WHAT DOES ON-POLICY DISTILLATION OPTIMIZE?

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation uses teacher feedback on student-generated prefixes, so an update changes both the student and the distribution of its later supervision. Yet practical methods often replace sequence gradients with token-local, bounded, support-restricted, or future-truncated updates. This raises a basic question: *what objective do these updates represent?* We answer this question at three levels: First, every continuous coordinate-local reward admits an explicit separable potential; we characterize its proper and Bregman subclasses. Next, universal autoregressive composition within the separable Bregman class forces the logarithmic generator, exposing a tradeoff between bounded rewards and sequence consistency. Finally, under student occupancy, an exact future-credit decomposition yields an intrinsic objective-fidelity coefficient, a finite-step descent guarantee, and circulation tests for integrability. To validate these results, exact finite trees verify these identities and separate composition from update fidelity. Complementing these exact checks, Qwen3.5 measurements on GSM8K, ARC-Challenge, and HellaSwag, using 0.8B-to-4B and 4B-to-9B model pairs, reveal horizon-dependent transitions from descent to near orthogonality or ascent. Matched training further shows that cached-prefix teacher fit and task utility can rank the same updates differently. Together, these results identify the objective each surrogate represents, the sequence terms it omits, and when its local update preserves progress.

1 INTRODUCTION

On-policy distillation (OPD) asks a teacher to supervise prefixes generated by the current student (Agarwal et al., 2024; Gu et al., 2024). This reduces the distribution gap from fixed teacher trajectories. However, it also makes supervision depend on the parameters being updated: a token chosen now determines where all later teacher feedback is evaluated.

Yet in practice, modern systems seldom optimize the resulting path-space objective literally. Sequence reverse KL admits a return-to-go policy-gradient estimator, but its long-horizon variance can be large (Czarnecki et al., 2019; Xu et al., 2026). Practical methods therefore keep only instantaneous token rewards, restrict the teacher distribution to a local top-K support, or replace the unbounded log ratio by a bounded transform (Zhao et al., 2026). Although these changes can improve numerical stability, their objective meaning remains unclear.

To resolve this ambiguity, we distinguish three mathematically different questions often grouped under *optimizing KL*. First, does the token reward define a proper objective when a prefix is fixed? Second, do these prefix objectives compose into a divergence over complete responses? Third, after the student itself supplies the prefix distribution, is the implemented update the gradient of that objective? Fig. 1 separates these levels.

At the first level, every continuous coordinate-local reward has an explicit separable potential. A simple sign condition makes it strictly proper, and the transform-then-subtract form $h(p_i) - h(q_i)$ characterizes its Bregman subclass. Consequently, strictly increasing bounded transforms retain a proper fixed-prefix objective (Fig. 1a).

Next, sequence composition imposes a stronger requirement. Classical scoring-rule and sufficiency results single out information divergence among Bregman divergences (Gneiting & Raftery, 2007; Harremoës, 2015; Ovcharov, 2018). We specialize this boundary to autoregressive OPD: logarithmic rewards inherit the path-space chain rule, whereas a bounded transform generally defines a sum of

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.

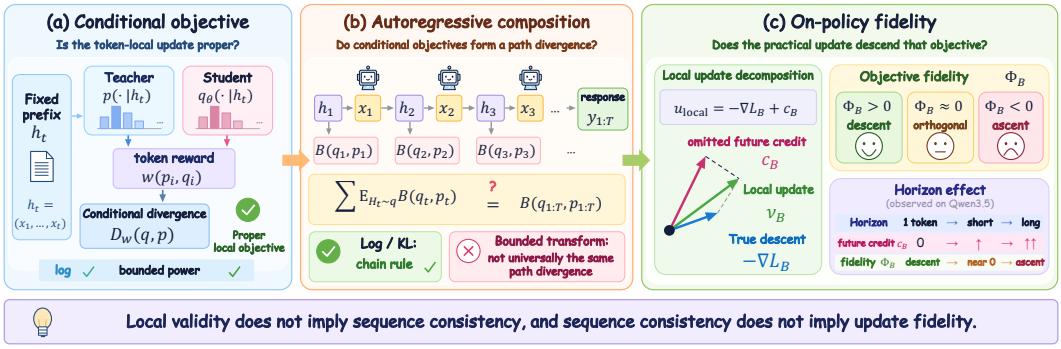

Figure 1: Three levels of OPD objective semantics: (a) proper conditional objectives at fixed prefixes; (b) universal composition into the same divergence over complete responses, satisfied by logarithmic KL but generally not bounded transforms; and (c) update fidelity under student occupancy, where omitted future credit separates local updates from exact descent.

occupancy-weighted conditional objectives instead of the same Bregman divergence over complete responses.

Finally, identifying a sequence objective does not ensure that a local update follows its gradient (Fig. 1c). Let L_B be the expected sum of a conditional divergence B_t under student trajectories. Its gradient contains the direct derivative of each B_t and a score term recording how an earlier action changes later prefixes. The local update keeps the direct term, and we show exactly that

$$u_{\mathrm{local}} = -\nabla L_B + c_B, \tag{1}$$

where c_B is the omitted future-credit vector. This identity induces an intrinsic objective-fidelity coefficient. It also yields a measurable antisymmetric Jacobian whose nonzero circulation certifies that the local field has no scalar potential in the measured region.

Accordingly, we first use exact finite trees to verify the identities and isolate occupancy feedback. We then move to Qwen3.5, measuring logarithmic, power, and local-support objectives on common trajectories and parameter directions. Across GSM8K, ARC-Challenge, and HellaSwag, local-log fidelity moves from descent toward near orthogonality or ascent as longer horizons expose more future credit. Increasing the model pair from 0.8B-to-4B to 4B-to-9B shifts this transition to longer horizons. Finally, matched training compares local, corrected, sequence Monte Carlo, bounded, and support-restricted updates, linking the measured geometry to optimization behavior.

These connections lead to three contributions:

- We characterize coordinate-local OPD potentials, their proper and Bregman subclasses, and the logarithmic boundary for universal composition.
- We derive objective fidelity, finite-step descent, future-neutrality, support omission, circulation, and approximate-integrability bounds.
- We validate the theory in exact systems and across three Transformer tasks and two model scales, then connect fidelity to training behavior.

2 Related Work

On-policy and sequence distillation. Classical knowledge distillation matches teacher predictions, with extensions to policy transfer and analyses of functional agreement and optimization variance (Hinton et al., 2015; Rusu et al., 2015; Stanton et al., 2021; Safaryan et al., 2023). In parallel, sequence-level distillation extends the unit from a token to a generated response (Kim & Rush, 2016). GKD and MiniLLM establish student-prefix distillation with generalized or reverse KL objectives (Agarwal et al., 2024; Gu et al., 2024). Sequence-level f-divergence decompositions show that prefix occupancy and conditional divergence must be paired correctly (Wen et al., 2023). DistiLLM trades exact on-policy computation for practical skewed or hybrid estimators (Ko et al., 2024). Our object is the population update represented after these approximations.

Objective consistency and credit assignment. Imitation learning and scheduled sampling expose distribution shift when training and rollout prefixes differ (Ross et al., 2011; Bengio et al., 2015; Huszár, 2015; Rajaraman et al., 2021). Decision-dependent stochastic optimization and state-dependent performative prediction formalize the broader coupling between parameters and their data distribution (Drusvyatskiy & Xiao, 2023; Li & Wai, 2022). Likelihood-ratio gradients provide the standard mechanism for assigning future consequences to sampled actions (Williams, 1992). Czarnecki et al. (2019) prove that a student-sampled local-loss field need not be conservative and derive a future-loss correction. β-OPSD applies return-to-go credit to a logarithmic sequence objective (Xu et al., 2026). Building on these foundations, we study bounded and support-restricted surrogates through one conditional, compositional, and differential framework.

Proper scores and divergence geometry. Proper scoring rules correspond to convex entropy and Bregman regret constructions (Gneiting & Raftery, 2007; Ovcharov, 2018). Sufficiency and locality distinguish logarithmic information divergence within this family (Harremoës, 2015). We use these classical results to state the sequence-composition boundary accurately, then derive OPD-specific fidelity and circulation quantities under parameter-dependent occupancy.

Stabilized OPD. Sampled-token log ratios expose a bias and variance tradeoff when future coupling is removed. Local support matching changes the represented token distribution, while bounded power rewards preserve tokenwise sign and sharply reduce gradient magnitude (Zhao et al., 2026). These properties concern stability and conditional geometry. Accordingly, our analysis determines when they also preserve progress on a declared sequence objective.

3 Three Levels of Objective Semantics

We now formalize the three levels introduced in Fig. 1: conditional propriety, autoregressive composition, and update integrability. Appendix A consolidates the notation used below and extends the setup to variable horizons, prompt distributions, and shared parameterizations.

Let $\mathcal{X}$ be a prompt space, $\mathcal{Y}$ a finite token alphabet, and T a finite response horizon. For one prompt x, the student and teacher factor as

$$q_\theta(y_{1:T} \mid x) = \prod_{t=1}^{T} q_\theta(y_t \mid h_t), \qquad p(y_{1:T} \mid x) = \prod_{t=1}^{T} p(y_t \mid h_t). \tag{2}$$

where $h_t = (x, y_{<t})$. The teacher is fixed during one update. The score of the sampled token at position t is

$$s_t(\theta, Y) = \nabla_\theta \log q_\theta(Y_t \mid H_t). \tag{3}$$

Definition 3.1 (Conditional propriety)**.** A token update is conditionally proper if, for every fixed prefix distribution, its population field is the negative gradient of a nonnegative conditional divergence minimized uniquely at $q_\theta(\cdot \mid h) = p(\cdot \mid h)$.

Definition 3.2 (Autoregressive composition)**.** A conditional divergence B composes if the corresponding divergence between complete responses equals the student-prefix expectation of the conditional divergences. Write $q_t = q(\cdot \mid H_t)$ and $p_t = p(\cdot \mid H_t)$. Then

$$B(q_{1:T}, p_{1:T}) = \sum_{t=1}^{T} \mathbb{E}_{H_t \sim q} B(q_t, p_t). \tag{4}$$

Definition 3.3 (Update integrability)**.** A continuously differentiable population update $u : \Theta \to \mathbb{R}^d$ is integrable on Θ if a scalar F exists with $u = -\nabla F$ throughout the domain.

These definitions concern different properties. A power reward is conditionally proper, yet its Bregman divergence need not satisfy Equation 4. A logarithmic conditional divergence does compose, yet a token-local update under student occupancy omits the derivative of that occupancy and can fail integrability.

To make the third question explicit, for a differentiable conditional divergence

$$B_t(\theta, h) = B(q_\theta(\cdot \mid h), p(\cdot \mid h)), \tag{5}$$

define the occupancy-weighted objective

$$L_B(\theta) = \mathbb{E}_{Y\sim q_\theta} \sum_{t=1}^{T} B_t(\theta, H_t). \tag{6}$$

When B is reverse KL, Equation 4 makes L_B the sequence reverse KL. For other B, Equation 6 remains a scalar objective with an explicit operational meaning. The next section asks when the practical token-local update actually follows the gradient of this objective.

4 Objective Fidelity in OPD

4.1 Proper conditional rewards

Begin with an arbitrary coordinate-local reward $w(p_i, q_i)$. Define its conditional potential by

$$D_w(q, p) = -\sum_{i\in\mathcal{Y}} \int_{p_i}^{q_i} w(p_i, z)\, dz. \tag{7}$$

Theorem 4.1 (Coordinate-local potential)**.** *For every continuous w and differentiable categorical model,*

$$\mathbb{E}_{I\sim q_\theta}\left[w(p_I, q_{\theta,I})\nabla_\theta \log q_{\theta,I}\right] = -\nabla_\theta D_w(q_\theta, p). \tag{8}$$

If $w(p, p) = 0$ and $(p - z)w(p, z) > 0$ for $z \neq p$, then D_w is strictly proper. Moreover, D_w is a separable Bregman divergence for every alphabet if and only if $w(p, z) = h(p) - h(z)$ for a single function h. Two rewards that differ by an action-independent baseline induce the same parameter field.

The first statement shows that coordinate locality itself is enough for a fixed-prefix scalar potential. Properness and compatibility with one common Bregman geometry impose additional restrictions. Appendix B proves the full characterization.

Let $h : (0, 1) \to \mathbb{R}$ be continuously differentiable and let $\psi' = h$. For $p, q \in \Delta^\circ(\mathcal{Y})$, define

$$B_\psi(q, p) = \sum_{i\in\mathcal{Y}} \big[\psi(q_i) - \psi(p_i) - h(p_i)(q_i - p_i)\big]. \tag{9}$$

The sampled-token reward is

$$w_h(i; p, q) = h(p_i) - h(q_i). \tag{10}$$

Corollary 4.2 (Transform Bregman objective)**.** *At a fixed prefix,*

$$\mathbb{E}_{I\sim q_\theta}\left[w_h(I; p, q_\theta)\nabla_\theta \log q_{\theta,I}\right] = -\nabla_\theta B_\psi(q_\theta, p). \tag{11}$$

If h is increasing, B_ψ is nonnegative. If h is strictly increasing, it is strictly proper on the simplex interior.

The theorem separates boundedness from conditional propriety. The logarithmic choice gives reverse KL. The power choice $h(z) = z^\alpha$, $\alpha > 0$, gives a strictly proper power-generated Bregman divergence with a bounded reward.

4.2 The path-space boundary

Conditional propriety alone does not provide a chain rule. The relevant classical characterization can be stated directly for the present setting.

Proposition 4.3 (Compositional boundary)**.** *Let B_ψ be a strictly proper separable Bregman divergence with a twice continuously differentiable generator on every finite alphabet with at least three symbols. If it satisfies Equation 4 for every pair of positive autoregressive distributions, then*

$$\psi(z) = az \log z + bz + c, \qquad a > 0. \tag{12}$$

Consequently, B_ψ is a positive multiple of information divergence.

Proposition 4.3 is a specialization of classical sufficiency characterizations (Harremoës, 2015). Its role is to delimit the OPD claim. It implies the following practical consequence.

Corollary 4.4 (Bounded local reward tradeoff)**.** *No bounded transform h in Equation 10 is simultaneously strictly proper at each prefix and universally compositional as the same separable Bregman divergence over complete autoregressive responses.*

Theorem 4.5 (Composition and fidelity separate)**.** *Autoregressive composition and local-update fidelity imply neither direction of one another.*

1. *A positive three-step binary model exists for which logarithmic KL has zero chain residual while the token-local field satisfies $\Phi_B < 0$.*
2. *A positive two-step binary model exists for which the power Bregman divergence has nonzero chain residual while the token-local field satisfies $\Phi_B = 1$.*

Theorem 4.5 rules out treating path-space semantics and optimization fidelity as one property. Appendix C gives the two explicit models.

4.3 Future credit and objective fidelity

Having separated composition from fidelity, we now identify their differential link. The next result holds whether or not B composes into a path-space Bregman divergence.

Theorem 4.6 (Future-credit decomposition)**.** *Assume differentiation may pass through the finite trajectory expectation. Define*

$$c_B(\theta) = \mathbb{E}_{Y\sim q_\theta} \sum_{s<t} s_s(\theta, Y) B_t(\theta, H_t). \tag{13}$$

The negative direct-gradient field

$$u_B(\theta) = -\mathbb{E}_{Y\sim q_\theta} \sum_{t=1}^{T} \nabla_\theta B_t(\theta, H_t) \tag{14}$$

satisfies

$$u_B(\theta) = -\nabla_\theta L_B(\theta) + c_B(\theta). \tag{15}$$

Appendix D proves the decomposition and its subsequent descent and neutrality consequences. The decomposition suggests a scale-free quantity.

Definition 4.7 (Objective fidelity)**.** When $\nabla L_B(\theta) \neq 0$, define

$$\Phi_B(\theta) = \frac{\langle -\nabla L_B, u_B\rangle}{\|\nabla L_B\|_2^2} = 1 - \frac{\langle \nabla L_B, c_B\rangle}{\|\nabla L_B\|_2^2}, \tag{16}$$

Proposition 4.8 (Intrinsic descent criterion)**.** *The local field is a strict descent direction for L_B exactly when $\Phi_B > 0$. Moreover,*

$$|\Phi_B - 1| \leq \frac{\|c_B\|_2}{\|\nabla L_B\|_2}, \tag{17}$$

and the bound is sharp for collinear future credit.

Proposition 4.9 (Finite-step descent)**.** *Suppose ∇L_B is M-Lipschitz along $\{\theta + t\eta u_B : 0 \leq t \leq 1\}$. If $\Phi_B > 0$, then every step size*

$$0 < \eta < \frac{2\Phi_B \|\nabla L_B\|_2^2}{M\|u_B\|_2^2} \tag{18}$$

strictly decreases L_B.

No empirical threshold enters Proposition 4.8. Zero is the sign boundary of the directional derivative. A value of one means that projected progress equals exact gradient progress; a negative value means that the practical field ascends the declared objective.

When local credit suffices. Future credit vanishes when every action at an occupied prefix has the same expected future divergence. Thus horizon exposes the opportunity for a mismatch, while action-dependent future costs determine whether it occurs. Appendix D states the exact advantage representation and the tabular converse.

4.4 Circulation of the practical field

Let $J_B = \nabla_\theta u_B$ and define its antisymmetric part

$$C_B = \frac{1}{2}(J_B - J_B^\top). \tag{19}$$

Theorem 4.10 (Score-gradient wedge)**.** *For the local field in Equation 14,*

$$2[C_B]_{ij} = -\mathbb{E}\sum_{s<t}\left[s_{s,j}\partial_i B_t - s_{s,i}\partial_j B_t\right]. \tag{20}$$

On a simply connected domain, u_B is integrable exactly when $C_B = 0$ throughout the domain.

Appendix E gives the differential proof, closed-loop expansion, and projected neural estimator.

For orthonormal directions a, b, let $\mathcal{R}_\varepsilon(a, b)$ be the oriented square of side ε. A local Stokes expansion gives

$$\oint_{\partial\mathcal{R}_\varepsilon} u_B^\top d\theta = -2\varepsilon^2 a^\top C_B b + O(\varepsilon^3). \tag{21}$$

Thus Equation 20 is estimated from one set of student trajectories, while Equation 21 supplies an independent closed-loop check. Appendix E also bounds the error of a local scalar potential when circulation is small.

4.5 What top-K matching omits

For a token subset S, let $\alpha = q(S)$, $\beta = p(S)$, and q_S, p_S denote normalized conditionals. Reverse KL satisfies

$$\begin{aligned} D_{\mathrm{KL}}(q\|p) = \mathrm{kl}(\alpha\|\beta) &+ \alpha D_{\mathrm{KL}}(q_S\|p_S) \\ &+ (1-\alpha) D_{\mathrm{KL}}(q_{S^c}\|p_{S^c}). \end{aligned} \tag{22}$$

A renormalized local-support loss keeps $D_{\mathrm{KL}}(q_S\|p_S)$ and removes the mass divergence, the outside conditional divergence, and the factor α. The omission is therefore observable from the same logits used to build the support.

Appendix G develops fixed-occupancy controls, variance and horizon consequences, and the local approximate-potential interpretation.

5 Experimental Design

The experiments test one theoretical quantity at a time. Exact systems validate identities; neural measurements estimate objective fidelity and circulation; training runs connect those quantities to optimization and task behavior.

Exact finite trees. We sample positive categorical distributions and shared-parameter binary autoregressive trees. The numerical checks cover 1,000 prefix trials and 128 trees. It compares logarithmic and power transforms, complete enumeration, finite-difference gradients, analytic wedges, and opposite loop orientations.

Transformer measurement. A Qwen3.5 0.8B student and 4B teacher use rank-eight LoRA on GSM8K, ARC-Challenge, and HellaSwag (Cobbe et al., 2021; Clark et al., 2018; Zellers et al., 2019). Sixteen seeds per task freeze prompt pools, rollouts, teacher distributions, and four parameter directions. Horizons extend from 1 to 64. A second FP32 panel uses a 4B student and 9B teacher with four seeds per task. Within the measured parameter subspaces, central differences estimate Φ_B, future credit, and the projected norm of Equation 20.

