation 15. Fix the sampled tokens and detached coefficients. For a standard causal transformer, each layer maps the collection of lower-layer activations to the next layer. Let

$$J^{(l)}_{k\leftarrow t} = \frac{\partial h^{(l+1)}_k}{\partial h^{(l)}_t}, \qquad J^{(l)}_{k\leftarrow t} = 0 \quad \text{for } k < t.$$

At the final layer, only the local log-probability depends directly on its row:

$$\hat{\lambda}^{(L)}_t = \frac{\partial L}{\partial h^{(L)}_t} = \widehat{A}_t \frac{\partial \ell_t}{\partial h^{(L)}_t}.$$

Final-layer outputs at different positions share parameters and ancestors, but neither is a differentiable parent of the other in this fixed-token graph. Consequently, for $k > t$, the activation-to-activation Jacobian $\partial h^{(L)}_k / \partial h^{(L)}_t$ is zero. Shared parameters do not create an additional edge between these outputs.

For any lower layer, reverse-mode differentiation sums over all child positions:

$$\frac{\partial L}{\partial h^{(l)}_t} = \sum_{k\geq t} \left(\frac{\partial h^{(l+1)}_k}{\partial h^{(l)}_t}\right)^\top \frac{\partial L}{\partial h^{(l+1)}_k}.$$

By causality the losses at positions before t do not contribute. This is exactly equation 15. Residual connections and positionwise normalization or feed-forward layers are included in the block Jacobians. □

A complete-state temporal representation. To obtain equation 16, choose a differentiable realization in which h_t contains all continuous quantities needed for the current distribution and subsequent deterministic computation. This includes the relevant cached keys/values and prediction features, with discrete prefix information treated as fixed side information. No required cache edge is detached in the computation being differentiated. Padding at a fixed maximum horizon gives common finite-dimensional coordinates; unused coordinates are not an omitted memory channel.

This complete state is not the single row $h_t^{(L)}$. Define the discounted suffix surrogate

$$\mathcal{L}_t := \sum_{k=t}^{T} \gamma^{k-t} \widehat{A}_k \ell_k, \qquad \hat{\lambda}_t := \frac{\partial \mathcal{L}_t}{\partial h_t}.$$

Differentiating $\mathcal{L}_t = \widehat{A}_t \ell_t + \gamma \mathcal{L}_{t+1}$ with tokens and weights fixed gives

$$\hat{\lambda}_t = \widehat{A}_t \frac{\partial \ell_t}{\partial h_t} + \gamma \left(\frac{\partial f_\theta(h_t, e_{o_t})}{\partial h_t} \right)^\top \hat{\lambda}_{t+1}.$$

For ordinary outcome-only GRPO, $\gamma = 1$. Group and length averaging multiply the trajectory contributions as in equation 1; the effective-target derivation above retains their possible dependence on the sampled completion.

The temporal recurrence and the layer–position recurrence are both valid for their respective computational variables. They do not identify a final-layer row with the complete recurrent state. Implementations that detach a cached state during the training backward pass define a different gradient graph and must use the Jacobians of that graph.

A.7 Proof of Proposition 2: sampling-gap bound

Proof. Fix the state and the explicit relaxation

$$p = \mathrm{softmax}(z/\tau_{\mathrm{rel}}), \qquad \tau_{\mathrm{rel}} > 0, \qquad e = E_{\mathrm{emb}} p.$$

Its Jacobian is

$$\frac{\partial p}{\partial z} = \frac{1}{\tau_{\mathrm{rel}}} \left(\mathrm{diag}(p) - pp^\top \right).$$

The matrix in parentheses is positive semidefinite, so its operator norm is at most its trace:

$$\left\| \mathrm{diag}(p) - pp^\top \right\| \le 1 - \sum_v p_v^2 \le 1.$$

Also, $-\log x \ge 1 - x$ for $0 < x \le 1$. Multiplying by p_v and summing, with the usual continuous convention at zero, gives

$$H(p) = -\sum_v p_v \log p_v \ge \sum_v p_v (1 - p_v) = 1 - \sum_v p_v^2.$$

Thus

$$\left\| \frac{\partial p}{\partial z} \right\| \le \frac{\min\{1, H(p)\}}{\tau_{\mathrm{rel}}}.$$

Apply the chain rule at the same relaxed state and embedding:

$$F_t^{\mathrm{rel}} - J_t^{\mathrm{stop}} = \frac{\partial f_\theta}{\partial e} E_{\mathrm{emb}} \frac{\partial p_t}{\partial z_t} \frac{\partial z_t}{\partial h_t}.$$

Submultiplicativity of the operator norm and the bound C_0 prove the first inequality in equation 20. Since $\min\{1, H\} \le \sqrt{H}$ for $H \ge 0$, the second inequality follows with

$$C = \frac{C_0 \sqrt{\log |V|}}{\tau_{\mathrm{rel}}}, \qquad |V| \ge 2.$$

The constants include the transition sensitivity, the embedding and logit maps, vocabulary size, and the fixed relaxation temperature. They cannot be described as depending only on embedding norms. □

What this proposition does not bound. The categorical draw is not replaced by an exact differentiable sampling map. For instance, if $U \sim \mathrm{Uniform}(0,1)$ and $a(h,U) = \mathbf{1}\{U < \sigma(h)\}$, its samplewise derivative is zero almost everywhere for fixed h, while

$$\frac{\partial}{\partial h}\mathbb{E}[a(h,U)] = \sigma(h)(1-\sigma(h)).$$

The score-function estimator accounts for this change in probability. There is no missing categorical pathwise derivative that Proposition 2 recovers.

If a Gumbel perturbation or another exogenous noise is included in the soft logits, the same matrix calculation applies conditionally on that noise using the entropy of the realized soft vector. Its entropy cannot be replaced without proof by the entropy of the unperturbed categorical policy. Likewise, changing the relaxation temperature changes both the entropy and the explicit $1/\tau_{\mathrm{rel}}$ factor. Low entropy therefore controls only the stated absolute local Jacobian, not relative accuracy, directional alignment, or the complete surrogate-to-return discrepancy.

A.8 Proof of Proposition 3: attention-pathway rank and magnitude

Proof. Consider one causal attention head with input vectors x_j and output at position i given by

$$q_i = W_Q x_i, \qquad k_j = W_K x_j, \qquad v_j = W_V x_j,$$
$$\alpha_{ij} = \frac{\exp(q_i^\top k_j/\sqrt{d_k})}{\sum_{u\le i}\exp(q_i^\top k_u/\sqrt{d_k})}, \qquad y_i = \sum_{j\le i}\alpha_{ij} v_j.$$

Fix an earlier input position $j < i$, so perturbing x_j does not perturb q_i within this sublayer. Differentiating both the values and the attention weights yields

$$\frac{\partial y_i}{\partial x_j} = \alpha_{ij} W_V + \frac{\alpha_{ij}}{\sqrt{d_k}}(v_j - y_i) q_i^\top W_K.$$

Hence

$$\left\| \frac{\partial y_i}{\partial x_j} \right\| \le \alpha_{ij}\left(\|W_V\| + \frac{\|v_j - y_i\|\,\|q_i\|\,\|W_K\|}{\sqrt{d_k}} \right).$$

The key-derivative contribution is necessary; treating attention weights as constants would omit it.

A head output has dimension d_{head}, so its cross-position Jacobian has rank at most d_{head}. Concatenating n_{heads} head outputs and applying the output projection gives rank at most $\min\{d_{\mathrm{model}}, n_{\mathrm{heads}} d_{\mathrm{head}}\}$. A norm bound follows by multiplying by the output-projection norm and, for example, summing the headwise bounds. Positionwise normalization before the attention sublayer multiplies the displayed Jacobian on the right by its own Jacobian; subsequent positionwise maps multiply on the left. Their norms must be included in a block-level bound.