Training panel. Six matched methods use regenerated trajectories: local log ratio, Rao–Blackwellized corrected log, sequence Monte Carlo return-to-go, local power, corrected power, and top-16 local-support KL. Plain SGD removes optimizer-memory differences. GSM8K and ARC-Challenge each use eight training seeds and 128 updates with paired schedules and independently restored adapters. Final adapters receive cached-prefix, fresh-sequence, task, and endpoint-geometry evaluations.

Appendices H, I, and J give the estimands, prompts, schedules, numerical dtypes, parsing rules, and statistical units.

Table 1: One-parameter finite trees separating composition from update fidelity. Both prefix objectives are strictly proper; constructions are in Appendix C.

Quantity	**Logarithmic KL**	**Power** $(h(z) = z)$
Sequence length T	3	2
Absolute chain residual $\lvert\Delta_\psi\rvert$	2.22×10^{-16}	0.09227593
Local update u_B	-0.09030669	-0.19157799
Objective descent direction $-\nabla L_B$	0.09356865	-0.19157799
Future credit c_B	-0.18387534	0
Objective fidelity Φ_B	-0.96513829	1
Local behavior for L_B	Ascent	Exact descent

6 RESULTS

We first separate conditional validity, composition, and update fidelity, then examine how horizon, scale, support, and training alter their empirical consequences.

Observation 1: Conditional validity, composition, and update fidelity are distinct. Table 1 isolates the distinction in two one-parameter finite trees with strictly proper conditional objectives. Logarithmic KL composes over complete responses, yet its local field has $\Phi_B = -0.965$ and opposes the objective descent direction. Conversely, the power example has chain residual magnitude 0.0923 while $u_B = -\nabla L_B$ and $\Phi_B = 1$. Thus composition and update fidelity imply neither direction of one another; conditional propriety alone decides neither. These are constructive counterexamples, not averages over random trees. Separately, 1,000 prefix trials and 128 trees check numerical consistency with the analytic identities; Appendix K reports all seven recorded residuals.

Observation 2: Increasing horizon can turn local descent into sequence ascent. At one token, future credit is empty and projected fidelity is exactly one. For the 0.8B student, local-log fidelity falls on ARC from 0.166 at horizon 8 to 0.024 at 16, -0.005 at 32, and -0.0069 at 64. The horizon-64 interval is $[-0.0125, -0.0011]$. On GSM8K it falls from 0.429 at horizon 8 to 0.142 at 16 and reaches the zero boundary by horizon 32. HellaSwag gives a sharper transition: fidelity is 1.18 at horizon 4, -0.025 at 8, and -0.0098 at 32, where the interval is $[-0.0184, -0.0012]$. Thus the practical field can pass from useful alignment, through near orthogonality, to sequence ascent as omitted future credit accumulates. However, horizon alone does not force this transition: future credit can vanish even for long sequences when later divergence is action independent. The relevant interaction is between the student's score sensitivity and action-dependent future mismatch, rather than token count alone (Appendix G).

Observation 3: Larger models shift the fidelity transition to longer horizons. The scale panel tests whether the same boundary persists after increasing student and teacher capacity. With a 4B student and 9B teacher, FP32 local-log fidelity at horizon 16 is 0.129 on ARC and 0.637 on GSM8K. At horizon 64 it is 0.012 and 0.191. The corresponding future credit is 9.53 and 3.61 times the local-field norm, compared with 23.24 and 12.11 for the 0.8B student. Fig. 2 shows the paired horizon curves and the independently measured antisymmetric Jacobian. Thus future-credit magnitude alone does not determine the sign of objective fidelity.

Eight-direction measurements preserve the same ordering, so the transition is not specific to one two-dimensional plane. On HellaSwag, the eight-direction horizon-8 estimate is -0.050 with interval $[-0.093, -0.0065]$. Power objectives also remain conditionally proper while exhibiting a different fidelity curve. Complete transform, direction, and seed tables are in Appendix K; Fig. 5 isolates the future-credit and circulation terms. These comparisons use common trajectories and parameter directions, so surrogate differences are paired. The additional directions probe sensitivity to the sampled LoRA subspace; they do not recover full-parameter fidelity. Similarly, a nonzero projected wedge witnesses nonintegrability, whereas a small value cannot establish that the full field is conservative. Appendix H specifies these estimators.

Observation 4: Greater support coverage need not improve sequence fidelity. Beyond horizon and scale, the support sweep makes the omitted mass measurable. Teacher top-4 support contains 95.2% of ARC probability and 98.4% of GSM8K probability on average at horizon 16, but only 77.7% and 91.2% of student probability. Top-64 captures more than 99.7% of teacher mass and 96.0% and 99.2% of student mass. Despite this greater coverage, its horizon-16 fidelity is lower

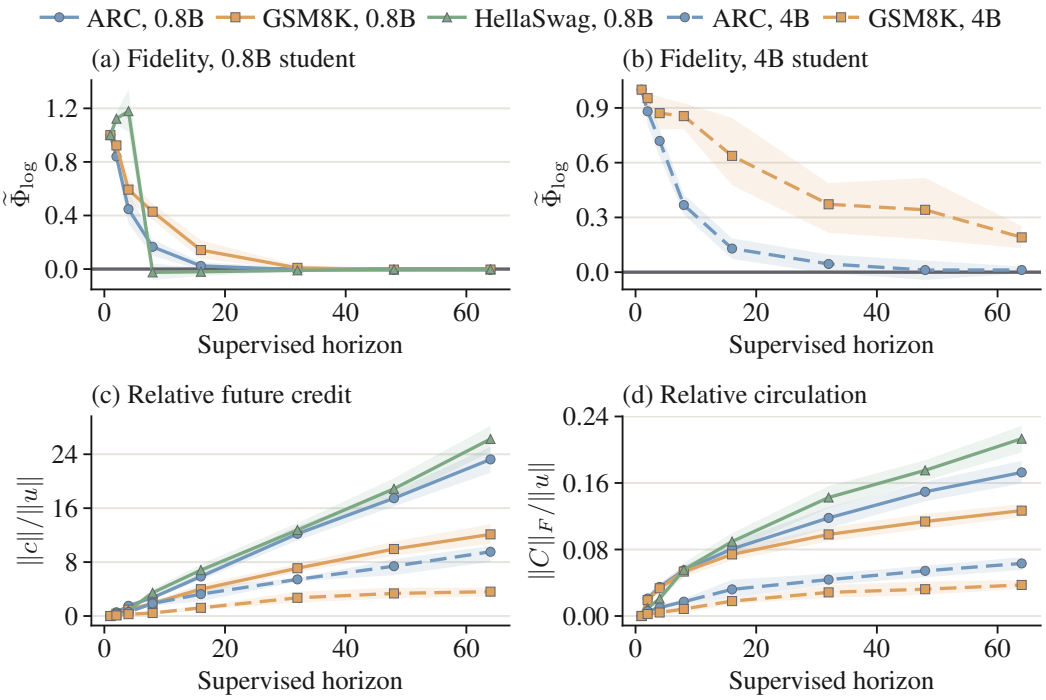

Figure 2: Horizon-dependent objective fidelity across tasks and model scales. The 0.8B student reaches task-dependent zero-fidelity boundaries. The 4B student stays aligned longer, although future credit and circulation still grow with horizon. Bands show 95% seed-bootstrap intervals.

in both tasks: 0.033 versus 0.066 on ARC and 0.073 versus 0.167 on GSM8K. Support restriction is sharper on HellaSwag: top-4 remains positive at 0.035, whereas top-64 is -0.014 at horizon 16. Support coverage therefore does not order sequence fidelity by itself. The renormalized conditional geometry and its future credit change together. Fig. 7 and Appendix F give the measured curves and the exact omitted-term identity.

Observation 5: Cached-prefix fit and task utility rank updates differently. To connect these diagnostics to optimization, the 96-run primary training matrix uses eight independent seeds for each task and method. Local log reduces the fixed-prefix evaluation teacher gap by 0.0138 on GSM8K and 0.0066 on ARC. Corrected log changes the same quantity by -0.0142 and -0.0280, while sequence log Monte Carlo changes it by -0.0017 and -0.0247. The methods therefore induce different orderings on the cached conditional objective. HellaSwag independently repeats the split: local log improves the gap by 0.0273, whereas corrected log and sequence Monte Carlo change it by -0.0205 and -0.0193.

Task behavior reverses part of that ordering. On ARC, corrected log improves accuracy over local log by 0.0566 with interval $[0.0059, 0.1094]$ and pass@4 by 0.148 with interval $[0.070, 0.234]$. Sequence log Monte Carlo improves accuracy by 0.0508 with interval $[0.0097, 0.0957]$ and pass@4 by 0.141 with interval $[0.0625, 0.219]$. On GSM8K, corrected log improves pass@4 over local log by 0.102 with interval $[0.0156, 0.188]$; its accuracy interval crosses zero. HellaSwag locates the complementary boundary: corrected log reduces accuracy relative to local log by 0.081 with interval $[-0.118, -0.0469]$ and raises the raw gradient norm from 22.6 to 161.1. Exact future credit therefore does not remove the finite-step variance tradeoff. Fig. 3 places fixed-prefix fit, task accuracy, and raw gradient scale on a common method axis.

Observation 6: Future correction increases raw update scale. On GSM8K, the raw gradient norm averages 14.6 for local log, 54.0 for sequence Monte Carlo, and 72.2 for corrected log; on ARC it averages 21.3, 123.4, and 161.6. Corrected log therefore has approximately 4.9 and 7.6 times the local-log mean norm on the two tasks. The 95% seed-bootstrap intervals support the same ordering: corrected log gives $[68.4, 76.3]$ on GSM8K and $[153.3, 169.8]$ on ARC, compared with $[13.8, 15.7]$ and $[20.6, 21.9]$ for local log. Sequence Monte Carlo falls between these methods, with intervals

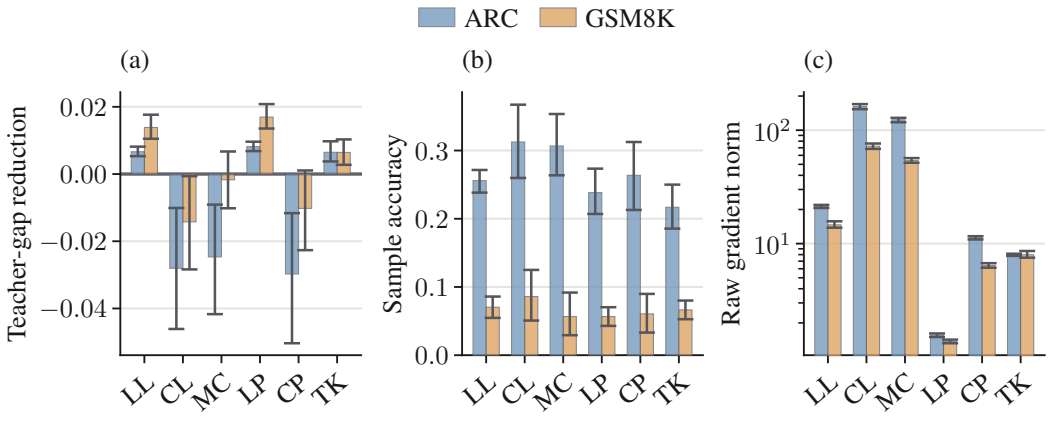

Figure 3: After 128 on-policy updates, local objectives improve cached-prefix teacher fit; future correction improves ARC task outcomes and GSM8K pass@4 in paired contrasts. Bars show means and 95% seed-bootstrap intervals. LL/CL: local/corrected log; MC: sequence log Monte Carlo; LP/CP: local/corrected power; TK: top-16 local-support KL.

$[51.7, 56.8]$ on GSM8K and $[117.4, 128.5]$ on ARC. All three raw means greatly exceed the shared clipping threshold of one. Clipping is active in every log update, so post-clipping norm is matched while stochastic direction and trajectory occupancy differ. Fig. 6 shows the 128-step dynamics; Tables 13 and 14 report the complete training and evaluation summaries.

These measurements establish a difference in raw update scale, without separating mean-gradient magnitude from stochastic variance. Their relevance follows from the finite-step analysis: for an unbiased stochastic field, fidelity controls expected first-order progress, while the smoothness penalty depends on its second moment (Appendix G). Both the mean field and stochastic fluctuations contribute to that second moment. This analysis concerns the field before clipping, rather than the clipped update itself. Consequently, restoring future credit does not by itself control update scale or the effect of clipping. This distinction connects the measured raw norms to the task-dependent outcomes in Observation 5, since a shared learning rate and clipping threshold do not equate the optimization trajectories.

Observation 7: Training changes the geometry in which later local updates act. Finally, trained endpoints confirm that the geometry itself moves. At horizon 16, ARC states produced by local log retain mean local-log fidelity 0.032, whereas corrected-log states have mean -0.021. On GSM8K the corresponding values are 0.177 and -0.031. These are measurements of the local field evaluated at each final state, not relabelings of the corrected field. At 4B scale, every method endpoint remains positive, spanning 0.256 to 0.327 on ARC and 0.502 to 0.617 on GSM8K, which independently locates the scale-dependent boundary. Every endpoint uses the same prompts, parameter directions, and finite-difference construction. The change therefore records path-dependent relocation of the occupied prefixes and local tangent geometry, rather than a change in measurement protocol. The training method determines both its immediate update and the region in which later local updates are evaluated. Fig. 8 and the paired endpoint tables in Appendix K show the full comparison.

7 Conclusion

Practical OPD connects conditional objectives, sequence composition, and optimization under student occupancy, but these properties need not coincide. Coordinate-local rewards admit a fixed-prefix potential; propriety and Bregman structure impose stronger conditions, and universal composition forces the logarithmic generator. Even then, local updates can omit future credit. Objective fidelity, finite-step descent, and circulation quantify the resulting mismatch. Consistent with this account, finite-tree checks and Transformer measurements reveal horizon-dependent transitions toward orthogonality or ascent, while matched training separates cached-prefix fit from task utility. Together, these results evaluate stabilized OPD through the objective it represents, the sequence terms it omits, and the progress its implemented update preserves. Accordingly, no single cached-prefix metric orders all objectives and task outcomes.

AI USAGE STATEMENT

We use generative AI as an aid to manuscript editing and research implementation, including language refinement, structural revision, code drafting and debugging, and figure formatting. The authors formulate the mathematical framework and equations and carry out the substantive analyses and interpretation. AI-assisted checks of mathematical expressions provide supplementary feedback on correctness and consistency; the authors independently assess this feedback and determine the final formulations. AI assistants used in preparing the work do not create or replace experimental measurements or determine scientific conclusions. Model-generated responses in the experiments instead follow the stated research protocol. The authors review assisted code against the intended computations, verify numerical claims against the recorded outputs, and revise all assisted prose. They retain full responsibility for the paper's arguments, equations, results, code, and accompanying materials.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv preprint arXiv:1506.03099*, 2015.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. doi: 10.48550/arXiv.1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Wojciech M. Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1331–1340. PMLR, 2019.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023. doi: 10.1287/moor.2022.1287.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Peter Harremoës. Proper scoring and sufficiency. *arXiv preprint arXiv:1507.07089*, 2015.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3164–3186. PMLR, 2022.

Evgeni Y. Ovcharov. Proper scoring rules and Bregman divergences. *Bernoulli*, 24(1):53–79, 2018.

Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, Jiantao Jiao, and Kannan Ramchandran. On the value of interaction and function approximation in imitation learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1325–1336, 2021.

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. PMLR, 2011.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Mher Safaryan, Alexandra Peste, and Dan Alistarh. Knowledge distillation performs partial variance reduction. *Advances in Neural Information Processing Systems*, 36, 2023.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*, 2021.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.605.

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

Jiawei Xu, Minghui Liu, Juzheng Zhang, Tom Goldstein, and Furong Huang. β-OPSD: Deriving with policy optimization, training with self-distillation. *arXiv preprint arXiv:2607.28582*, 2026.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472.

Anhao Zhao, Junlong Tong, Yingqi Fan, Ping Nie, Wenjie Li, and Xiaoyu Shen. PowerOPD: Stabilizing on-policy distillation with bounded power transformation. *arXiv preprint arXiv:2606.17199*, 2026.

A Notation and Scope

Table 2 lists the principal objects.

A.1 Finite and variable horizons

The main analysis uses a deterministic finite horizon so that every expectation is a finite sum. An end-of-sequence symbol converts variable-length generation into this setting. Once termination occurs, later conditionals are point masses on a padding state and contribute zero divergence. The same identities therefore apply to bounded variable-length responses.

A.2 Conditioning on prompts

All identities hold pointwise in x. Averaging over a prompt distribution independent of θ preserves them by linearity. If prompt selection also depends on the current student, its score contribution enters the future-credit term exactly as another occupancy variable.

Table 2: Notation used in the theory and experiments.

Symbol	Type	Meaning
x	prompt	Fixed conditioning input for one response distribution
$Y_{1:T}$	trajectory	Student-generated token sequence of horizon T
H_t	prefix	Random prefix $(x, Y_{<t})$ occupied before token t
q_θ	distribution	Current student policy
p	distribution	Teacher policy held fixed during one update
s_t	vector	Sampled-token score $\nabla_\theta \log q_\theta(Y_t \mid H_t)$
h	scalar function	Probability transform used by a sampled-token reward
ψ	scalar function	Convex generator satisfying $\psi' = h$
B_t	scalar	Conditional divergence at occupied prefix H_t
L_B	scalar	Student-occupancy expectation of the sum of B_t
u_B	vector	Negative direct-gradient field used by a local update
c_B	vector	Future-credit contribution omitted by u_B
Φ_B	scalar	Objective fidelity of u_B relative to $-\nabla L_B$
C_B	matrix	Antisymmetric Jacobian of the local field

A.3 Teacher dependence

The teacher may depend on the occupied prefix and on fixed privileged information. It is held fixed with respect to θ during one student update. A moving teacher creates a two-timescale system; its parameter derivative is separate from the student-occupancy derivative analyzed here.

A.4 Parameterization

The prefix theorem holds for every differentiable mapping from parameters to the simplex. The converse in Proposition D.1 uses independent tabular softmax coordinates to make the score span complete. The forward implication and every decomposition remain valid under shared neural parameters.

B Conditional Propriety Proofs

B.1 Proof of Theorem 4.1

Set $w_I = w(p_I, q_{\theta,I})$. The score identity converts the expected update into a differential on the probability simplex:

$$\begin{aligned}\mathbb{E}_{I\sim q_\theta}\left[w_I \nabla_\theta \log q_{\theta,I}\right] &= \sum_i w(p_i, q_{\theta,i}) \nabla_\theta q_{\theta,i} \\ &= -\nabla_\theta D_w(q_\theta, p). \end{aligned} \tag{23}$$

If $q_i > p_i$, the sign condition gives $w(p_i, z) < 0$ between p_i and q_i. If $q_i < p_i$, it gives $w(p_i, z) > 0$ on the reversed interval. Every nonzero coordinate contribution to Equation 7 is therefore positive, which proves strict propriety.

Now suppose $D_w = B_\psi$ for a differentiable separable generator. Comparing coordinate derivatives in the positive orthant gives

$$-w(p, z) = \psi'(z) - \psi'(p). \tag{24}$$

Hence $w(p, z) = h(p) - h(z)$ with $h = \psi'$. Conversely, integrating any reward of this form recovers B_ψ up to affine generator terms. On the simplex, adding a common action-independent reward baseline changes neither field because of Equation 31.

B.2 Proof of Corollary 4.2

Let $q = q_\theta$ and $r_I = h(p_I) - h(q_I)$. The expected sampled-token field equals

$$\begin{aligned}\mathbb{E}_{I\sim q}\left[r_I \nabla_\theta \log q_I\right] &= \sum_i q_i (h(p_i) - h(q_i)) \frac{\nabla_\theta q_i}{q_i} \\ &= \sum_i (h(p_i) - h(q_i)) \nabla_\theta q_i. \end{aligned} \tag{25}$$

Differentiating Equation 9 gives

$$\nabla_\theta B_\psi(q_\theta, p) = \sum_i (h(q_i) - h(p_i)) \nabla_\theta q_i. \tag{26}$$

Equations 25 and 26 prove the identity.

If h is increasing, then ψ is convex and the supporting-hyperplane inequality gives

$$\psi(q_i) \geq \psi(p_i) + h(p_i)(q_i - p_i). \tag{27}$$

Summing Equation 27 gives nonnegativity. Strict increase makes ψ strictly convex, so equality on the simplex interior holds only at $q = p$.

B.3 Examples

Logarithmic reward. For $h(z) = \log z$ we may take $\psi(z) = z \log z - z$. The affine terms cancel on the simplex and

$$B_\psi(q, p) = \sum_i q_i \log \frac{q_i}{p_i} = D_{\mathrm{KL}}(q \| p). \tag{28}$$

Power reward. For $h(z) = z^\alpha$, $\alpha > 0$, take

$$\psi_\alpha(z) = \frac{z^{\alpha+1}}{\alpha + 1}. \tag{29}$$

The resulting divergence is

$$B_{\psi_\alpha}(q, p) = \sum_i \left[\frac{q_i^{\alpha+1} - p_i^{\alpha+1}}{\alpha + 1} - p_i^\alpha (q_i - p_i) \right]. \tag{30}$$

At $\alpha = 1$, this reduces to one half of squared Euclidean distance on the probability simplex.

B.4 Reward baselines and uniqueness

The simplex constraint permits adding the same scalar to every reward because

$$\sum_i q_i \nabla_\theta \log q_i = \nabla_\theta \sum_i q_i = 0. \tag{31}$$

Hence the conditional objective is unique only up to affine generator terms and action-independent baselines.

C Autoregressive Composition

C.1 Reverse KL

Using Equation 2, write $q_t = q(\cdot \mid H_t)$, $p_t = p(\cdot \mid H_t)$, and $r_t = \log q(Y_t \mid H_t) - \log p(Y_t \mid H_t)$. Then

$$\begin{aligned} D_{\mathrm{KL}}(q_{1:T} \| p_{1:T}) &= \mathbb{E}_{Y\sim q} \sum_{t=1}^{T} r_t \\ &= \sum_{t=1}^{T} \mathbb{E}_{H_t \sim q} D_{\mathrm{KL}}(q_t \| p_t). \end{aligned} \tag{32}$$

This is the chain rule used by sequence reverse-KL OPD.

C.2 Proof of Proposition 4.3

Write the coordinate contribution as

$$d_\psi(x, y) = \psi(x) - \psi(y) - \psi'(y)(x - y). \tag{33}$$

Consider a binary root whose first outcome is refined into two outcomes in the next step. Give the student and teacher the same conditional split $(c, 1 - c)$, but root probabilities x and y. The conditional divergence of the split is zero. Universal composition therefore requires

$$d_\psi(cx, cy) + d_\psi((1 - c)x, (1 - c)y) = d_\psi(x, y) \tag{34}$$

for every positive x, y, c for which the resulting three-outcome distributions are interior.

Differentiate Equation 34 twice with respect to x and then set $x = y$. With $f = \psi''$, this gives

$$c^2 f(cy) + (1 - c)^2 f((1 - c)y) = f(y). \tag{35}$$

Set $u = cy$ and $v = (1 - c)y$. Multiplication by $y = u + v$ transforms the identity into

$$u^2 f(u) + v^2 f(v) = (u + v)^2 f(u + v). \tag{36}$$

Thus $g(z) = z^2 f(z)$ is additive on the positive interval. Continuity implies $g(z) = az$, hence $\psi''(z) = a/z$. Two integrations yield Equation 12; strict convexity gives $a > 0$. Direct substitution verifies that this generator satisfies the chain rule.

C.3 Relation to the classical characterization

A chain rule implies invariance under refinements that retain all information relevant to a pair of distributions. Such invariance is the sufficiency condition studied in proper scoring and information geometry. On finite alphabets with at least three outcomes, a differentiable Bregman divergence satisfying sufficiency is proportional to information divergence (Harremoës, 2015). The generator is therefore Equation 12 up to affine terms.

The OPD-specific conclusion is the mapping from a practical reward transform to its conditional Bregman generator, followed by the observation that bounded transforms lie outside the compositional class.

C.4 A two-step diagnostic

For a binary two-step tree, let $q_{ij} = q_i q_{j|i}$ and $p_{ij} = p_i p_{j|i}$. Using Equation 33, define the empirical chain residual

$$\begin{aligned}\Delta_\psi(q, p) = &\sum_{i,j} d_\psi(q_{ij}, p_{ij}) - \sum_i d_\psi(q_i, p_i) \\ &- \sum_i q_i \sum_j d_\psi(q_{j|i}, p_{j|i}).\end{aligned} \tag{37}$$

The logarithmic generator makes Δ_ψ identically zero. Power generators produce nonzero residuals for generic positive conditionals. The finite-tree checks evaluate Equation 37 after enumerating complete trees, rather than inferring failure from optimization behavior.

C.5 Boundedness consequence

Equation 12 has derivative

$$h(z) = a(1 + \log z) + b. \tag{38}$$

For $a > 0$, this derivative is unbounded below as z approaches zero. A transform bounded on the closed probability interval cannot have this form. Conditional strict propriety and boundedness remain compatible by Theorem 4.2; boundedness conflicts only with universal composition as the same separable path-space Bregman divergence.

C.6 Proof of Theorem 4.5

For the first construction, take a one-parameter, three-step binary student. Its root feature is one. At depth one the features following actions zero and one are minus one and one. Every depth-two feature is one. Set $\theta = 2.6344758262$. The teacher probabilities at the root, depth-one prefixes, and depth-two prefixes are

$$\begin{aligned}&(0.8575418901),\\ &(0.5789768196,\ 0.5587632575),\\ &(0.0400315682,\ 0.0314507936),\\ &(0.9938537965,\ 0.9933607070).\end{aligned} \tag{39}$$

All probabilities are strictly between zero and one. Complete enumeration with the logarithmic generator gives

$$\begin{aligned}u_B &= -0.09030669,\\ c_B &= -0.18387534,\\ -\nabla L_B &= 0.09356865.\end{aligned} \tag{40}$$

The chain residual is below 3×10^{-16} by the KL chain rule, while

$$\Phi_B = \frac{u_B}{-\nabla L_B} = -0.96513829. \tag{41}$$

Thus a perfectly compositional conditional divergence can have a token-local field that ascends its sequence objective.

For the second construction, take a two-step binary student with one parameter. The root feature is zero, and the two depth-one features are one and minus one. Set $\theta = 0.4$ and teacher probabilities to 0.3 at the root and $(0.2, 0.8)$ at depth one. The root occupancy is parameter independent, so $c_B = 0$. With the power generator $\psi(z) = z^2/2$,

$$\begin{aligned}u_B &= -\nabla L_B = -0.19157799,\\ \Phi_B &= 1.\end{aligned} \tag{42}$$

The path-space power Bregman divergence is 0.10667593, whereas the occupancy-weighted conditional objective is 0.19895185. Their residual magnitude is

$$|\Delta_\psi| = 0.09227593. \tag{43}$$

This proves the reverse nonimplication.

D Future Credit and Fidelity Proofs

D.1 When does local credit suffice?

Define the divergence cost-to-go after position s as

$$Q_s^B(h, a) = \mathbb{E}_{q_\theta}\left[\sum_{t>s} B_t(\theta, H_t) \,\middle|\, H_s = h, Y_s = a\right]. \tag{44}$$

Let $V_s^B(h) = \mathbb{E}_{A\sim q_\theta(\cdot|h)} Q_s^B(h, A)$ and $A_s^B = Q_s^B - V_s^B$.

Proposition D.1 (Future neutrality)**.** *The future-credit term admits the advantage form*

$$c_B = \mathbb{E}\sum_{s=1}^{T-1} s_s A_s^B. \tag{45}$$

It vanishes whenever $Q_s^B(h, a)$ is independent of a at every occupied prefix. Under independent full-rank tabular softmax parameters, the converse holds almost surely.

The action dependence of future divergence, instead of horizon alone, determines the need for future credit. Cauchy–Schwarz gives the measurable upper bound

$$\|c_B\|_2 \le \sum_{s=1}^{T-1} \sqrt{\mathbb{E}\|s_s\|_2^2}\sqrt{\mathbb{E}(A_s^B)^2}. \tag{46}$$

D.2 Proof of Theorem 4.6

Write the finite trajectory probability as $q_\theta(y)$ and let $G_B(y) = \sum_t B_t(\theta, H_t)$. The log-derivative identity gives

$$\nabla_\theta L_B = \sum_y q_\theta(y) \nabla_\theta G_B(y) + \sum_y q_\theta(y) G_B(y) \nabla_\theta \log q_\theta(y). \tag{47}$$

The complete trajectory score is

$$\nabla_\theta \log q_\theta(y) = \sum_{s=1}^{T} s_s. \tag{48}$$

For $s \geq t$, B_t is measurable before action Y_s is drawn. Conditional score centering yields

$$\mathbb{E}[s_s B_t \mid H_s] = B_t \mathbb{E}[s_s \mid H_s] = 0. \tag{49}$$

Only terms with $s < t$ remain, giving

$$\nabla_\theta L_B = \mathbb{E} \sum_t \nabla_\theta B_t + \mathbb{E} \sum_{s<t} s_s B_t. \tag{50}$$

Substituting Equations 13 and 14 proves Equation 15.

D.3 Proof of Proposition 4.8

The directional derivative along the update field is

$$\begin{aligned} \langle \nabla L_B, u_B \rangle &= -\|\nabla L_B\|_2^2 + \langle \nabla L_B, c_B \rangle \\ &= -\Phi_B \|\nabla L_B\|_2^2. \end{aligned} \tag{51}$$

It is negative exactly when $\Phi_B > 0$. Cauchy–Schwarz gives

$$\left| \frac{\langle \nabla L_B, c_B \rangle}{\|\nabla L_B\|_2^2} \right| \leq \frac{\|c_B\|_2}{\|\nabla L_B\|_2}. \tag{52}$$

Equality holds when c_B is collinear with ∇L_B.

D.4 Proof of Proposition 4.9

The descent lemma and Equation 51 give

$$\begin{aligned} L_B(\theta + \eta u_B) - L_B(\theta) &\leq \eta \langle \nabla L_B, u_B \rangle \\ &\quad + \frac{M\eta^2}{2} \|u_B\|_2^2 \\ &= -\eta \Phi_B \|\nabla L_B\|_2^2 \\ &\quad + \frac{M\eta^2}{2} \|u_B\|_2^2. \end{aligned} \tag{53}$$

The right-hand side is negative precisely throughout the interval in Equation 18.

D.5 Proof of Proposition D.1

Starting from Equation 13, condition on (H_s, Y_s) and collect all later divergences:

$$c_B = \mathbb{E} \sum_s s_s Q_s^B(H_s, Y_s). \tag{54}$$

Subtracting the prefix baseline $V_s^B(H_s)$ does not change the expectation by Equation 31. This proves Equation 45.

If $Q_s^B(h, a)$ is action independent, its centered advantage is zero. For the converse, use independent tabular logits at each prefix. Conditional on an occupied h, the score span is the tangent space of the simplex. The expectation of score times Q_s^B vanishes for every logit coordinate only when the centered vector of action values is zero. Positivity of the occupied-prefix probability then establishes action independence almost surely.

D.6 Proof of Equation 46

Apply the triangle inequality to Equation 45, then apply Cauchy–Schwarz to each position:

$$\begin{aligned}\|c_B\|_2 &\le \sum_s \left\|\mathbb{E}[s_s A_s^B]\right\|_2 \\ &\le \sum_s \sqrt{\mathbb{E}\|s_s\|_2^2}\sqrt{\mathbb{E}(A_s^B)^2}. \end{aligned} \tag{55}$$

The factors distinguish parameter sensitivity from action-dependent future mismatch.

E Integrability and Circulation Proofs

E.1 Proof of Theorem 4.10

Let $u_i = -\mathbb{E}\sum_t \partial_i B_t$. Differentiating the expectation gives

$$\partial_j u_i = -\mathbb{E}\sum_t \left[\partial_{ji} B_t + \sum_{s<t} s_{s,j}\partial_i B_t\right]. \tag{56}$$

Terms with $s \ge t$ vanish by the same conditional-centering argument as Equation 49. Interchanging i and j and subtracting cancels the symmetric Hessian:

$$\partial_j u_i - \partial_i u_j = -\mathbb{E}\sum_{s<t}\left[s_{s,j}\partial_i B_t - s_{s,i}\partial_j B_t\right]. \tag{57}$$

This is Equation 20. The Poincaré lemma gives integrability if and only if the antisymmetric Jacobian vanishes throughout a simply connected domain.

E.2 Local loop expansion

Let a, b be orthonormal and define

$$f(r,s) = u_B(\theta + ra + sb). \tag{58}$$

A first-order Taylor expansion on the four edges of a square gives

$$\begin{aligned}\oint u_B^\top d\theta &= \varepsilon^2\left[b^\top J_B a - a^\top J_B b\right] + O(\varepsilon^3) \\ &= -2\varepsilon^2 a^\top C_B b + O(\varepsilon^3). \end{aligned} \tag{59}$$

Reversing the orientation changes the sign. Applying the same calculation to an exact gradient field yields a zero quadratic term because its Jacobian is symmetric.

E.3 Approximate integrability

The same tensor also controls approximate, rather than exact, integrability. For a star-shaped region Ω centered at θ_0, define the radial potential

$$F_{\theta_0}(\theta) = -\int_0^1 u_B(\theta_0 + t(\theta-\theta_0))^\top(\theta-\theta_0)\,dt. \tag{60}$$

Proposition E.1 (Approximate potential)**.** *If $\|C_B(\vartheta)\|_{\mathrm{op}} \le \varepsilon$ throughout Ω, then*

$$\|u_B(\theta) + \nabla F_{\theta_0}(\theta)\|_2 \le \varepsilon\|\theta - \theta_0\|_2. \tag{61}$$

E.4 Proof of Proposition E.1

Write $x = \theta - \theta_0$ and differentiate Equation 60 under the integral. The identity

$$\begin{aligned}\frac{d}{dt}\left[t u_{B,j}(\theta_0 + tx)\right] &= u_{B,j}(\theta_0 + tx) \\ &\quad + t x_i \partial_i u_{B,j}(\theta_0 + tx). \end{aligned} \tag{62}$$

yields

$$-\nabla F_{\theta_0}(\theta) - u_B(\theta) = \int_0^1 2t\, C_B(\theta_0 + tx)^\top x\, dt. \tag{63}$$

Taking norms, applying the uniform operator-norm bound, and using $\int_0^1 2t\, dt = 1$ proves Equation 61.

E.5 Projected neural estimator

For directions $a_1, \ldots, a_m$, define directional scores and divergence derivatives

$$\sigma_{t,r} = a_r^\top s_t, \qquad \delta_{t,r} = a_r^\top \nabla B_t. \tag{64}$$

The projected antisymmetric matrix has entries

$$2[\widetilde{C}_B]_{rs} = -\mathbb{E} \sum_{v<t} \left[\sigma_{v,s}\delta_{t,r} - \sigma_{v,r}\delta_{t,s}\right]. \tag{65}$$

Central finite differences evaluate both quantities from common trajectories. The experiment reports the projected Frobenius norm and does not interpret it as the full-parameter norm.

E.6 Finite-difference convergence

For a three-times differentiable scalar function, the centered directional derivative has truncation error $O(\varepsilon^2)$. Neural floating-point arithmetic adds a small-ε roundoff term. The pilot therefore evaluates three relative step sizes in FP32 and checks stability of the reported future-credit, fidelity, and circulation statistics before freezing the middle value.

F Support Restriction Identities

F.1 Proof of Equation 22

For $i \in S$, write $q_i = \alpha q_S(i)$ and $p_i = \beta p_S(i)$. Then

$$\sum_{i\in S} q_i \log \frac{q_i}{p_i} = \alpha \log \frac{\alpha}{\beta} + \alpha D_{\mathrm{KL}}(q_S \| p_S). \tag{66}$$

Applying the same calculation to S^c gives

$$\begin{aligned}\sum_{i\notin S} q_i \log \frac{q_i}{p_i} &= (1-\alpha)\log\frac{1-\alpha}{1-\beta} \\ &\quad + (1-\alpha) D_{\mathrm{KL}}(q_{S^c} \| p_{S^c}).\end{aligned} \tag{67}$$

Adding Equations 66 and 67 proves the result.

F.2 Exact omitted quantity

Let the renormalized local-support objective be

$$L_S(q,p) = D_{\mathrm{KL}}(q_S \| p_S). \tag{68}$$

Let $D_{S^c} = D_{\mathrm{KL}}(q_{S^c} \| p_{S^c})$. Its difference from full reverse KL is

$$\begin{aligned}D_{\mathrm{KL}}(q\|p) - L_S(q,p) &= \mathrm{kl}(\alpha\|\beta) \\ &\quad - (1-\alpha) L_S(q,p) \\ &\quad + (1-\alpha) D_{S^c}.\end{aligned} \tag{69}$$

The right side may have either sign because the practical objective also removes the factor α multiplying the inside conditional divergence. The appropriate statement is an identity, not a universal ordering.

F.3 Support changes

When S is selected from teacher logits, it is fixed with respect to student parameters during an update. The conditional objective is smooth as long as the teacher ranking has no tie at the boundary. Student-selected support introduces piecewise-smooth regions whose boundary terms require separate treatment. The primary experiment uses teacher-selected support.

G Further Consequences of the Decomposition

The identities in the main text have several consequences that are useful when designing or interpreting an OPD surrogate.

G.1 Fixed occupancy is an exact control

Suppose trajectories are drawn from a distribution μ independent of θ, while the conditional divergence still uses q_θ. Define

$$L_B^\mu(\theta) = \mathbb{E}_{Y\sim\mu} \sum_t B_t(\theta, H_t). \tag{70}$$

There is no likelihood-ratio derivative of μ, and therefore

$$u_B^\mu(\theta) = -\nabla L_B^\mu(\theta). \tag{71}$$

This explains why teacher-forced or replay-buffer objectives can be well-defined even when the analogous on-policy token-local field is not. The difference is occupancy differentiation, not conditional propriety.

G.2 A recursive view of future credit

Let $G_{s+1}^B = \sum_{t>s} B_t$ be the future divergence return. The omitted term can be written

$$c_B = \sum_{s=1}^{T-1} \mathbb{E}[s_s G_{s+1}^B]. \tag{72}$$

Conditioning on (H_s, Y_s) replaces G_{s+1}^B by Q_s^B without changing the expectation. Conditioning once more on H_s and subtracting V_s^B gives Proposition D.1. Thus three estimators target the same population term: a raw sampled return, a conditional cost-to-go, and a centered conditional advantage.

The conditional advantage is the variance-minimizing action-independent baseline. For any prefix baseline $b(H_s)$,

$$\mathbb{E}[s_s b(H_s)] = 0. \tag{73}$$

Minimizing the conditional second moment of $s_s(Q_s^B - b(H_s))$ yields a score-weighted value baseline. When score norms are action independent, it reduces to V_s^B. This variance statement does not alter the target future-credit vector.

G.3 Horizon is a carrier, not a sufficient cause

Equation 46 grows as more future positions become available, but it can remain zero at an arbitrary horizon when later divergence is action independent. Conversely, one highly action-sensitive transition can create large future credit at a short horizon. Define

$$\chi_s^B = \sqrt{\mathbb{E}\|s_s\|_2^2}\sqrt{\mathbb{E}(A_s^B)^2}. \tag{74}$$

Then $\|c_B\|_2 \le \sum_s \chi_s^B$. The first factor measures how strongly the student can move probability mass at position s; the second measures how much that action changes future teacher mismatch. Their product, rather than token count alone, is the local carrier of sequence inconsistency.

G.4 Composition does not remove occupancy derivatives

When B is reverse KL, L_B equals a complete path-space divergence. The chain rule changes the interpretation of the scalar but not its derivative. The derivative still contains both the direct conditional term and the likelihood-ratio term from earlier actions. Therefore

$$\text{composition} \not\Rightarrow \text{local fidelity}. \tag{75}$$

The first construction in Theorem 4.5 makes this nonimplication strict with $\Phi_B < 0$.

G.5 Conditional propriety does not determine the path objective

For a bounded increasing transform, Theorem 4.2 provides a strictly proper divergence at each occupied prefix. The scalar L_B in Equation 6 is therefore meaningful and is minimized when every occupied conditional matches the teacher. However, it is generally not the same Bregman divergence applied to the two distributions over complete responses. This distinction permits a bounded method to have high local fidelity to its own occupancy objective without inheriting a sequence chain rule.