Across multiple layers, derivatives are sums of products over the actual layer–position paths. Rank and magnitude therefore cannot be inferred from a single attention heatmap or from the number of layers alone. The trivial rank bound for a full-stack map between two d_{model}-dimensional activations is d_{model}; the single-sublayer proposition makes no stronger full-stack claim. □

A.9 Proof of Theorem 1: conditional approximation of BPTT costates

Proof. Fix a time t, its discrete history, and a complete continuous starting state $h_t = h$. On a common probability space, sample a discrete suffix and a relaxed suffix with their specified marginal laws, both initialized at h. The relaxed future-noise law is the one in Definition 1. Intermediate states may differ; their tangent coordinates are identified in the stated common state space. The theorem is a comparison for this coupled pair of suffixes, not an assumption that their future states have the same law.

Write the two recursions along their respective suffixes as

$$\hat{\lambda}_k = b_k + \gamma J_k^\top \hat{\lambda}_{k+1}, \qquad \lambda_k = d_k + \gamma F_k^\top \lambda_{k+1}.$$

Here b_k is the actual detached-coefficient score source in the complete-state surrogate; d_k is the total local reward derivative of the specified relaxed return. No equality between these sources is assumed.

Set $e_k = \hat{\lambda}_k - \lambda_k$ and

$$\eta_k := b_k - d_k + \gamma(J_k - F_k)^\top \lambda_{k+1}.$$

Subtracting the recursions gives the exact pathwise identity

$$e_k = \eta_k + \gamma J_k^\top e_{k+1}, \qquad \|\eta_k\| \le \|b_k - d_k\| + \gamma \|J_k - F_k\| \, \|\lambda_{k+1}\| = \epsilon_k.$$

Define the forward product

$$\Phi_{t,k} := J_{k-1} \cdots J_t \quad (k > t), \qquad \Phi_{t,t} := I.$$

Backward substitution, with both terminal adjoints zero, yields

$$e_t = \sum_{k=t}^{T} \gamma^{k-t} \Phi_{t,k}^\top \eta_k.$$

This ordering is important: the reverse-mode product is $J_t^\top \cdots J_{k-1}^\top$, not an arbitrary power of one sampled Jacobian. Consequently,

$$\|e_t\| \le \sum_{k=t}^{T} \gamma^{k-t} \left(\prod_{j=t}^{k-1} \|J_j\| \right) \epsilon_k.$$

All factors here can be random and mutually dependent.

The correct relaxed marginal and the common initialization allow equation 9 to be applied at the starting state:

$$\mathbb{E}[\lambda_t \mid h_t = h] = G_t^{\mathrm{rel}}(h).$$

Conditional Jensen and the preceding pathwise inequality now give

$$\begin{aligned} \|\mathbb{E}[\hat{\lambda}_t \mid h_t] - G_t^{\mathrm{rel}}(h_t)\| &= \|\mathbb{E}[e_t \mid h_t]\| \\ &\le \mathbb{E}[\|e_t\| \mid h_t] \\ &\le \mathbb{E}\left[\sum_{k=t}^{T} \gamma^{k-t} \left(\prod_{j=t}^{k-1} \|J_j\| \right) \epsilon_k \,\middle|\, h_t \right]. \end{aligned}$$

This is equation 22. No random Jacobian is taken outside a conditional expectation, and no independence between a Jacobian and a future costate is required. □

Terminal and uniform-bound checks. At $t = T$, the bound retains $\mathbb{E}[\|b_T - d_T\| \mid h_T]$. Thus it does not assert exact agreement at the terminal step without a corresponding source condition. If nonzero terminal payoffs are represented through terminal adjoints rather than included in the last reward, the exact expansion has the additional term

$$\gamma^{T+1-t} \Phi_{t,T+1}^\top (\hat{\lambda}_{T+1} - \lambda_{T+1}),$$

and its norm must be included.

If $\|J_j\| \le M$ and $\epsilon_k \le \bar{\epsilon}$ almost surely, the theorem implies the simpler sufficient bound

$$\|\mathbb{E}[\hat{\lambda}_t \mid h_t] - G_t^{\mathrm{rel}}(h_t)\| \le \bar{\epsilon} \sum_{r=0}^{T-t} (\gamma M)^r.$$

This is at most $\bar{\epsilon}/(1-\gamma M)$ when $\gamma M < 1$, and is $(T - t + 1)\bar{\epsilon}$ when $\gamma M = 1$. Small local residuals need not imply a small long-horizon bound if the Jacobian products grow rapidly.

Separating the sampling component from the full transition residual. Let (h_k, e_{a_k}) be the discrete evaluation point and $(\widetilde{h}_k, \widetilde{e}_k)$ the relaxed one. With all quantities expressed in the common coordinates,

$$\begin{aligned} F_k - J_k = &\left(F_k - J_k^{\mathrm{stop}}(\widetilde{h}_k, \widetilde{e}_k)\right) \\ &+ \left(J_k^{\mathrm{stop}}(\widetilde{h}_k, \widetilde{e}_k) - J_k^{\mathrm{disc}}(h_k, e_{a_k})\right). \end{aligned}$$

Proposition 2 controls the first parenthesis when the stated soft relaxation and norm assumptions hold. The second includes different sampled embeddings, different evaluation states, and any other difference between the two deterministic realizations. It requires additional bounds or measurements. Similarly, $b_k - d_k$ can reflect advantage normalization, target choice, reward derivatives, and estimator fluctuations. A score-function/pathwise equality in expectation for a particular target does not make this entire pathwise residual vanish.

The theorem is therefore a sufficient error decomposition, not a proof that every pretrained transformer is close to the relaxed regime. Its conditional-mean left-hand side can also be much smaller than its right-hand side because the latter bounds norms before averaging.

A.10 Proof of the usable-signal lower bound

Proof. Let

$$u_t^{(m)} := \mathbb{E}[\hat{\lambda}_t^{(m)} \mid h_t], \qquad v_t := G_t.$$

Then

$$\langle u_t^{(m)}, v_t \rangle = \|v_t\|_2^2 + \langle u_t^{(m)} - v_t, v_t \rangle.$$

By Cauchy–Schwarz,

$$\langle u_t^{(m)} - v_t, v_t \rangle \geq -\|u_t^{(m)} - v_t\|_2 \, \|v_t\|_2.$$

Therefore,

$$\langle u_t^{(m)}, v_t \rangle \geq \|G_t\|_2^2 - \|G_t\|_2 \left\| \mathbb{E}[\hat{\lambda}_t^{(m)} \mid h_t] - G_t \right\|_2.$$

Taking expectation over q, τ, t gives

$$\mathbb{E}_{q,\tau,t}\left[\left\langle \mathbb{E}[\hat{\lambda}_t^{(m)} \mid h_t], G_t \right\rangle \right] \geq \Sigma(\theta) - \Lambda_m(\theta),$$

with $\Sigma(\theta)$ and $\Lambda_m(\theta)$ defined in equation 25. This proves equation 24. □

A.11 From hidden-state alignment to an actual parameter update

The usable-signal inequality is an inner-product inequality in specified hidden-state coordinates. It is exact in those coordinates, but does not alone imply a corresponding lower bound on the improvement of the model's expected reward. This distinction does not alter the diagnostic definition equation 26.

To see the missing geometry, consider a differentiable computation with a nonredundant intermediate vector $h = h(\theta)$ such that both a target F and a surrogate L depend on θ only through h. Let

$$K = \frac{\partial h}{\partial \theta}, \qquad v = \frac{\partial F}{\partial h}, \qquad u = \frac{\partial L}{\partial h}.$$

Then $\partial_\theta F = K^\top v$ and $\partial_\theta L = K^\top u$. For an update preconditioner P, their first-order parameter-space alignment is

$$\left(\frac{\partial F}{\partial \theta}\right)^\top P \frac{\partial L}{\partial \theta} = v^\top K P K^\top u,$$

not generally $v^\top u$. The intermediate vector must be an actual nonredundant cut of the computation, not a concatenation of dependent intermediate nodes whose gradients would be counted twice. Explicit parameter dependencies require their own terms. For stochastic objectives, the full expected parameter gradients and their joint dependence on sampled computation must be used; the deterministic identity is an illustration of the geometry rather than an automatic factorization of these expectations.