G.6 Finite-step versus infinitesimal progress

Objective fidelity is an instantaneous coefficient. Proposition 4.9 converts it to a finite-step guarantee under local smoothness. The admissible step interval contracts when Φ_B approaches zero, when the update norm grows relative to the objective gradient, or when curvature increases. This explains why two methods with the same sign of Φ_B can behave differently under one shared learning rate.

For a stochastic estimate $\widehat{u}_B = u_B + \xi$ with $\mathbb{E}[\xi \mid \theta] = 0$, the expected first-order progress remains

$$\mathbb{E}[\langle \nabla L_B, \widehat{u}_B \rangle \mid \theta] = -\Phi_B \|\nabla L_B\|_2^2. \tag{76}$$

The second-order penalty depends on $\mathbb{E}\|\widehat{u}_B\|_2^2$, so future correction can improve objective fidelity while worsening the usable step size through variance. The training panel measures both sides of this tradeoff.

G.7 Approximate potential is local and testable

Proposition E.1 gives a constructive scalar F_{θ_0} on any star-shaped neighborhood. Small operator norm of the antisymmetric Jacobian makes the field uniformly close to its negative gradient, with error increasing at most linearly in distance from the center. Closed-loop work supplies the complementary falsification test. For a square of side r in directions a, b,

$$\frac{1}{r^2} \oint_{\partial \mathcal{R}_r} u_B^\top d\theta \longrightarrow -2a^\top C_B b \quad \text{as } r \to 0. \tag{77}$$

The exact experiment checks the limit, its sign reversal under opposite orientation, and the zero result for the corrected gradient field.

H Neural Estimands and Estimators

This appendix states exactly how the population objects in Section 4 become finite neural measurements. The construction uses the same trajectory, teacher logits, and parameter directions for every surrogate being compared.

H.1 Projected directional derivatives

Let $a_1, \ldots, a_m$ be orthonormal directions in the trainable LoRA subspace. For a frozen student trajectory y and occupied prefix H_t, the directional score and divergence derivative are

$$\widehat{s}_{t,r} = \frac{\log q_{\theta+\epsilon a_r}(y_t \mid H_t) - \log q_{\theta-\epsilon a_r}(y_t \mid H_t)}{2\epsilon}, \tag{78}$$

$$\widehat{d}^{B}_{t,r} = \frac{B_t(\theta + \epsilon a_r, H_t) - B_t(\theta - \epsilon a_r, H_t)}{2\epsilon}. \tag{79}$$

The perturbations change the evaluated conditional distribution but not the stored trajectory. This is deliberate: the trajectory score in Equation 13 accounts analytically for the derivative of student occupancy. Regenerating a different discrete trajectory at each perturbation would instead mix the target derivative with sampling noise.

For one trajectory, write

$$\widehat{S}_{t,r} = \sum_{s<t} \widehat{s}_{s,r}. \tag{80}$$

The projected direct, future-credit, and exact-gradient vectors are then

$$[\widehat{u}_B]_r = -\sum_t \widehat{d}^B_{t,r}, \tag{81}$$

$$[\widehat{c}_B]_r = \sum_t \widehat{S}_{t,r} B_t(\theta, H_t), \tag{82}$$

$$[-\widehat{\nabla L_B}]_r = [\widehat{u}_B]_r - [\widehat{c}_B]_r. \tag{83}$$

Equation 83 evaluates the decomposition from two separately measurable terms. It does not estimate an unobserved full-parameter gradient and then project it after the fact.

H.2 Projected fidelity and wedge

For each prompt trajectory, the reported fidelity is

$$\widehat{\Phi}_B = \frac{\langle \widehat{u}_B - \widehat{c}_B, \widehat{u}_B \rangle}{\|\widehat{u}_B - \widehat{c}_B\|_2^2}. \tag{84}$$

The corresponding projected antisymmetric matrix is

$$[\widehat{W}_B]_{rs} = -\sum_t \left(\widehat{d}^B_{t,r} \widehat{S}_{t,s} - \widehat{S}_{t,r} \widehat{d}^B_{t,s} \right). \tag{85}$$

By Theorem 4.10, $\widehat{W}_B$ estimates $2C_B$ restricted to the chosen subspace. We report its Frobenius norm and normalize by $\|\widehat{u}_B\|_2$ when comparing horizons. This normalization removes the most direct scale change between logarithmic and bounded transforms.

The projection cannot certify that the full-space wedge is zero. A nonzero projected value does certify a nonzero full-space wedge, while a small value only describes the measured plane. Direction-robustness runs increase m from four to eight and regenerate the orthonormal basis from an independent seed schedule.

H.3 Horizon nesting

All requested horizons for a trajectory are nested prefixes of one response. If a response terminates before the requested horizon, the realized horizon is its available token count. For a requested H, Equations 81 through 85 use only positions $t \le H$. Consequently, differences between adjacent horizons are paired within trajectory, seed, task, model pair, transform, and parameter basis.

This nesting makes the one-token cell an exact internal control. At $H = 1$, $\widehat{S}_{1,r} = 0$ for every direction, so

$$\widehat{c}_B = 0, \qquad \widehat{W}_B = 0, \qquad \widehat{\Phi}_B = 1 \tag{86}$$

whenever the projected direct gradient is nonzero. Every neural run checks this identity before contributing to an aggregate.

H.4 Common random numbers

The central differences share all non-parameter randomness. The plus and minus evaluations use identical token sequences, masks, teacher logits, and direction tensors. Surrogate comparisons reuse the same baseline student logits. Horizon comparisons reuse the same longer response. These couplings reduce estimator variance without increasing the independent sample size.

The statistical unit is therefore the training or measurement seed. Prompt trajectories and directions within a seed are averaged before uncertainty is computed. Percentile intervals resample seed-level records with replacement. No prompt, token, response, or parameter direction is counted as an independent replicate.

H.5 Trained-endpoint measurement

For a training method M, let $\theta_M^{(s)}$ denote the final adapter from seed s. The endpoint panel restores that adapter and repeats the frozen measurement using the same seed-specific prompt trajectories and directions as at initialization. The paired change is

$$\Delta\widehat{\Phi}_{B,M}^{(s)} = \widehat{\Phi}_B(\theta_M^{(s)}) - \widehat{\Phi}_B(\theta_0^{(s)}). \tag{87}$$

This quantity describes the local field that would be applied at the trained state. It is not relabeled as the fidelity of the method that produced the state. In particular, a corrected method can reach a state where switching back to local credit would have low fidelity.

H.6 Fresh-sequence evaluation

Cached prefix metrics are complemented by new trajectories sampled from each final adapter. On every fresh trajectory, the evaluator computes the sampled sequence log ratio, the Rao–Blackwellized sum of conditional reverse KL, the occupancy-weighted power objective, and the teacher-selected support objective. Four responses are drawn for each of sixteen heldout prompts per training seed. Common prompt and generation-seed schedules make method contrasts paired, while uncertainty again uses the training seed as the independent unit.

I Implementation and Evaluation Details

I.1 Models and trainable subspace

The primary neural panel uses a Qwen3.5 0.8B student and a Qwen3.5 4B teacher. The scale panel uses a 4B student and a 9B teacher. Teachers are frozen. Each student receives rank-eight LoRA adapters with scaling factor sixteen on all linear modules. Every seed begins from a freshly restored adapter tensor, and the serialized tensor hash is checked after restoration.

Frozen-checkpoint measurements evaluate the primary student in FP32 with eager attention. The scale measurement also uses FP32 after a memory pilot confirmed that the 4B and 9B pair fits on one 40 GB accelerator. Training uses BF16 and scaled dot-product attention. TensorFloat-32 is disabled for central finite differences and enabled for training.

I.2 Tasks and prompts

GSM8K supplies open-form numerical reasoning questions. ARC-Challenge supplies multiple-choice science questions, and HellaSwag supplies multiple-choice commonsense continuations. Each prompt requests concise reasoning and a final answer introduced by the string ####. Numerical answers remove currency symbols and thousands separators before comparison. Multiple-choice answers retain the final option letter. The training and heldout pools are sorted by hashes of the task tag and question, then sliced before any model is evaluated.

Training responses contain at most 32 new tokens. Frozen measurements and task evaluations contain at most 64. Maximum prompt length is 384 tokens. Sampling uses temperature 0.8 and nucleus probability 0.95. The generation seed is a deterministic hash of the experimental seed, stage name, and item index.

I.3 Training estimators

Let $r_t = \log q_\theta(Y_t \mid H_t) - \log p(Y_t \mid H_t)$ and let $B_t^{\log}$ and B_t^{pow} denote the conditional reverse-KL and power divergences. The local logarithmic loss whose negative gradient gives the token-local

update is

$$\widehat{\ell}_{\text{local-log}} = \frac{1}{T}\sum_t \text{sg}(r_t)\log q_\theta(Y_t \mid H_t). \tag{88}$$

The Rao–Blackwellized correction adds the conditional future divergence:

$$\widehat{\ell}_{\text{corrected-log}} = \frac{1}{T}\sum_t \text{sg}\left(r_t + \sum_{v>t} B_v^{\log}\right) \cdot \log q_\theta(Y_t \mid H_t). \tag{89}$$

The sequence Monte Carlo control instead uses the sampled return

$$\widehat{\ell}_{\text{sequence-log}} = \frac{1}{T}\sum_t \text{sg}\left(\sum_{v\geq t} r_v\right) \cdot \log q_\theta(Y_t \mid H_t). \tag{90}$$

For the power transform $h(z) = z$, the sampled local reward is $p(Y_t \mid H_t) - q_\theta(Y_t \mid H_t)$. Its corrected version adds the future power divergence with the sign prescribed by Equation 15. Let $G_{t+1}^{\text{pow}} = \sum_{v>t} B_v^{\text{pow}}$. In the implementation,

$$\widehat{\ell}_{\text{local-pow}} = -\frac{1}{T}\sum_t \text{sg}(p_{t,Y_t} - q_{t,Y_t})\log q_{t,Y_t}, \tag{91}$$

and define

$$\widehat{\ell}_{\text{credit-pow}} = \frac{1}{T}\sum_t \text{sg}(G_{t+1}^{\text{pow}})\log q_{t,Y_t}. \tag{92}$$

The corrected loss is

$$\widehat{\ell}_{\text{corrected-pow}} = \widehat{\ell}_{\text{local-pow}} + \widehat{\ell}_{\text{credit-pow}}. \tag{93}$$

The support method computes reverse KL after renormalizing teacher and student probabilities on the teacher's top-sixteen set S_t:

$$\widehat{\ell}_{\text{top-}K} = \frac{1}{T}\sum_t D_{\text{KL}}(q_{t,S_t} \| p_{t,S_t}). \tag{94}$$

All methods use plain SGD with learning rate 0.003 for 128 on-policy updates. Global gradient norm is clipped at one. Plain SGD removes optimizer-state memory from the comparison. Raw pre-clipping norm, clipping incidence, response length, conditional divergences, future conditional divergence, and support mass are recorded at every update.

I.4 Cached and fresh evaluations

At powers of two and at the terminal update, each run evaluates restored common sequences with cached teacher top-sixteen distributions. Cross entropy minus the entropy of the cached teacher distribution defines the conditional teacher gap. This metric fixes the trajectory while changing the student.

Task evaluation is separate. Each final adapter generates four responses for each of sixteen heldout questions. Mean exact accuracy treats each generation as one outcome. Pass@4 treats a question as solved when at least one of its four responses is correct. An independently restored initial adapter receives the same prompt and generation-seed schedule, so every method-to-initial contrast is paired by training seed.

Fresh-sequence evaluation again generates four responses for sixteen heldout questions, but retains teacher and student logits at every generated token. It reports sequence reverse KL, sampled sequence log ratio, occupancy power objective, support-restricted objective, and response length. These quantities distinguish an improvement on cached prefixes from an improvement under the final student's own occupancy.

I.5 Numerical precision and failure handling

The primary finite-difference step is 3×10^{-4} times the larger of the adapter norm and one. Three-step pilot sweeps verify stability before the formal run. The scale pair is measured in FP32 at the same relative step after a BF16 two-step sweep established sign agreement but showed avoidable quantization variation in magnitude.

Each result is written to a temporary file, flushed and synchronized, then atomically renamed. A result contributes only when its schema and terminal status match the aggregator contract. Missing cells prevent aggregation. Training adapters and JSON summaries are keyed by task, seed, and method; central-difference measurements are keyed additionally by transform and horizon. This makes interrupted runs resumable without treating partial files as observations.

J Complete Experimental Protocol

J.1 Exact categorical trials

The categorical checks use seed 271828 and 1,000 trials. Alphabet size is sampled uniformly from two through eight. Student logits are standard Gaussian and the teacher is drawn from a symmetric Dirichlet distribution. The transforms are logarithmic and power transforms with exponents 0.25, 1, and 2. An additional reward $(p - z)(1 + pz)$ lies outside the transform-difference Bregman subclass and has an analytic coordinate-local potential. Every trial records the expected update, potential gradient, divergence nonnegativity, and the value at the teacher.

J.2 Exact autoregressive trees

The tree checks cover 128 independent shared-parameter binary trees. Horizon is sampled from two through six and parameter dimension from two through six. Each prefix receives a normalized Gaussian feature vector. The student probability is logistic in the shared parameter vector; teacher logits are independent Gaussian draws.

Every complete trajectory is enumerated. The numerical checks evaluate conditional and path-space divergences, the future-credit decomposition, central finite-difference gradients and Jacobians, the analytic wedge, and loop integrals in two orthonormal random directions. The objective-gradient finite-difference step is 2×10^{-6}, the Jacobian step is 2×10^{-5}, and the loop side is 10^{-3}.

J.3 Transformer models

The primary neural student is Qwen3.5 0.8B and the teacher is Qwen3.5 4B. A scale panel pairs a 4B student with a 9B teacher. Students use rank-eight LoRA with scale 16 on all linear modules. Frozen measurement uses FP32 student weights, BF16 teacher weights, eager attention, and disabled TF32. Training uses BF16, scaled dot-product attention, and TF32 matrix multiplication. Model revisions and adapter hashes are stored in the result records.

J.4 Frozen prompt and trajectory construction

GSM8K, ARC-Challenge, and HellaSwag use deterministic train and test splits. The primary GSM8K and ARC assets retain their common warm-start adapters. HellaSwag and the scale panel begin from fresh common LoRA adapters. Each seed fixes its design and evaluation prompts before a surrogate is compared. Student trajectories, teacher distributions, prompt hashes, and adapter tensors are frozen before objective measurement.

J.5 Projected derivatives

Each seed uses four deterministic Rademacher directions normalized in the complete trainable LoRA parameter space. The relative finite-difference step is 3×10^{-4} times the adapter norm. A three-step pilot at 10^{-4}, 3×10^{-4}, and 10^{-3} checks convergence. Every central difference restores the exact baseline adapter before evaluating the opposite side.

Table 3: Maximum recorded residuals for seven numerical consistency checks.

Numerical check	Recorded maximum	Relation
General local potential	7.11×10^{-11}	Eq. 8
Prefix Bregman identity	0	Eq. 11
Log chain rule	2.00×10^{-15}	Eq. 32
Corrected gradient	8.71×10^{-10}	Eq. 15
Analytic wedge	3.53×10^{-11}	Eq. 20
Corrected-field antisymmetry	4.79×10^{-11}	Eq. 19
Corrected-field loop	1.50×10^{-14}	Eq. 21

For each trajectory and transform, the program retains the conditional objective, projected local field, projected future-credit vector, projected exact field, objective fidelity, and projected antisymmetric matrix at horizons 1, 2, 4, 8, 16, 32, 48, and 64. The direction-robustness panels use eight directions at horizons 4, 8, and 16. Support panels compare teacher top-4 and top-64 sets with the top-16 primary method. The independent statistical unit is a frozen training seed; prompts and directions remain paired within seed.

J.6 Training methods

The formal matrix compares six methods. Local log uses the instantaneous sampled log ratio. Corrected log adds the Rao–Blackwellized future conditional KL. Sequence log Monte Carlo uses the sampled return-to-go with discount one. Local power uses $h(z) = z$, whose conditional Bregman divergence is one half squared probability distance. Corrected power adds its future divergence. Top-K log minimizes renormalized reverse KL on the teacher's top 16 tokens.

All methods use 128 regenerated on-policy updates, SGD learning rate 0.003, gradient clipping at one, the same question schedule, and the same generation seeds within training seed. The primary GSM8K and ARC-Challenge panels use eight seeds each. Checkpoints are evaluated after updates 1, 2, 4, 8, 16, 32, 64, and 128.

J.7 Uncertainty

Means and percentile intervals use 10,000 bootstrap resamples of the independent seed. Method comparisons are paired within seed and task. Every registered method, seed, and horizon enters the aggregate.

K Complete Result Tables

K.1 Numerical consistency checks

Table 3 reports the recorded maxima from 1,000 categorical trials and 128 fully enumerated trees. These checks assess numerical consistency with the analytic identities, separately from the constructed counterexamples in Table 1. Appendix J specifies the protocol and finite-difference settings.

The chain-rule and loop entries are absolute residuals. A reported zero denotes zero in the numerical output, rather than an additional exact-arithmetic proof. These residuals assess the implemented identities; the analytic proofs establish the theoretical claims.

K.2 Diagnostic Figures

Method abbreviations throughout the figures and result tables are LL: local log; CL: corrected log; MC: sequence log Monte Carlo; LP: local power; CP: corrected power; TK: top-16 local-support KL.

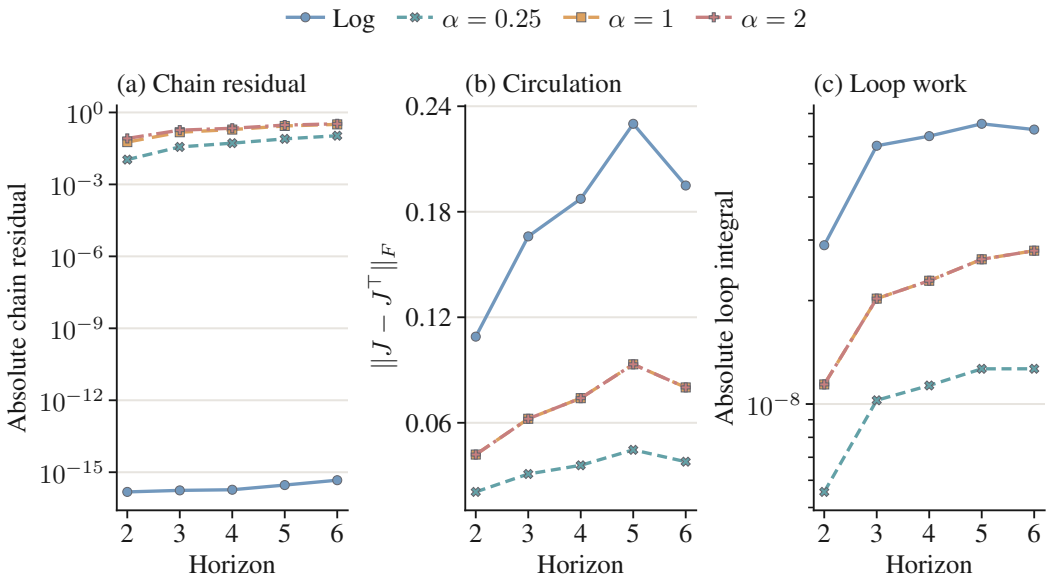

Figure 4: Exact finite-tree objective geometry. Logarithmic conditional KL composes over complete paths, while power Bregman objectives have nonzero chain residuals. Shared-parameter occupancy creates an antisymmetric local field and nonzero closed-loop work.

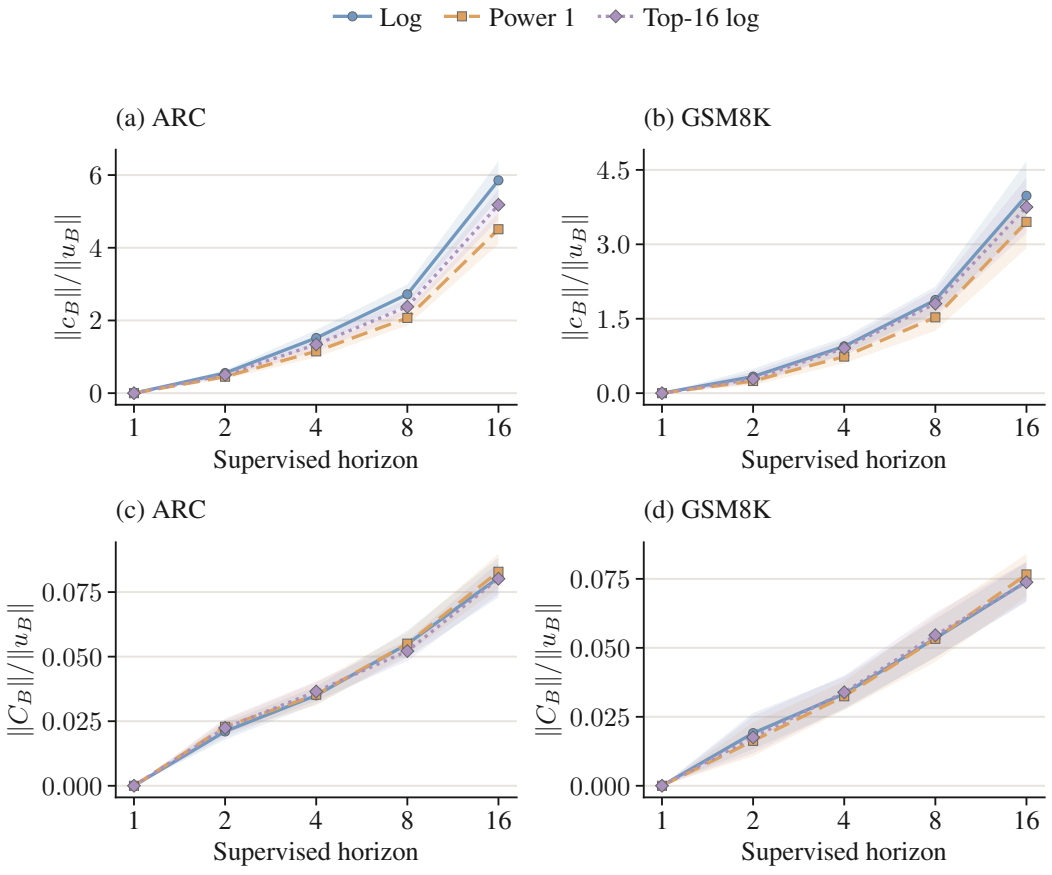

Figure 5: Future credit and the projected antisymmetric Jacobian in the primary model pair. Both grow relative to the token-local update as supervised horizon increases.

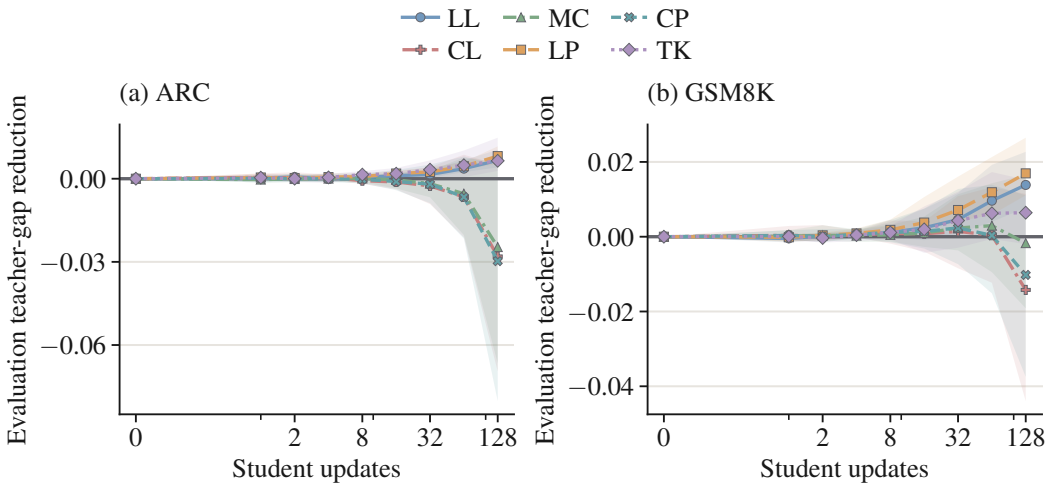

Figure 6: Cached-prefix evaluation gap during formal training. Local and future-corrected methods enter different trajectory regimes. Curves show seed means and 95% bootstrap intervals.

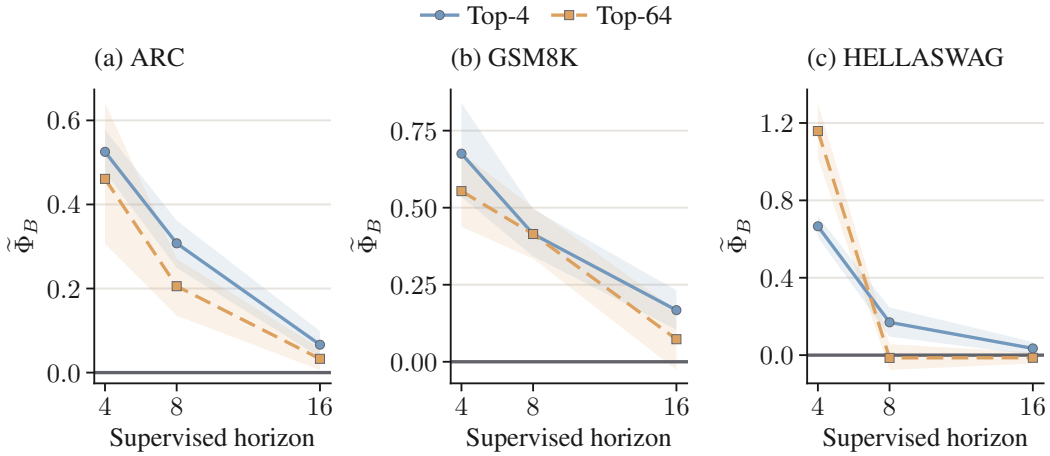

Figure 7: Objective fidelity under teacher-selected top-4 and top-64 support. Greater support mass does not by itself order sequence fidelity because the renormalized conditional objective and future-credit term change together.

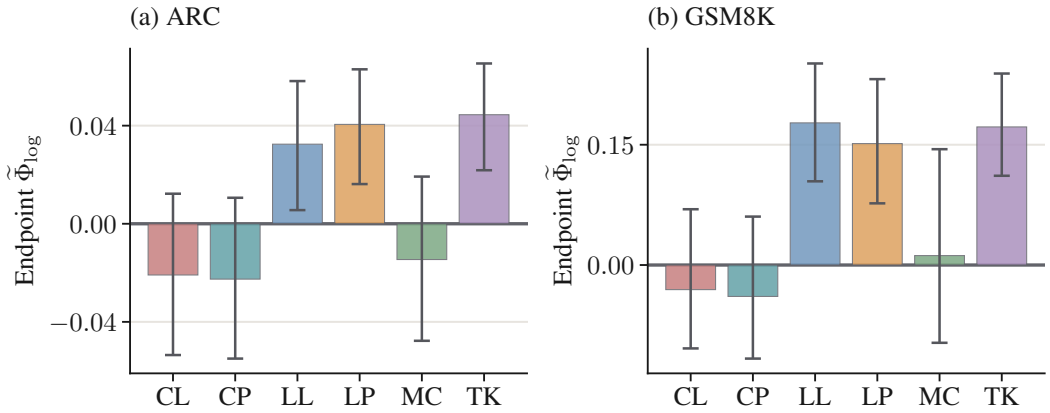

Figure 8: Local-log fidelity at the final adapter of each training method. The bars evaluate one common local field at different trained states; they do not relabel the update used to reach each state.

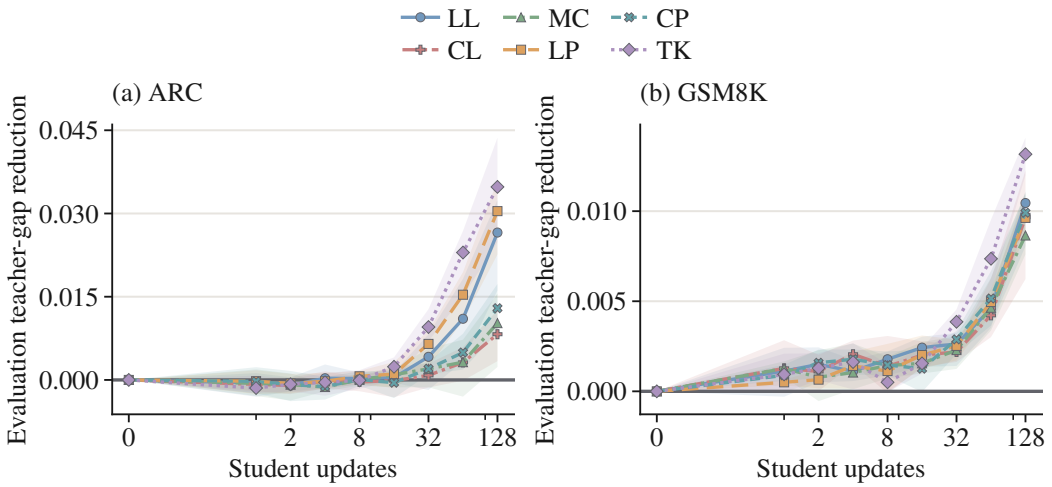

Figure 9: Cached-prefix teacher-gap dynamics for the 4B student and 9B teacher. All six methods improve this metric, while local and support-restricted objectives retain the largest reductions.

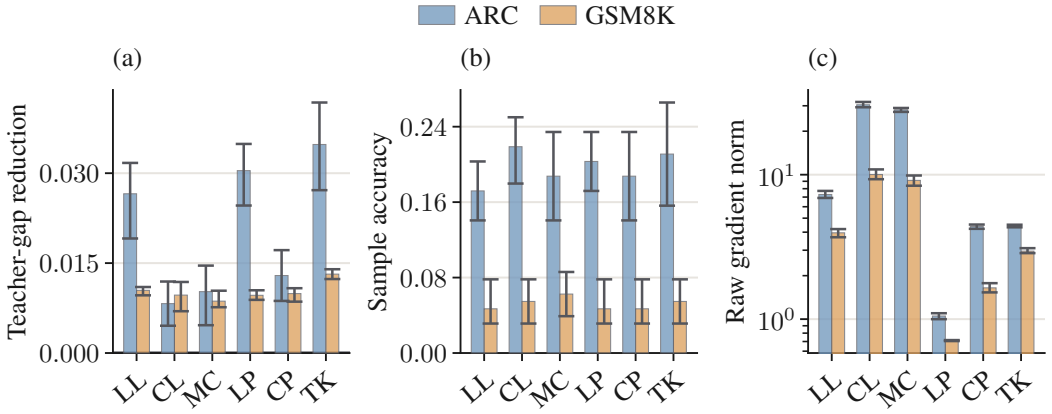

Figure 10: Training and task frontier for the 4B-to-9B extension. The task panel contains four training seeds per task and eight heldout questions per seed.

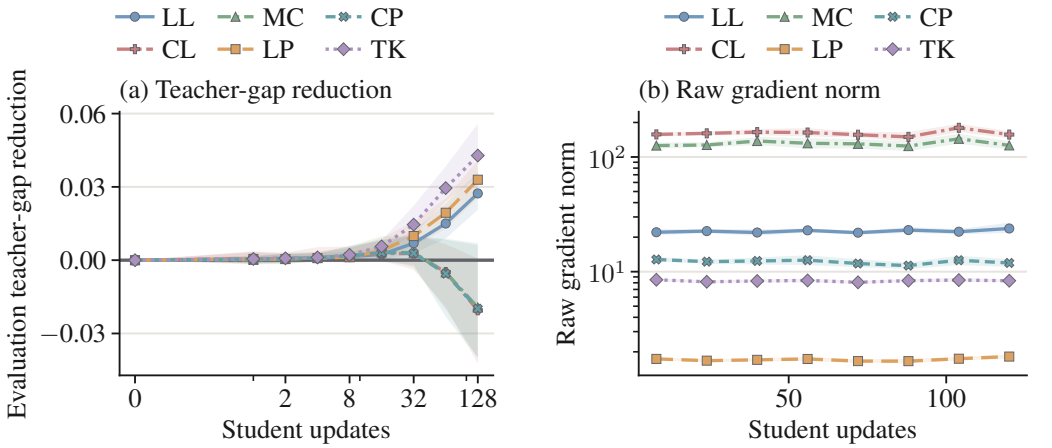

Figure 11: HellaSwag optimization dynamics across sixteen seeds. Left: cached-prefix teacher-gap reduction. Right: block-averaged raw gradient norm with 95% seed-bootstrap intervals. Future correction changes both the optimization trajectory and its gradient scale.

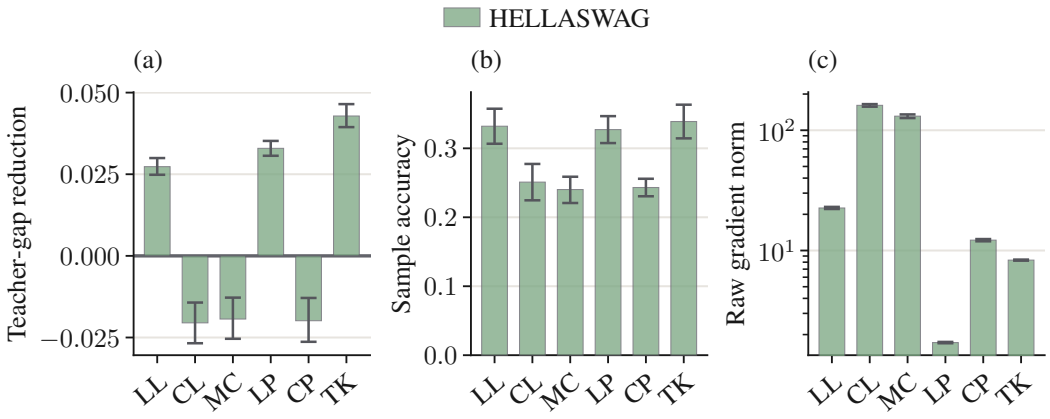

Figure 12: HellaSwag optimization and task frontier across sixteen seeds. Exact future correction reduces task accuracy in this setting while increasing raw gradient scale, exposing the finite-step variance boundary.

Table 4: Exact-tree means by transform and realized horizon.

Transform	T	Chain residual	$\|c_B\|$	$\|J - J^\top\|_F$	Loop
Log	2	1.51×10^{-16}	0.064	0.109	2.90×10^{-8}
Log	3	1.75×10^{-16}	0.106	0.166	5.64×10^{-8}
Log	4	1.86×10^{-16}	0.114	0.187	6.02×10^{-8}
Log	5	2.91×10^{-16}	0.132	0.230	6.53×10^{-8}
Log	6	4.65×10^{-16}	0.125	0.195	6.29×10^{-8}
Power 0.25	2	0.011	0.011	0.021	5.55×10^{-9}
Power 0.25	3	0.037	0.017	0.031	1.03×10^{-8}
Power 0.25	4	0.052	0.020	0.036	1.13×10^{-8}
Power 0.25	5	0.079	0.023	0.045	1.27×10^{-8}
Power 0.25	6	0.107	0.021	0.038	1.27×10^{-8}
Power 1	2	0.058	0.020	0.042	1.14×10^{-8}
Power 1	3	0.150	0.030	0.062	2.03×10^{-8}
Power 1	4	0.193	0.036	0.074	2.28×10^{-8}
Power 1	5	0.273	0.041	0.093	2.64×10^{-8}
Power 1	6	0.318	0.037	0.080	2.79×10^{-8}
Power 2	2	0.083	0.020	0.042	1.14×10^{-8}
Power 2	3	0.182	0.030	0.062	2.03×10^{-8}
Power 2	4	0.219	0.036	0.074	2.28×10^{-8}
Power 2	5	0.297	0.041	0.093	2.64×10^{-8}
Power 2	6	0.336	0.037	0.080	2.79×10^{-8}

Table 5: Projected objective measurements on ARC across 16 seeds.

Transform	H	$\widetilde{\Phi}_B$	Alignment	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Log	2	0.840 [0.794, 0.897]	0.801 [0.755, 0.844]	0.554 [0.481, 0.631]	0.021 [0.018, 0.024]
Log	4	0.448 [0.339, 0.583]	0.418 [0.329, 0.502]	1.519 [1.305, 1.740]	0.035 [0.032, 0.038]
Log	8	0.158 [0.076, 0.221]	0.291 [0.216, 0.366]	2.719 [2.447, 2.991]	0.055 [0.050, 0.060]
Log	16	0.025 [-0.007, 0.053]	0.100 [0.011, 0.182]	5.855 [5.355, 6.379]	0.081 [0.073, 0.088]
Power 0.25	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 0.25	2	0.825 [0.765, 0.884]	0.847 [0.795, 0.891]	0.513 [0.441, 0.589]	0.022 [0.019, 0.026]
Power 0.25	4	0.522 [0.438, 0.606]	0.562 [0.497, 0.624]	1.373 [1.176, 1.584]	0.036 [0.032, 0.041]
Power 0.25	8	0.275 [0.193, 0.383]	0.373 [0.307, 0.436]	2.509 [2.285, 2.745]	0.056 [0.051, 0.060]
Power 0.25	16	0.048 [0.022, 0.074]	0.196 [0.109, 0.286]	5.683 [5.244, 6.143]	0.083 [0.077, 0.091]
Power 0.5	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]

Continued on next page

Table 5 (continued)

Transform	H	$\widetilde{\Phi}_B$	Alignment	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Power 0.5	2	0.814 [0.759, 0.867]	0.871 [0.829, 0.910]	0.468 [0.400, 0.537]	0.023 [0.020, 0.026]
Power 0.5	4	0.535 [0.469, 0.598]	0.626 [0.564, 0.689]	1.225 [1.055, 1.405]	0.036 [0.031, 0.041]
Power 0.5	8	0.252 [0.197, 0.309]	0.414 [0.339, 0.483]	2.251 [2.041, 2.466]	0.055 [0.051, 0.060]
Power 0.5	16	0.067 [0.042, 0.092]	0.248 [0.171, 0.331]	5.096 [4.657, 5.547]	0.084 [0.077, 0.091]
Power 1	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 1	2	0.858 [0.795, 0.924]	0.892 [0.849, 0.929]	0.453 [0.384, 0.530]	0.023 [0.020, 0.025]
Power 1	4	0.553 [0.484, 0.621]	0.664 [0.596, 0.730]	1.150 [0.985, 1.324]	0.035 [0.031, 0.040]
Power 1	8	0.274 [0.217, 0.330]	0.430 [0.349, 0.503]	2.071 [1.861, 2.300]	0.055 [0.051, 0.059]
Power 1	16	0.079 [0.050, 0.109]	0.278 [0.197, 0.364]	4.509 [4.098, 4.960]	0.083 [0.076, 0.090]
Power 2	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 2	2	0.809 [0.752, 0.860]	0.860 [0.805, 0.905]	0.501 [0.413, 0.607]	0.024 [0.021, 0.027]
Power 2	4	0.548 [0.480, 0.616]	0.648 [0.573, 0.722]	1.243 [1.047, 1.448]	0.036 [0.032, 0.040]
Power 2	8	0.277 [0.227, 0.328]	0.408 [0.329, 0.484]	2.120 [1.881, 2.377]	0.056 [0.052, 0.061]
Power 2	16	0.074 [0.041, 0.108]	0.257 [0.173, 0.348]	4.252 [3.853, 4.676]	0.082 [0.076, 0.089]
Top-16 log	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Top-16 log	2	0.818 [0.778, 0.860]	0.828 [0.781, 0.872]	0.505 [0.431, 0.578]	0.022 [0.019, 0.026]
Top-16 log	4	0.477 [0.388, 0.570]	0.510 [0.435, 0.582]	1.340 [1.144, 1.556]	0.036 [0.033, 0.040]
Top-16 log	8	0.198 [0.151, 0.244]	0.327 [0.270, 0.388]	2.378 [2.168, 2.578]	0.052 [0.048, 0.056]
Top-16 log	16	0.049 [0.022, 0.076]	0.166 [0.070, 0.257]	5.182 [4.777, 5.601]	0.080 [0.074, 0.086]

Table 6: Projected objective measurements on GSM8K across 16 seeds.