For a direct parameter-space statement, let $F(\theta)$ have an L_F-Lipschitz gradient on the region traversed by a step, and let g be the actual random update direction before a fixed preconditioner P. Conditional on the checkpoint and the fixed optimizer state, write $\bar{g} = \mathbb{E}[g]$ and suppose the required second moment is finite. Smoothness gives

$$F(\theta + \eta P g) - F(\theta) \geq \eta \, \partial_\theta F(\theta)^\top P g - \frac{L_F \eta^2}{2} \|Pg\|^2.$$

Taking expectations yields

$$\mathbb{E}[F(\theta+\eta Pg)-F(\theta)] \geq \eta\,\partial_\theta F(\theta)^\top P\bar{g} - \frac{L_F\eta^2}{2}\left(\|P\bar{g}\|^2 + \mathbb{E}\|P(g-\bar{g})\|^2\right).$$

This follows by integrating the gradient along the step and using the Lipschitz bound, followed by the variance decomposition of the squared norm. It shows why step size, curvature, optimizer geometry, and estimator variance matter in addition to hidden-state alignment. A general optimizer can instead be analyzed through its actual update vector; the displayed preconditioned form applies only when the update has that form.

In particular, $S_m(\theta)$ may be negative: it is a lower-bound diagnostic and has not been clipped to a positive progress guarantee. The empirical product equation 29 requires its own finite-budget validation. It does not follow by multiplying the one-step improvement inequality by a headroom statistic.

A.12 Variational meaning of reachable headroom

Fix a prompt, let $\mu = \pi_\theta(\cdot \mid q)$ be its trajectory law, and take a bounded measurable return R and $\alpha > 0$. Set

$$Z = \mathbb{E}_\mu[e^{\alpha R}], \qquad \frac{d\nu^\star}{d\mu} = \frac{e^{\alpha R}}{Z}.$$

For any $\nu \ll \mu$ with finite relative entropy,

$$\mathrm{KL}(\nu\|\nu^\star) = \mathrm{KL}(\nu\|\mu) - \alpha\mathbb{E}_\nu[R] + \log Z.$$

Rearranging and minimizing the nonnegative left-hand side gives

$$\frac{1}{\alpha}\log Z = \sup_{\nu\ll\mu}\left\{\mathbb{E}_\nu[R] - \frac{1}{\alpha}\mathrm{KL}(\nu\|\mu)\right\},$$

with equality at $\nu^\star$. Subtracting the original mean return and averaging over prompts proves that equation 27 is the best *KL-penalized* gain from unrestricted reweighting of each prompt's existing support. It need not be attained by a shared parameterized language model under a finite optimization budget.

Jensen's inequality implies $\mathcal{H}_\alpha \geq 0$. For a fixed prompt it vanishes exactly when R is constant μ-almost surely. If uniformly bounded reward distributions converge to any constant, bounded convergence gives vanishing headroom at fixed α. This justifies the low-signal and saturation intuition under reward-concentration assumptions, not for arbitrary changing reward distributions.

For example, for a genuine Bernoulli rollout reward with success probability p, the promptwise quantity is

$$\mathcal{H}_\alpha(p) = \frac{1}{\alpha}\log(1-p+pe^\alpha) - p.$$

It is positive for $0 < p < 1$ and vanishes at both endpoints. This does not by itself imply that a particular pretraining sequence or its RL-gain curve is hump-shaped. That requires how its probabilities, gradients, and optimization dynamics actually evolve.

For bounded rewards, expansion near $\alpha = 0$ also gives

$$\mathcal{H}_\alpha(q) = \frac{\alpha}{2}\,\mathrm{Var}_\mu(R) + O(\alpha^2).$$

Thus the quantity depends on the chosen reward scale and on α; these must be fixed when comparing checkpoints.

A.13 Theoretical scope of the experimental diagnostics

The figures and numerical correlations in §6 describe checkpoint-dependent diagnostics. Relating those measurements to the formal quantities requires specifying the training reward, the gradient target, the sampling axes, and the full bound assumptions. The following distinctions delimit their theoretical interpretation.