Transform	H	$\widetilde{\Phi}_B$	Alignment	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Log	2	0.923 [0.863, 0.969]	0.951 [0.915, 0.976]	0.334 [0.221, 0.491]	0.019 [0.013, 0.026]
Log	4	0.593 [0.509, 0.678]	0.701 [0.625, 0.775]	0.943 [0.785, 1.127]	0.033 [0.028, 0.040]
Log	8	0.429 [0.366, 0.489]	0.530 [0.462, 0.602]	1.878 [1.615, 2.147]	0.053 [0.047, 0.060]
Log	16	0.142 [0.080, 0.213]	0.226 [0.138, 0.324]	3.979 [3.370, 4.674]	0.074 [0.067, 0.081]
Power 0.25	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 0.25	2	0.930 [0.866, 0.976]	0.964 [0.937, 0.983]	0.284 [0.184, 0.429]	0.017 [0.012, 0.024]
Power 0.25	4	0.650 [0.588, 0.718]	0.753 [0.699, 0.808]	0.896 [0.745, 1.067]	0.034 [0.029, 0.039]
Power 0.25	8	0.457 [0.392, 0.520]	0.586 [0.534, 0.642]	1.805 [1.542, 2.075]	0.054 [0.046, 0.062]
Power 0.25	16	0.245 [0.131, 0.389]	0.276 [0.187, 0.364]	3.979 [3.405, 4.608]	0.076 [0.069, 0.083]
Power 0.5	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 0.5	2	0.925 [0.867, 0.967]	0.968 [0.948, 0.984]	0.256 [0.168, 0.383]	0.016 [0.011, 0.023]
Power 0.5	4	0.731 [0.659, 0.808]	0.818 [0.775, 0.861]	0.813 [0.674, 0.970]	0.033 [0.028, 0.038]
Power 0.5	8	0.485 [0.425, 0.543]	0.638 [0.589, 0.692]	1.658 [1.383, 1.946]	0.053 [0.045, 0.061]
Power 0.5	16	0.224 [0.140, 0.322]	0.323 [0.230, 0.413]	3.739 [3.194, 4.327]	0.076 [0.069, 0.084]
Power 1	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 1	2	0.915 [0.857, 0.959]	0.966 [0.940, 0.986]	0.245 [0.161, 0.367]	0.016 [0.011, 0.023]
Power 1	4	0.796 [0.719, 0.870]	0.862 [0.822, 0.900]	0.737 [0.602, 0.893]	0.032 [0.027, 0.038]
Power 1	8	0.530 [0.465, 0.596]	0.686 [0.629, 0.745]	1.528 [1.257, 1.812]	0.053 [0.045, 0.062]
Power 1	16	0.233 [0.161, 0.311]	0.360 [0.262, 0.454]	3.452 [2.921, 4.049]	0.077 [0.069, 0.084]
Power 2	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 2	2	0.916 [0.857, 0.959]	0.970 [0.953, 0.984]	0.272 [0.176, 0.413]	0.018 [0.012, 0.025]
Power 2	4	0.775 [0.698, 0.856]	0.848 [0.802, 0.888]	0.736 [0.597, 0.887]	0.033 [0.028, 0.039]
Power 2	8	0.507 [0.449, 0.563]	0.665 [0.600, 0.730]	1.519 [1.252, 1.806]	0.055 [0.046, 0.063]
Power 2	16	0.237 [0.176, 0.302]	0.363 [0.274, 0.450]	3.296 [2.775, 3.896]	0.077 [0.069, 0.085]
Top-16 log	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Top-16 log	2	0.940 [0.887, 0.979]	0.959 [0.928, 0.979]	0.287 [0.187, 0.428]	0.018 [0.012, 0.025]
Top-16 log	4	0.590 [0.510, 0.676]	0.702 [0.624, 0.778]	0.907 [0.763, 1.067]	0.034 [0.028, 0.040]
Top-16 log	8	0.426 [0.365, 0.483]	0.532 [0.468, 0.599]	1.804 [1.549, 2.079]	0.055 [0.047, 0.063]
Top-16 log	16	0.144 [0.065, 0.229]	0.223 [0.131, 0.319]	3.750 [3.224, 4.344]	0.074 [0.067, 0.081]

Table 7: Per-seed projected objective fidelity on ARC for Log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$
7201	1.000	0.761	0.254	0.161	0.073
7307	1.000	0.700	0.210	0.194	0.023
7411	1.000	0.711	0.457	0.053	0.017
7523	1.000	0.798	0.161	0.044	0.025
7643	1.000	0.881	0.646	0.274	0.065
7753	1.000	0.776	0.460	0.221	-0.049
7867	1.000	0.780	0.434	0.329	0.052
7979	1.000	1.151	1.230	-0.345	-0.008
8101	1.000	0.786	0.367	0.233	0.060
8213	1.000	0.935	0.520	0.254	0.136
8329	1.000	0.828	0.580	0.166	0.063
8443	1.000	0.927	0.262	0.134	0.085
8563	1.000	0.880	0.336	0.211	-0.031
8681	1.000	0.829	0.255	0.226	-0.138
8803	1.000	0.807	0.657	0.274	-0.011
8923	1.000	0.899	0.331	0.102	0.033

Table 8: Per-seed projected objective fidelity on ARC for Power 1.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$
7201	1.000	0.778	0.526	0.115	0.164
7307	1.000	0.778	0.413	0.229	0.049
7411	1.000	0.560	0.582	0.094	0.014
7523	1.000	0.889	0.281	0.143	-0.032
7643	1.000	1.214	0.520	0.208	0.079
7753	1.000	0.876	0.349	0.187	0.037
7867	1.000	0.745	0.604	0.464	0.123
7979	1.000	0.908	0.646	0.315	0.128
8101	1.000	0.798	0.674	0.444	0.060
8213	1.000	0.899	0.700	0.329	0.114
8329	1.000	0.915	0.578	0.255	0.135
8443	1.000	0.890	0.663	0.361	0.197
8563	1.000	0.760	0.587	0.377	0.026
8681	1.000	0.822	0.518	0.357	0.060
8803	1.000	0.941	0.834	0.371	0.014
8923	1.000	0.948	0.380	0.130	0.101

Table 9: Per-seed projected objective fidelity on ARC for Top-16 log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$
7201	1.000	0.825	0.522	0.167	0.120
7307	1.000	0.683	0.230	0.117	0.033
7411	1.000	0.707	0.558	0.156	0.008
7523	1.000	0.904	0.179	0.062	-0.004
7643	1.000	0.858	0.592	0.302	0.093
7753	1.000	0.780	0.474	0.236	-0.025
7867	1.000	0.733	0.486	0.358	0.082
7979	1.000	1.001	0.498	0.018	0.052
8101	1.000	0.808	0.416	0.238	0.086
8213	1.000	0.877	0.754	0.247	0.149
8329	1.000	0.797	0.896	0.353	0.086
8443	1.000	0.836	0.467	0.206	0.106
8563	1.000	0.853	0.285	0.186	0.005
8681	1.000	0.702	0.241	0.098	-0.053
8803	1.000	0.806	0.656	0.251	-0.022

Continued on next page

Table 9 (continued)

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$
8923	1.000	0.917	0.385	0.165	0.072

Table 10: Per-seed projected objective fidelity on GSM8K for Log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$
7201	1.000	1.029	0.746	0.655	0.296
7307	1.000	0.963	0.520	0.396	0.125
7411	1.000	1.017	0.456	0.482	0.212
7523	1.000	0.964	0.260	0.323	0.162
7643	1.000	0.934	0.872	0.425	0.158
7753	1.000	0.946	0.552	0.250	0.032
7867	1.000	0.824	0.546	0.532	-0.012
7979	1.000	0.962	0.497	0.346	0.139
8101	1.000	1.005	0.641	0.614	0.511
8213	1.000	1.022	0.620	0.349	0.024
8329	1.000	0.938	0.681	0.496	0.166
8443	1.000	0.945	0.758	0.473	-0.035
8563	1.000	0.956	0.917	0.370	0.102
8681	1.000	0.565	0.328	0.150	0.074
8803	1.000	0.860	0.564	0.510	0.312
8923	1.000	0.843	0.534	0.488	0.004

Table 11: Per-seed projected objective fidelity on GSM8K for Power 1.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$
7201	1.000	0.957	0.899	0.782	0.431
7307	1.000	0.915	0.530	0.449	0.136
7411	1.000	1.033	0.796	0.684	0.072
7523	1.000	0.982	0.882	0.591	0.125
7643	1.000	0.924	0.941	0.381	0.556
7753	1.000	0.979	0.856	0.450	0.104
7867	1.000	0.801	0.931	0.745	0.294
7979	1.000	0.981	0.582	0.416	0.224
8101	1.000	0.972	0.779	0.548	0.419
8213	1.000	0.963	0.671	0.396	0.001
8329	1.000	0.956	0.817	0.629	0.351
8443	1.000	0.947	0.901	0.589	0.174
8563	1.000	0.924	1.110	0.420	0.067
8681	1.000	0.569	0.519	0.289	0.193
8803	1.000	0.910	0.779	0.537	0.407
8923	1.000	0.834	0.738	0.582	0.179

Table 12: Per-seed projected objective fidelity on GSM8K for Top-16 log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$
7201	1.000	1.030	0.745	0.655	0.318
7307	1.000	0.970	0.532	0.402	0.122
7411	1.000	1.015	0.426	0.406	0.106
7523	1.000	0.984	0.353	0.355	0.313
7643	1.000	0.936	0.882	0.434	-0.174
7753	1.000	0.921	0.527	0.228	0.015
7867	1.000	0.979	0.580	0.512	-0.003
7979	1.000	0.961	0.510	0.336	0.169
8101	1.000	0.997	0.620	0.611	0.532
8213	1.000	1.021	0.479	0.404	0.057

Continued on next page

Table 12 (continued)

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$
8329	1.000	0.946	0.703	0.508	0.179
8443	1.000	0.959	0.769	0.419	-0.012
8563	1.000	0.961	0.952	0.393	0.166
8681	1.000	0.615	0.334	0.179	0.093
8803	1.000	0.895	0.496	0.467	0.375
8923	1.000	0.844	0.537	0.505	0.046

Table 13: Formal ARC training and evaluation across 8 seeds.

Method	Δ train gap	Δ eval gap	Grad. norm	Tokens	Accuracy	Pass@k
LL	0.009 [0.008, 0.011]	0.007 [0.005, 0.008]	21.314 [20.623, 21.944]	31.968 [31.903, 32.000]	0.256 [0.238, 0.271]	0.438 [0.383, 0.500]
CL	-0.030 [-0.045, -0.014]	-0.028 [-0.046, -0.010]	161.596 [153.319, 169.776]	31.688 [31.354, 31.931]	0.312 [0.260, 0.367]	0.586 [0.484, 0.688]
MC	-0.026 [-0.040, -0.013]	-0.025 [-0.042, -0.009]	123.356 [117.410, 128.492]	31.805 [31.541, 31.982]	0.307 [0.264, 0.354]	0.578 [0.484, 0.680]
LP	0.013 [0.012, 0.015]	0.008 [0.007, 0.010]	1.556 [1.502, 1.614]	31.959 [31.895, 31.996]	0.238 [0.207, 0.273]	0.414 [0.359, 0.461]
CP	-0.031 [-0.048, -0.015]	-0.030 [-0.050, -0.012]	11.256 [10.884, 11.621]	31.697 [31.365, 31.933]	0.264 [0.213, 0.312]	0.547 [0.445, 0.641]
TK	0.015 [0.013, 0.018]	0.006 [0.004, 0.010]	7.972 [7.791, 8.142]	31.974 [31.931, 32.000]	0.217 [0.186, 0.250]	0.383 [0.328, 0.445]

Table 14: Formal GSM8K training and evaluation across 8 seeds.

Method	Δ train gap	Δ eval gap	Grad. norm	Tokens	Accuracy	Pass@k
LL	0.015 [0.013, 0.018]	0.014 [0.010, 0.018]	14.632 [13.812, 15.733]	31.973 [31.918, 32.000]	0.070 [0.055, 0.086]	0.125 [0.102, 0.148]
CL	-0.017 [-0.032, -0.005]	-0.014 [-0.028, -0.001]	72.247 [68.393, 76.310]	31.656 [31.176, 32.000]	0.086 [0.051, 0.125]	0.227 [0.141, 0.320]
MC	-0.004 [-0.013, 0.003]	-0.002 [-0.010, 0.007]	53.982 [51.659, 56.770]	31.903 [31.774, 32.000]	0.057 [0.029, 0.092]	0.156 [0.086, 0.242]
LP	0.019 [0.017, 0.022]	0.017 [0.014, 0.021]	1.373 [1.323, 1.424]	31.973 [31.918, 32.000]	0.057 [0.043, 0.070]	0.125 [0.102, 0.148]
CP	-0.012 [-0.025, -0.001]	-0.010 [-0.023, 0.001]	6.413 [6.132, 6.725]	31.732 [31.369, 32.000]	0.061 [0.033, 0.090]	0.172 [0.094, 0.266]
TK	0.011 [0.007, 0.016]	0.006 [0.003, 0.010]	7.972 [7.514, 8.607]	31.973 [31.918, 32.000]	0.066 [0.053, 0.080]	0.117 [0.102, 0.125]

K.3 Long-Horizon Measurements

Table 15: Long-horizon objective geometry on ARC. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	8	0.166 [0.098, 0.220]	2.720 [2.450, 2.981]	0.055 [0.050, 0.060]
Log	16	0.024 [-0.008, 0.052]	5.865 [5.360, 6.401]	0.081 [0.073, 0.089]
Log	32	-0.005 [-0.019, 0.009]	12.203 [11.191, 13.332]	0.118 [0.106, 0.131]
Log	48	-0.007 [-0.015, 0.001]	17.475 [16.253, 18.543]	0.149 [0.138, 0.162]
Log	64	-0.007 [-0.012, -0.001]	23.237 [21.263, 25.135]	0.173 [0.160, 0.187]
Power 1	8	0.271 [0.215, 0.327]	2.069 [1.861, 2.299]	0.055 [0.051, 0.059]
Power 1	16	0.077 [0.047, 0.106]	4.526 [4.117, 4.978]	0.083 [0.076, 0.090]
Power 1	32	0.026 [0.011, 0.039]	10.198 [8.805, 11.815]	0.123 [0.112, 0.136]
Power 1	48	0.011 [0.002, 0.018]	15.435 [13.487, 17.515]	0.153 [0.140, 0.166]
Power 1	64	2.03×10^{-4} [-0.016, 0.011]	19.993 [17.883, 22.235]	0.173 [0.156, 0.190]
Top-16 log	8	0.199 [0.156, 0.244]	2.366 [2.175, 2.556]	0.052 [0.048, 0.056]
Top-16 log	16	0.048 [0.020, 0.075]	5.185 [4.810, 5.592]	0.080 [0.075, 0.087]
Top-16 log	32	0.008 [-0.008, 0.023]	10.955 [10.190, 11.822]	0.115 [0.106, 0.124]

Continued on next page

Table 15 (continued)

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Top-16 log	48	-0.002 [-0.010, 0.006]	16.491 [15.382, 17.642]	0.142 [0.133, 0.153]
Top-16 log	64	-0.006 [-0.013, 0.001]	22.085 [20.444, 23.854]	0.172 [0.157, 0.188]

Table 16: Per-seed long-horizon objective geometry fidelity on ARC for Log.

Seed	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
7201	0.162	0.069	0.010	0.008	0.016
7307	0.196	0.024	0.006	0.009	0.007
7411	0.050	0.023	-0.059	-0.028	-0.018
7523	0.042	0.024	-0.013	-0.005	-0.013
7643	0.289	0.066	0.029	0.019	-0.001
7753	0.218	-0.046	-0.033	-0.019	-0.019
7867	0.328	0.049	0.033	0.015	0.008
7979	-0.212	-0.017	-0.005	-0.019	-0.015
8101	0.229	0.060	0.038	0.003	-0.003
8213	0.255	0.131	0.001	0.015	0.004
8329	0.154	0.054	-0.033	-0.013	-0.023
8443	0.125	0.084	0.015	-0.009	-0.003
8563	0.209	-0.024	-0.021	-0.030	-0.018
8681	0.233	-0.146	-0.058	-0.038	-0.023
8803	0.270	-0.009	-1.15×10^{-4}	-0.015	-0.011
8923	0.106	0.037	0.010	-0.006	0.002

Table 17: Per-seed long-horizon objective geometry fidelity on ARC for Power 1.

Seed	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
7201	0.112	0.159	0.035	0.024	0.018
7307	0.238	0.047	-0.004	0.002	0.003
7411	0.084	0.024	-0.016	-0.005	-0.002
7523	0.138	-0.041	0.025	0.019	0.008
7643	0.214	0.085	0.055	0.005	-0.011
7753	0.174	0.033	-0.033	-0.001	0.002
7867	0.461	0.128	0.061	0.030	0.018
7979	0.323	0.125	0.062	0.025	0.006
8101	0.438	0.059	0.057	0.020	0.010
8213	0.328	0.096	0.043	0.039	0.029
8329	0.253	0.123	0.027	0.018	-0.102
8443	0.361	0.195	0.036	-0.001	0.006
8563	0.372	0.028	0.006	0.001	0.009
8681	0.356	0.042	-0.021	-0.030	-0.017
8803	0.344	0.012	0.041	0.018	0.020
8923	0.148	0.111	0.036	0.004	0.006

Table 18: Per-seed long-horizon objective geometry fidelity on ARC for Top-16 log.

Seed	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
7201	0.181	0.114	0.027	0.010	0.003
7307	0.134	0.040	0.002	2.81×10^{-4}	-2.43×10^{-4}
7411	0.156	0.018	-0.055	-0.028	-0.018
7523	0.062	-0.010	0.019	0.012	-0.004
7643	0.315	0.095	0.058	0.014	-0.004
7753	0.225	-0.024	-0.023	-0.002	-0.003
7867	0.360	0.081	0.041	0.014	0.006
7979	0.044	0.021	1.04×10^{-4}	-0.022	-0.026
8101	0.236	0.092	0.053	0.014	0.004

Continued on next page

Table 18 (continued)

Seed	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
8213	0.247	0.146	0.022	0.027	0.015
8329	0.333	0.077	0.003	-0.004	-0.044
8443	0.207	0.108	0.018	-0.015	-0.005
8563	0.182	0.009	-0.024	-0.018	-0.008
8681	0.088	-0.061	-0.053	-0.032	-0.020
8803	0.249	-0.018	0.008	-0.008	-4.72×10^{-5}
8923	0.170	0.074	0.030	0.009	0.010

Table 19: Long-horizon objective geometry on GSM8K. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	8	0.429 [0.366, 0.489]	1.878 [1.623, 2.154]	0.053 [0.047, 0.060]
Log	16	0.142 [0.078, 0.215]	3.979 [3.365, 4.695]	0.074 [0.067, 0.081]
Log	32	0.009 [-0.008, 0.027]	7.089 [6.469, 7.757]	0.098 [0.090, 0.107]
Log	48	-0.004 [-0.017, 0.008]	9.934 [8.929, 11.040]	0.114 [0.105, 0.123]
Log	64	-0.004 [-0.016, 0.009]	12.113 [10.738, 13.629]	0.127 [0.119, 0.135]
Power 1	8	0.527 [0.464, 0.595]	1.532 [1.256, 1.825]	0.054 [0.046, 0.062]
Power 1	16	0.236 [0.156, 0.321]	3.463 [2.921, 4.084]	0.077 [0.069, 0.084]
Power 1	32	0.028 [0.006, 0.050]	6.603 [5.980, 7.306]	0.110 [0.100, 0.119]
Power 1	48	0.015 [3.07×10^{-4}, 0.032]	9.041 [8.196, 9.936]	0.131 [0.120, 0.141]
Power 1	64	0.013 [-0.001, 0.029]	11.100 [9.886, 12.460]	0.144 [0.132, 0.157]
Top-16 log	8	0.427 [0.368, 0.484]	1.806 [1.552, 2.083]	0.055 [0.047, 0.063]
Top-16 log	16	0.136 [0.060, 0.220]	3.761 [3.229, 4.349]	0.074 [0.067, 0.081]
Top-16 log	32	0.012 [-0.009, 0.033]	6.940 [6.339, 7.609]	0.098 [0.089, 0.107]
Top-16 log	48	-0.001 [-0.012, 0.010]	9.748 [8.815, 10.806]	0.116 [0.108, 0.124]
Top-16 log	64	-0.001 [-0.014, 0.012]	11.949 [10.762, 13.320]	0.131 [0.123, 0.139]

Table 20: Per-seed long-horizon objective geometry fidelity on GSM8K for Log.

Seed	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
7201	0.662	0.300	0.096	0.049	0.040
7307	0.399	0.124	0.007	0.002	-0.001
7411	0.478	0.217	-0.020	0.008	0.007
7523	0.324	0.120	-0.038	-0.011	0.004
7643	0.421	0.197	-0.006	-0.006	0.004
7753	0.250	0.029	-0.012	-0.009	-0.006
7867	0.547	-0.016	-0.005	-0.013	-0.001
7979	0.354	0.136	0.043	0.012	-0.016
8101	0.611	0.512	0.027	-0.035	-0.050
8213	0.347	0.024	-0.041	-0.063	-0.063
8329	0.488	0.174	-0.012	-0.001	0.005
8443	0.476	-0.033	-0.015	-0.004	-0.004
8563	0.363	0.100	-0.003	-0.006	-0.009
8681	0.152	0.076	0.039	0.017	0.003
8803	0.511	0.311	0.038	0.028	0.045
8923	0.474	-0.002	0.046	-0.033	-0.016

Table 21: Per-seed long-horizon objective geometry fidelity on GSM8K for Power 1.

Seed	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
7201	0.792	0.432	0.017	-0.008	-0.008
7307	0.445	0.137	0.006	-0.023	-0.019
7411	0.696	0.085	-0.035	-0.007	-0.025
7523	0.573	0.130	0.006	0.019	0.032

Continued on next page

Table 21 (continued)

Seed	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
7643	0.384	0.646	-0.012	-0.027	0.006
7753	0.422	0.100	0.033	0.009	-1.29×10^{-4}
7867	0.739	0.296	-0.026	-0.018	0.007
7979	0.427	0.227	0.052	0.035	0.009
8101	0.551	0.420	0.049	0.009	-0.011
8213	0.400	-0.013	0.035	0.013	-0.020
8329	0.627	0.349	0.040	0.030	0.035
8443	0.588	0.166	-0.009	0.001	0.008
8563	0.421	0.031	-0.023	0.005	-2.17×10^{-4}
8681	0.298	0.196	0.116	0.106	0.082
8803	0.494	0.381	0.085	0.051	0.070
8923	0.575	0.193	0.110	0.047	0.047

Table 22: Per-seed long-horizon objective geometry fidelity on GSM8K for Top-16 log.

Seed	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
7201	0.662	0.317	0.097	0.042	0.041
7307	0.406	0.122	0.007	-0.005	-0.008
7411	0.404	0.106	-0.013	0.003	-0.005
7523	0.359	0.240	-0.017	0.004	0.017
7643	0.430	-0.168	-0.001	-4.73×10^{-4}	0.008
7753	0.225	0.010	-0.019	-0.019	-0.012
7867	0.524	-0.010	-0.007	-0.012	0.002
7979	0.344	0.169	0.044	0.019	-0.015
8101	0.609	0.532	0.042	-0.029	-0.047
8213	0.405	0.055	-0.035	-0.043	-0.045
8329	0.500	0.186	-0.010	0.012	0.021
8443	0.429	-0.023	-0.076	-0.033	-0.020
8563	0.388	0.158	-0.009	-0.007	-0.016
8681	0.183	0.094	0.050	0.022	0.006
8803	0.464	0.349	0.052	0.037	0.055
8923	0.498	0.044	0.081	-0.009	-0.001

K.4 Direction-Robustness Measurements

Table 23: Eight-direction objective geometry on ARC. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	4	0.368 [0.255, 0.475]	1.624 [1.290, 2.012]	0.060 [0.055, 0.066]
Log	8	0.139 [0.086, 0.194]	2.781 [2.500, 3.112]	0.088 [0.084, 0.092]
Log	16	0.009 [-0.008, 0.026]	5.801 [5.221, 6.468]	0.131 [0.121, 0.141]
Power 1	4	0.511 [0.413, 0.594]	1.196 [1.046, 1.371]	0.065 [0.060, 0.071]
Power 1	8	0.230 [0.154, 0.300]	2.059 [1.870, 2.248]	0.089 [0.085, 0.094]
Power 1	16	0.061 [0.021, 0.103]	4.456 [3.941, 5.073]	0.132 [0.121, 0.144]
Top-16 log	4	0.401 [0.299, 0.510]	1.349 [1.064, 1.673]	0.061 [0.056, 0.068]
Top-16 log	8	0.167 [0.118, 0.219]	2.356 [2.184, 2.545]	0.085 [0.082, 0.088]
Top-16 log	16	0.030 [0.005, 0.054]	5.092 [4.719, 5.508]	0.133 [0.123, 0.143]

Table 24: Per-seed eight-direction objective geometry fidelity on ARC for Log.

Seed	$H = 4$	$H = 8$	$H = 16$
7201	0.241	0.136	0.035
7307	0.215	0.129	0.020
7411	0.573	0.037	-0.035

Continued on next page

Table 24 (continued)

Seed	$H = 4$	$H = 8$	$H = 16$
7523	0.088	0.035	-0.008
7643	0.442	0.150	-0.001
7753	0.542	0.253	0.044
7867	0.374	0.253	-0.009
7979	0.470	0.119	0.029

Table 25: Per-seed eight-direction objective geometry fidelity on ARC for Power 1.

Seed	$H = 4$	$H = 8$	$H = 16$
7201	0.434	0.213	0.141
7307	0.459	0.243	0.028
7411	0.690	0.017	-0.027
7523	0.225	0.140	0.039
7643	0.539	0.236	-0.008
7753	0.544	0.351	0.096
7867	0.577	0.372	0.072
7979	0.618	0.270	0.147

Table 26: Per-seed eight-direction objective geometry fidelity on ARC for Top-16 log.

Seed	$H = 4$	$H = 8$	$H = 16$
7201	0.254	0.137	0.074
7307	0.282	0.167	0.019
7411	0.674	0.088	-0.026
7523	0.187	0.062	0.003
7643	0.437	0.175	0.008
7753	0.547	0.291	0.080
7867	0.408	0.269	0.017
7979	0.416	0.145	0.062

Table 27: Eight-direction objective geometry on GSM8K. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	4	0.636 [0.572, 0.694]	0.879 [0.742, 1.047]	0.047 [0.040, 0.056]
Log	8	0.437 [0.360, 0.516]	1.788 [1.440, 2.101]	0.080 [0.065, 0.096]
Log	16	0.102 [0.066, 0.137]	3.923 [3.136, 4.824]	0.119 [0.102, 0.133]
Power 1	4	0.764 [0.668, 0.853]	0.811 [0.620, 1.039]	0.053 [0.044, 0.063]
Power 1	8	0.530 [0.423, 0.649]	1.646 [1.240, 2.025]	0.085 [0.069, 0.099]
Power 1	16	0.181 [0.109, 0.269]	3.676 [3.023, 4.386]	0.126 [0.114, 0.138]
Top-16 log	4	0.640 [0.577, 0.695]	0.856 [0.719, 1.019]	0.048 [0.040, 0.057]
Top-16 log	8	0.436 [0.353, 0.518]	1.707 [1.381, 2.026]	0.081 [0.066, 0.098]
Top-16 log	16	0.101 [0.064, 0.139]	3.734 [3.044, 4.543]	0.119 [0.104, 0.133]

Table 28: Per-seed eight-direction objective geometry fidelity on GSM8K for Log.

Seed	$H = 4$	$H = 8$	$H = 16$
7201	0.714	0.549	0.166
7307	0.612	0.529	0.098
7411	0.717	0.551	0.157
7523	0.601	0.388	0.122
7643	0.706	0.299	0.024
7753	0.567	0.319	0.048
7867	0.709	0.557	0.050
7979	0.459	0.304	0.150

Table 29: Per-seed eight-direction objective geometry fidelity on GSM8K for Power 1.

Seed	$H = 4$	$H = 8$	$H = 16$
7201	0.937	0.808	0.438
7307	0.557	0.476	0.142
7411	0.830	0.736	0.090
7523	0.790	0.490	0.164
7643	0.884	0.336	0.048
7753	0.667	0.365	0.099
7867	0.869	0.628	0.262
7979	0.578	0.401	0.205

Table 30: Per-seed eight-direction objective geometry fidelity on GSM8K for Top-16 log.

Seed	$H = 4$	$H = 8$	$H = 16$
7201	0.708	0.542	0.178
7307	0.621	0.531	0.106
7411	0.708	0.578	0.082
7523	0.625	0.394	0.124
7643	0.711	0.293	0.028
7753	0.535	0.294	0.040
7867	0.731	0.551	0.068
7979	0.481	0.307	0.179

K.5 HellaSwag Direction Robustness

Table 31: HellaSwag eight-direction objective geometry on HELLASWAG. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	4	1.172 [1.066, 1.282]	0.777 [0.672, 0.892]	0.034 [0.030, 0.039]
Log	8	-0.050 [-0.093, -0.007]	3.179 [2.857, 3.515]	0.081 [0.073, 0.090]
Log	16	-0.029 [-0.045, -0.014]	6.513 [6.052, 6.947]	0.132 [0.123, 0.142]
Power 1	4	0.609 [0.552, 0.669]	1.070 [0.872, 1.327]	0.063 [0.055, 0.072]
Power 1	8	0.069 [0.025, 0.113]	2.892 [2.680, 3.108]	0.109 [0.101, 0.116]
Power 1	16	0.014 [-0.008, 0.035]	5.517 [5.234, 5.857]	0.159 [0.152, 0.168]
Top-16 log	4	0.937 [0.851, 1.017]	0.823 [0.709, 0.958]	0.040 [0.036, 0.045]
Top-16 log	8	0.009 [-0.031, 0.046]	2.964 [2.756, 3.189]	0.093 [0.086, 0.100]
Top-16 log	16	-0.004 [-0.019, 0.011]	5.917 [5.631, 6.201]	0.145 [0.138, 0.152]

Table 32: Per-seed hellaswag eight-direction objective geometry fidelity on HELLASWAG for Log.

Seed	$H = 4$	$H = 8$	$H = 16$
10103	0.950	-0.063	-0.010
10211	1.201	-0.111	-0.043
10321	1.489	-0.220	-0.079
10433	0.916	0.114	0.014
10501	1.385	0.063	0.018
10613	1.344	-0.050	-0.059
10723	1.071	-0.019	-0.010
10831	0.800	-0.102	-0.045
10909	0.863	0.061	0.014
11027	1.275	0.015	-0.043
11131	1.169	-0.048	-0.006
11243	1.171	-0.132	-0.080
11353	1.198	-0.043	-0.020
11467	1.552	-0.081	-0.037
11579	0.936	0.002	-0.020

Continued on next page

Table 32 (continued)

Seed	$H = 4$	$H = 8$	$H = 16$
11689	1.440	-0.189	-0.066

Table 33: Per-seed hellaswag eight-direction objective geometry fidelity on HELLASWAG for Power 1.

Seed	$H = 4$	$H = 8$	$H = 16$
10103	0.533	0.136	0.036
10211	0.658	-0.102	-0.033
10321	0.443	-0.009	0.035
10433	0.569	0.146	0.071
10501	0.408	-0.005	0.006
10613	0.879	0.103	0.012
10723	0.646	0.221	0.104
10831	0.601	0.139	0.005
10909	0.693	0.136	0.048
11027	0.719	0.135	0.015
11131	0.481	0.012	-0.041
11243	0.669	-0.021	-0.093
11353	0.749	0.072	0.030
11467	0.597	0.011	0.030
11579	0.480	0.182	0.013
11689	0.621	-0.047	-0.006

Table 34: Per-seed hellaswag eight-direction objective geometry fidelity on HELLASWAG for Top-16 log.

Seed	$H = 4$	$H = 8$	$H = 16$
10103	0.798	-0.019	0.013
10211	1.183	-0.119	-0.041
10321	1.036	-0.125	-0.015
10433	0.707	0.065	0.043
10501	1.138	0.072	0.036
10613	0.982	0.047	-0.014
10723	0.894	0.124	0.038
10831	0.591	0.034	-0.049
10909	0.811	0.076	0.028
11027	1.106	0.124	-0.010
11131	0.984	-0.082	-0.035
11243	1.020	-0.047	-0.054
11353	1.077	0.032	0.018
11467	0.884	-0.102	0.005
11579	0.703	0.067	-0.004
11689	1.075	-0.008	-0.018

K.6 Model-Scale Measurements

Table 35: Qwen3.5 4B-to-9B objective geometry on ARC for Log. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Log	2	0.881 [0.805, 0.958]	0.356 [0.307, 0.405]	0.006 [0.005, 0.007]
Log	4	0.719 [0.633, 0.783]	0.598 [0.430, 0.775]	0.010 [0.008, 0.013]
Log	8	0.368 [0.316, 0.427]	1.856 [1.512, 2.201]	0.017 [0.015, 0.021]
Log	16	0.129 [0.071, 0.185]	3.248 [2.181, 4.315]	0.032 [0.021, 0.044]
Log	32	0.045 [-0.006, 0.096]	5.432 [5.120, 5.660]	0.044 [0.038, 0.051]
Log	48	0.012 [-0.042, 0.064]	7.391 [6.060, 8.442]	0.054 [0.046, 0.061]
Log	64	0.012 [-0.008, 0.031]	9.528 [8.031, 10.414]	0.063 [0.055, 0.071]

Table 36: Qwen3.5 4B-to-9B objective geometry on ARC for Power 1. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Power 1	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 1	2	0.917 [0.860, 0.961]	0.254 [0.197, 0.332]	0.006 [0.005, 0.007]
Power 1	4	0.749 [0.669, 0.820]	0.451 [0.281, 0.666]	0.010 [0.007, 0.012]
Power 1	8	0.506 [0.454, 0.549]	1.670 [1.410, 1.948]	0.020 [0.017, 0.023]
Power 1	16	0.306 [0.182, 0.403]	2.994 [2.077, 3.952]	0.029 [0.024, 0.035]
Power 1	32	0.094 [0.035, 0.154]	4.917 [4.229, 5.659]	0.043 [0.041, 0.045]
Power 1	48	0.051 [-0.005, 0.104]	7.326 [5.591, 10.465]	0.059 [0.056, 0.063]
Power 1	64	0.031 [0.015, 0.052]	9.431 [7.939, 11.787]	0.066 [0.057, 0.074]

Table 37: Qwen3.5 4B-to-9B objective geometry on ARC for Top-16 log. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Top-16 log	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Top-16 log	2	0.973 [0.905, 1.030]	0.291 [0.254, 0.324]	0.005 [0.004, 0.006]
Top-16 log	4	0.732 [0.614, 0.814]	0.527 [0.378, 0.670]	0.009 [0.007, 0.010]
Top-16 log	8	0.388 [0.345, 0.427]	1.775 [1.455, 2.096]	0.017 [0.015, 0.021]
Top-16 log	16	0.140 [0.066, 0.213]	3.165 [2.180, 4.149]	0.031 [0.022, 0.040]
Top-16 log	32	0.046 [-0.016, 0.107]	5.176 [4.931, 5.443]	0.044 [0.041, 0.048]
Top-16 log	48	0.013 [-0.043, 0.069]	7.338 [6.632, 8.309]	0.054 [0.049, 0.058]
Top-16 log	64	0.020 [0.004, 0.035]	9.431 [8.771, 10.091]	0.061 [0.052, 0.069]

Table 38: Per-seed qwen3.5 4b-to-9b objective geometry fidelity on ARC for Log.

Seed	$H=1$	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$	$H=48$	$H=64$
8609	1.000	0.985	0.739	0.298	0.040	-0.014	-0.070	-0.001
8719	1.000	0.931	0.798	0.369	0.166	0.086	0.087	0.032
8831	1.000	0.797	0.596	0.451	0.107	0.001	-0.011	-0.014
8941	1.000	0.813	0.744	0.353	0.205	0.107	0.041	0.031

Table 39: Per-seed qwen3.5 4b-to-9b objective geometry fidelity on ARC for Power 1.

Seed	$H=1$	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$	$H=48$	$H=64$
8609	1.000	0.972	0.729	0.429	0.114	0.020	-0.039	0.013
8719	1.000	0.927	0.848	0.528	0.420	0.122	0.112	0.031
8831	1.000	0.934	0.627	0.503	0.303	0.050	0.035	0.018
8941	1.000	0.835	0.793	0.565	0.385	0.185	0.096	0.064

Table 40: Per-seed qwen3.5 4b-to-9b objective geometry fidelity on ARC for Top-16 log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
8609	1.000	1.009	0.777	0.321	0.039	-0.025	-0.074	0.013
8719	1.000	0.962	0.833	0.415	0.197	0.082	0.084	0.033
8831	1.000	0.870	0.560	0.439	0.094	-0.007	-0.012	-0.004
8941	1.000	1.051	0.756	0.379	0.229	0.132	0.053	0.038

Table 41: Qwen3.5 4B-to-9B objective geometry on GSM8K for Log. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Log	2	0.954 [0.901, 0.984]	0.080 [0.025, 0.162]	0.002 [0.001, 0.005]
Log	4	0.871 [0.782, 0.960]	0.274 [0.087, 0.461]	0.004 [0.002, 0.007]
Log	8	0.855 [0.784, 0.926]	0.445 [0.165, 0.726]	0.008 [0.004, 0.014]
Log	16	0.637 [0.476, 0.845]	1.213 [0.918, 1.504]	0.018 [0.014, 0.021]
Log	32	0.372 [0.215, 0.488]	2.701 [1.987, 3.601]	0.028 [0.022, 0.037]
Log	48	0.342 [0.180, 0.516]	3.350 [2.609, 4.217]	0.032 [0.025, 0.042]
Log	64	0.191 [0.131, 0.253]	3.615 [2.941, 4.215]	0.037 [0.033, 0.043]

Table 42: Qwen3.5 4B-to-9B objective geometry on GSM8K for Power 1. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Power 1	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 1	2	0.957 [0.888, 0.992]	0.129 [0.023, 0.311]	0.004 [0.001, 0.010]
Power 1	4	0.872 [0.770, 0.975]	0.332 [0.075, 0.588]	0.007 [0.002, 0.011]
Power 1	8	0.855 [0.757, 0.952]	0.449 [0.140, 0.758]	0.010 [0.004, 0.018]
Power 1	16	0.632 [0.502, 0.799]	1.287 [1.014, 1.559]	0.019 [0.015, 0.023]
Power 1	32	0.283 [0.171, 0.396]	2.592 [1.849, 3.448]	0.030 [0.024, 0.037]
Power 1	48	0.223 [0.108, 0.337]	3.203 [2.579, 3.987]	0.032 [0.027, 0.037]
Power 1	64	0.209 [0.147, 0.277]	3.335 [2.656, 4.080]	0.036 [0.033, 0.039]

Table 43: Qwen3.5 4B-to-9B objective geometry on GSM8K for Top-16 log. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Top-16 log	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Top-16 log	2	0.956 [0.904, 0.986]	0.076 [0.024, 0.152]	0.002 [0.001, 0.005]
Top-16 log	4	0.874 [0.785, 0.962]	0.268 [0.079, 0.456]	0.004 [0.002, 0.007]
Top-16 log	8	0.855 [0.783, 0.927]	0.435 [0.154, 0.716]	0.008 [0.003, 0.014]
Top-16 log	16	0.633 [0.486, 0.830]	1.204 [0.920, 1.516]	0.018 [0.014, 0.021]
Top-16 log	32	0.350 [0.220, 0.425]	2.678 [1.997, 3.524]	0.028 [0.022, 0.037]
Top-16 log	48	0.310 [0.186, 0.416]	3.324 [2.588, 4.174]	0.032 [0.025, 0.041]
Top-16 log	64	0.194 [0.136, 0.255]	3.558 [2.883, 4.181]	0.037 [0.033, 0.043]