A differentiable evaluation score does not specify a rollout reward. The stated score is $R(\theta; q) = p_\theta(y^\star \mid q)$. It is useful as a differentiable measure of a checkpoint's output, but it has no sampled-completion argument. Taken literally as the same reward for every completion of a fixed prompt at a fixed checkpoint, it gives

$$r_1 = \cdots = r_G = c, \qquad \widehat{A}_{i,t} = \frac{c - c}{0 + \xi_{\text{num}}} = 0.$$

The reward-driven GRPO actor term is then identically zero. A separate KL term can still contribute an update, but this is not a reward-driven GRPO signal. Without centering, the expected detached-score term is also zero:

$$\mathbb{E}_{o \sim \pi_\theta(\cdot|q)}[R(\theta; q)\, \partial_\theta \log \pi_\theta(o \mid q)] = R(\theta; q)\, \partial_\theta \sum_o \pi_\theta(o \mid q) = 0.$$

Directly differentiating $R(\theta; q)$ is a different update. Alternatively, a completion-dependent reward, such as sampled label correctness, can have $p_\theta(y^\star \mid q)$ as its expected score. The experimental description does not specify which such completion-dependent training reward or direct-gradient rule was used, so neither alternative is silently substituted here. This implementation detail must be resolved before the reported runs are identified with the reward-driven GRPO mechanism analyzed in this paper.

The derivative $\partial R(\theta; q)/\partial h_t$ is therefore called a differentiable-score gradient in §6. Identifying it with a trajectory value gradient requires a specified rollout objective. In a one-step label-correctness problem it can indeed be the exact derivative of the expected reward, but that is conditional on using that particular rollout reward and intervention, not on differentiability alone.

Conditional bias and samplewise discrepancy are different quantities. For a fixed value-gradient field $G_t = G_t(h_t)$, let

$$\widetilde{\Lambda}_m := \mathbb{E}\big[\|G_t\|\, \|\hat{\lambda}_t^{(m)} - G_t\|\big].$$

Conditional Jensen gives

$$\widetilde{\Lambda}_m \geq \mathbb{E}\big[\|G_t\|\, \|\mathbb{E}[\hat{\lambda}_t^{(m)} \mid h_t] - G_t\|\big] = \Lambda_m.$$

Thus the samplewise statistic described in §6 is not an unbiased estimate of the conditional-bias quantity equation 25. It includes Monte Carlo variability and supplies a conservative population surrogate for the same fixed field: $\Sigma - \widetilde{\Lambda}_m \leq S_m$. It can be large even when the conditional mean is exactly correct. Matching prompts and hidden states does not establish equality between a discrete objective and a relaxed objective, nor does it remove this variance contribution.

Within-prompt headroom and variation across prompts. If $R(\theta; q)$ is constant over the sampled trajectories for fixed q, its within-prompt headroom in equation 27 is exactly

$$\frac{1}{\alpha} \log e^{\alpha R(\theta;q)} - R(\theta; q) = 0.$$

Computing $\alpha^{-1} \log \mathbb{E}_q[e^{\alpha R(\theta;q)}] - \mathbb{E}_q[R(\theta; q)]$ instead can give a nonzero number, but it measures variation across prompts, not the within-prompt reweighting headroom defined in the theory. The sampling axis in the reported estimator must therefore be specified before its nonzero plotted values are given the interpretation of equation 27.

Interpreting the entropy diagnostic. Proposition 2 specifies a local Jacobian bound for a particular soft relaxation, with temperature and surrounding operator norms explicit. Theorem 1 additionally uses local-signal errors, coupled evaluation-point errors, future costate norms, and products of Jacobian norms. A costate-distance curve and a rescaled entropy curve alone do not establish this inequality. Reporting the actual operators, coordinates, coupling, temperature, and full residual terms is necessary for theorem-level verification. No fitted constant can be interpreted as a proved uniform upper bound without the corresponding norm control.

Calibration and empirical interpretation. The reported Spearman coefficients and plots remain descriptive observations. When an affine calibration uses the same checkpoints on which it is evaluated, agreement is not an out-of-sample prediction. A predictive evaluation requires separating calibration from evaluation checkpoints, keeping the task and optimization budget controlled, and comparing with competence-only baselines. The reported results do not by themselves establish a causal explanation of critic-free RL, the smallness of every term in equation 22, or the finite-budget product law as a theorem.