Table 44: Per-seed qwen3.5 4b-to-9b objective geometry fidelity on GSM8K for Log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
8201	1.000	0.979	0.786	0.800	0.964	0.428	0.303	0.194
8307	1.000	0.876	0.779	0.768	0.489	0.144	0.121	0.109
8411	1.000	0.987	0.951	0.939	0.464	0.517	0.587	0.182
8523	1.000	0.972	0.969	0.913	0.632	0.399	0.355	0.277

Table 45: Per-seed qwen3.5 4b-to-9b objective geometry fidelity on GSM8K for Power 1.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
8201	1.000	0.992	0.748	0.775	0.898	0.405	0.339	0.228
8307	1.000	0.854	0.791	0.740	0.503	0.193	0.156	0.131
8411	1.000	0.992	0.963	0.942	0.501	0.148	0.059	0.162
8523	1.000	0.990	0.987	0.961	0.628	0.388	0.335	0.315

Table 46: Per-seed qwen3.5 4b-to-9b objective geometry fidelity on GSM8K for Top-16 log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
8201	1.000	0.980	0.786	0.797	0.941	0.422	0.297	0.193
8307	1.000	0.879	0.784	0.769	0.495	0.153	0.129	0.117
8411	1.000	0.989	0.954	0.939	0.477	0.429	0.456	0.188
8523	1.000	0.975	0.971	0.916	0.618	0.398	0.357	0.277

K.7 HellaSwag Measurements

Table 47: HellaSwag objective geometry on HELLASWAG for Log. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Log	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Log	2	1.123 [1.079, 1.167]	0.223 [0.178, 0.286]	0.009 [0.008, 0.011]
Log	4	1.179 [1.024, 1.340]	0.799 [0.679, 0.930]	0.021 [0.018, 0.025]
Log	8	-0.025 [-0.081, 0.041]	3.454 [2.969, 3.991]	0.056 [0.048, 0.064]
Log	16	-0.020 [-0.050, 0.007]	6.810 [6.041, 7.671]	0.090 [0.080, 0.100]
Log	32	-0.010 [-0.018, -0.001]	12.754 [11.791, 13.794]	0.143 [0.130, 0.156]
Log	48	-0.002 [-0.008, 0.003]	18.873 [17.336, 20.441]	0.175 [0.163, 0.187]
Log	64	-0.001 [-0.007, 0.004]	26.262 [24.316, 28.224]	0.213 [0.197, 0.229]

Table 48: HellaSwag objective geometry on HELLASWAG for Power 1. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Power 1	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Power 1	2	1.039 [0.980, 1.107]	0.360 [0.238, 0.524]	0.019 [0.015, 0.024]
Power 1	4	0.556 [0.488, 0.623]	0.992 [0.828, 1.193]	0.036 [0.031, 0.041]
Power 1	8	0.049 [-0.030, 0.126]	2.798 [2.458, 3.150]	0.066 [0.059, 0.073]
Power 1	16	0.017 [-0.014, 0.046]	5.448 [4.813, 6.153]	0.099 [0.091, 0.106]
Power 1	32	0.016 [0.003, 0.027]	10.773 [9.779, 11.864]	0.147 [0.135, 0.160]
Power 1	48	0.011 [0.001, 0.021]	15.809 [14.255, 17.426]	0.183 [0.167, 0.201]
Power 1	64	0.008 [0.001, 0.014]	21.741 [19.778, 23.864]	0.218 [0.201, 0.236]

Table 49: HellaSwag objective geometry on HELLASWAG for Top-16 log. Cells are seed means with 95% bootstrap intervals.

Transform	H	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$	$\|C_B\|/\|u_B\|$
Top-16 log	1	1.000 [1.000, 1.000]	0 [0, 0]	0 [0, 0]
Top-16 log	2	1.079 [1.022, 1.143]	0.264 [0.206, 0.340]	0.011 [0.009, 0.013]
Top-16 log	4	0.970 [0.795, 1.192]	0.878 [0.759, 1.006]	0.027 [0.023, 0.031]
Top-16 log	8	0.075 [-0.051, 0.266]	3.168 [2.782, 3.580]	0.064 [0.055, 0.073]
Top-16 log	16	0.010 [-0.019, 0.040]	6.277 [5.581, 7.085]	0.098 [0.090, 0.107]
Top-16 log	32	0.007 [-0.001, 0.016]	11.756 [10.804, 12.733]	0.141 [0.125, 0.157]
Top-16 log	48	0.009 [0.002, 0.015]	17.721 [16.202, 19.317]	0.176 [0.164, 0.188]
Top-16 log	64	0.005 [-0.002, 0.011]	25.637 [23.358, 28.205]	0.212 [0.198, 0.226]

Table 50: Per-seed hellaswag objective geometry fidelity on HELLASWAG for Log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
10103	1.000	1.012	0.899	-0.063	0.039	0.010	0.004	-0.006
10211	1.000	1.091	1.159	-0.130	-0.061	-0.024	-0.011	-0.013
10321	1.000	1.244	1.554	0.008	0.047	0.004	0.004	0.002
10433	1.000	1.140	0.996	0.064	0.051	0.006	-0.005	-0.004
10501	1.000	1.258	1.671	0.352	0.064	-0.035	0.002	-1.18×10^{-4}
10613	1.000	1.171	1.188	-0.051	-0.047	-0.038	-0.015	0.001
10723	1.000	1.101	1.077	-0.129	-0.006	-0.016	-0.006	-4.15×10^{-4}
10831	1.000	1.125	0.745	-0.166	-0.029	-0.024	-0.012	-0.010
10909	1.000	1.121	0.743	0.082	0.033	0.023	0.016	0.011
11027	1.000	1.061	1.223	0.046	-0.019	-0.004	0.009	0.008
11131	1.000	0.934	1.139	-0.146	-0.012	0.017	0.011	0.016
11243	1.000	1.112	1.273	-0.112	-0.173	-0.005	-0.002	0.002
11353	1.000	1.039	1.247	0.010	-0.028	-0.007	0.013	0.017
11467	1.000	1.289	1.827	-0.057	-0.037	-0.011	-3.47×10^{-4}	-0.006
11579	1.000	1.086	0.652	0.033	-0.070	-0.028	-0.028	-0.030
11689	1.000	1.182	1.468	-0.141	-0.078	-0.026	-0.017	-0.011

Table 51: Per-seed hellaswag objective geometry fidelity on HELLASWAG for Power 1.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
10103	1.000	0.955	0.557	0.064	0.039	0.032	0.017	0.002
10211	1.000	1.069	0.405	-0.057	-0.032	-0.003	0.008	-0.003
10321	1.000	0.960	0.304	0.080	0.103	0.027	0.017	0.007
10433	1.000	0.996	0.449	0.012	0.059	0.017	0.007	0.003
10501	1.000	1.405	0.378	-0.082	-0.099	0.033	0.037	0.017
10613	1.000	0.996	0.698	0.111	0.023	-0.004	0.009	0.015
10723	1.000	0.941	0.520	0.233	0.088	0.035	0.032	0.017
10831	1.000	1.089	0.607	0.029	0.001	-0.043	-0.029	-0.018
10909	1.000	1.086	0.665	0.085	0.051	0.034	0.034	0.030
11027	1.000	1.010	0.735	0.129	-0.027	0.025	0.041	0.023
11131	1.000	0.793	0.357	-0.161	-0.114	-0.033	-0.025	-0.007
11243	1.000	1.079	0.603	0.118	0.076	0.018	0.016	0.016
11353	1.000	0.991	0.683	0.058	0.021	0.002	0.006	0.017
11467	1.000	1.123	0.773	0.088	0.078	0.046	0.022	0.015
11579	1.000	0.962	0.574	0.405	-0.009	0.037	-0.015	-0.021
11689	1.000	1.162	0.591	-0.331	0.014	0.025	0.001	0.010

Table 52: Per-seed hellaswag objective geometry fidelity on HELLASWAG for Top-16 log.

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
10103	1.000	1.002	0.723	-0.019	0.054	0.038	0.017	0.001
10211	1.000	1.086	1.015	-0.057	-0.015	-0.012	-5.73×10^{-5}	-0.003
10321	1.000	1.007	0.950	0.087	0.075	0.029	0.018	0.010
10433	1.000	1.053	0.620	-0.100	0.066	0.009	-0.001	-0.003
10501	1.000	1.372	2.275	1.353	0.155	0.005	0.024	0.016
10613	1.000	1.076	0.771	0.079	-0.011	-0.016	0.002	0.011
10723	1.000	1.001	0.976	-0.049	0.030	0.008	0.011	0.005
10831	1.000	1.090	0.506	-0.017	-0.023	-0.019	-0.008	-0.009
10909	1.000	1.104	0.722	0.098	0.056	0.030	0.028	0.020
11027	1.000	1.029	1.142	0.188	-0.028	0.009	0.021	0.010
11131	1.000	0.855	1.232	-0.245	-0.077	0.011	0.021	0.026
11243	1.000	1.095	0.962	-0.103	-0.060	0.033	0.021	0.009
11353	1.000	1.035	1.194	0.033	0.011	-0.008	0.013	0.019
11467	1.000	1.292	0.726	-0.079	0.012	0.010	4.63×10^{-4}	0.002
11579	1.000	0.965	0.530	0.132	-0.010	0.001	-0.018	-0.032

Continued on next page

Table 52 (continued)

Seed	$H = 1$	$H = 2$	$H = 4$	$H = 8$	$H = 16$	$H = 32$	$H = 48$	$H = 64$
11689	1.000	1.207	1.181	-0.101	-0.080	-0.018	-0.006	-0.002

K.8 Support Sweep

Table 53: Teacher-selected support sweep on ARC. Cells are seed means with 95% bootstrap intervals.

K	H	Teacher mass	Student mass	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$
4	4	0.955 [0.948, 0.962]	0.745 [0.718, 0.770]	0.525 [0.478, 0.578]	1.232 [1.022, 1.456]
4	8	0.962 [0.955, 0.968]	0.789 [0.765, 0.812]	0.308 [0.249, 0.362]	2.170 [1.999, 2.429]
4	16	0.952 [0.945, 0.958]	0.777 [0.763, 0.789]	0.066 [0.037, 0.099]	4.483 [3.843, 5.161]
64	4	0.998 [0.998, 0.999]	0.960 [0.953, 0.967]	0.461 [0.307, 0.642]	1.482 [1.178, 1.815]
64	8	0.998 [0.998, 0.999]	0.967 [0.960, 0.973]	0.206 [0.135, 0.271]	2.629 [2.355, 2.896]
64	16	0.997 [0.996, 0.998]	0.960 [0.955, 0.964]	0.033 [0.006, 0.058]	5.607 [4.959, 6.245]

Table 54: Teacher-selected support sweep on GSM8K. Cells are seed means with 95% bootstrap intervals.

K	H	Teacher mass	Student mass	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$
4	4	0.986 [0.983, 0.988]	0.931 [0.915, 0.945]	0.675 [0.535, 0.840]	0.839 [0.668, 1.006]
4	8	0.988 [0.985, 0.991]	0.934 [0.921, 0.949]	0.415 [0.342, 0.496]	1.655 [1.319, 1.948]
4	16	0.984 [0.980, 0.988]	0.912 [0.902, 0.925]	0.167 [0.104, 0.232]	3.634 [2.800, 4.643]
64	4	1.000 [0.999, 1.000]	0.995 [0.993, 0.997]	0.554 [0.438, 0.678]	0.878 [0.686, 1.083]
64	8	1.000 [0.999, 1.000]	0.994 [0.992, 0.995]	0.415 [0.334, 0.500]	1.857 [1.521, 2.152]
64	16	0.999 [0.999, 1.000]	0.992 [0.990, 0.993]	0.073 [-0.027, 0.167]	4.071 [3.118, 5.153]

K.9 HellaSwag Support Sweep

Table 55: Teacher-selected support sweep on HELLASWAG. Cells are seed means with 95% bootstrap intervals.

K	H	Teacher mass	Student mass	$\widetilde{\Phi}_B$	$\|c_B\|/\|u_B\|$
4	4	0.957 [0.952, 0.961]	0.808 [0.798, 0.818]	0.666 [0.610, 0.719]	0.941 [0.797, 1.084]
4	8	0.936 [0.932, 0.941]	0.729 [0.717, 0.743]	0.170 [0.095, 0.247]	2.405 [2.149, 2.671]
4	16	0.932 [0.926, 0.939]	0.740 [0.724, 0.758]	0.035 [4.39×10^{-4}, 0.068]	4.830 [4.268, 5.451]
64	4	0.999 [0.998, 0.999]	0.982 [0.980, 0.983]	1.158 [1.017, 1.304]	0.768 [0.660, 0.889]
64	8	0.996 [0.996, 0.997]	0.946 [0.942, 0.951]	-0.014 [-0.077, 0.057]	3.159 [2.731, 3.615]
64	16	0.995 [0.995, 0.996]	0.945 [0.940, 0.950]	-0.014 [-0.044, 0.015]	6.439 [5.730, 7.223]

K.10 Complete Training Outcomes

Table 56: Per-seed ARC training outcomes, method group 1.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
7201	Local log	0.008	0.281	0.562	22.487	1.000
7201	Corrected log	-0.035	0.375	0.625	162.547	1.000
7201	Sequence log MC	-0.027	0.312	0.562	121.215	1.000
7307	Local log	0.005	0.266	0.438	20.495	1.000
7307	Corrected log	-0.019	0.250	0.625	151.036	1.000
7307	Sequence log MC	-0.016	0.328	0.688	122.110	1.000
7411	Local log	0.004	0.266	0.562	20.490	1.000
7411	Corrected log	-0.072	0.391	0.750	164.921	1.000
7411	Sequence log MC	-0.072	0.438	0.812	120.061	1.000
7523	Local log	0.007	0.234	0.312	21.455	1.000
7523	Corrected log	-0.010	0.219	0.375	161.588	1.000

Continued on next page

Table 56 (continued)

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
7523	Sequence log MC	-0.013	0.234	0.375	122.864	1.000
7643	Local log	0.007	0.250	0.375	22.369	1.000
7643	Corrected log	0.003	0.219	0.375	155.031	1.000
7643	Sequence log MC	0.003	0.266	0.438	130.903	1.000
7753	Local log	0.011	0.281	0.438	21.754	1.000
7753	Corrected log	-0.040	0.406	0.812	179.421	1.000
7753	Sequence log MC	-0.039	0.344	0.625	134.373	1.000
7867	Local log	0.006	0.266	0.438	19.542	1.000
7867	Corrected log	0.006	0.266	0.500	140.949	1.000
7867	Sequence log MC	0.007	0.219	0.438	106.456	1.000
7979	Local log	0.005	0.203	0.375	21.921	1.000
7979	Corrected log	-0.057	0.375	0.625	177.275	1.000
7979	Sequence log MC	-0.039	0.312	0.688	128.867	1.000

Table 57: Per-seed ARC training outcomes, method group 2.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
7201	Local power	0.009	0.188	0.375	1.577	0.883
7201	Corrected power	-0.031	0.266	0.562	10.821	1.000
7201	Top-16 log	0.008	0.281	0.500	8.169	1.000
7307	Local power	0.007	0.312	0.500	1.442	0.891
7307	Corrected power	-0.021	0.203	0.500	10.815	1.000
7307	Top-16 log	0.003	0.188	0.312	7.537	1.000
7411	Local power	0.008	0.219	0.438	1.555	0.930
7411	Corrected power	-0.085	0.375	0.750	11.542	1.000
7411	Top-16 log	0.016	0.203	0.500	8.039	1.000
7523	Local power	0.005	0.234	0.438	1.583	0.938
7523	Corrected power	-0.011	0.250	0.562	11.046	1.000
7523	Top-16 log	0.004	0.172	0.375	8.352	1.000
7643	Local power	0.009	0.234	0.375	1.710	0.945
7643	Corrected power	0.001	0.219	0.375	11.817	1.000
7643	Top-16 log	0.005	0.234	0.375	8.176	1.000
7753	Local power	0.012	0.312	0.438	1.512	0.914
7753	Corrected power	-0.040	0.344	0.688	11.826	1.000
7753	Top-16 log	0.011	0.297	0.438	7.803	1.000
7867	Local power	0.009	0.234	0.500	1.463	0.922
7867	Corrected power	0.006	0.141	0.312	10.354	1.000
7867	Top-16 log	0.003	0.203	0.312	7.732	1.000
7979	Local power	0.007	0.172	0.250	1.607	0.914
7979	Corrected power	-0.056	0.312	0.625	11.827	1.000
7979	Top-16 log	0.003	0.156	0.250	7.972	1.000

Table 58: Paired method contrasts on ARC. Cells are seed means with 95% bootstrap intervals.

Contrast	Δ eval gap	Accuracy	Pass@k	Grad. norm
CL−LL	-0.035 [-0.052, -0.017]	0.057 [0.006, 0.109]	0.148 [0.070, 0.234]	140.282 [132.483, 148.196]
MC−LL	-0.031 [-0.048, -0.015]	0.051 [0.010, 0.096]	0.141 [0.062, 0.219]	102.042 [96.580, 106.762]
LP−LL	0.002 [4.01×10^{-4}, 0.003]	-0.018 [-0.047, 0.012]	-0.023 [-0.094, 0.047]	-19.758 [-20.359, -19.106]
CP−LP	-0.038 [-0.058, -0.019]	0.025 [-0.039, 0.088]	0.133 [0.008, 0.250]	9.700 [9.358, 10.035]

Continued on next page

Table 58 (continued)

Contrast	Δ eval gap	Accuracy	Pass@k	Grad. norm
TK−LL	-1.60×10^{-4} [-0.002, 0.003]	-0.039 [-0.061, -0.016]	-0.055 [-0.102, -0.008]	-13.342 [-13.892, -12.723]

Table 59: Task outcomes relative to the restored initial adapter on ARC.

Method	Δ accuracy	Δ pass@k	Δ tokens
Local log	-0.004 [-0.027, 0.020]	-0.031 [-0.078, 0.016]	0.320 [-0.049, 0.822]
Corrected log	0.053 [-0.020, 0.123]	0.117 [0.023, 0.211]	-6.166 [-9.412, -3.127]
Sequence log MC	0.047 [-0.010, 0.109]	0.109 [0.031, 0.195]	-5.156 [-8.094, -2.594]
Local power	-0.021 [-0.051, 0.010]	-0.055 [-0.133, 0.016]	0.373 [0.023, 0.773]
Corrected power	0.004 [-0.062, 0.074]	0.078 [0, 0.164]	-5.232 [-8.367, -2.512]
Top-16 log	-0.043 [-0.082, -0.006]	-0.086 [-0.148, -0.031]	0.645 [0.080, 1.356]

Table 60: Per-seed GSM8K training outcomes, method group 1.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
7201	Local log	0.014	0.047	0.125	14.397	1.000
7201	Corrected log	-2.25×10^{-4}	0.047	0.062	74.460	1.000
7201	Sequence log MC	0.008	0.016	0.062	56.237	1.000
7307	Local log	0.013	0.078	0.188	13.923	1.000
7307	Corrected log	-0.016	0.172	0.438	65.991	1.000
7307	Sequence log MC	-0.003	0.078	0.125	49.218	1.000
7411	Local log	0.008	0.031	0.062	18.130	1.000
7411	Corrected log	-0.044	0.062	0.250	79.497	1.000
7411	Sequence log MC	-0.015	0.031	0.125	62.142	1.000
7523	Local log	0.009	0.078	0.125	12.908	1.000
7523	Corrected log	-0.042	0.156	0.375	81.628	1.000
7523	Sequence log MC	-0.020	0.156	0.375	54.273	1.000
7643	Local log	0.024	0.062	0.125	14.607	1.000
7643	Corrected log	0.012	0.062	0.188	71.047	1.000
7643	Sequence log MC	0.018	0.031	0.125	53.434	1.000
7753	Local log	0.009	0.078	0.125	14.780	1.000
7753	Corrected log	-0.025	0.125	0.312	72.663	1.000
7753	Sequence log MC	-0.011	0.094	0.312	50.579	1.000
7867	Local log	0.019	0.109	0.125	14.229	1.000
7867	Corrected log	0.009	0.016	0.062	68.173	1.000
7867	Sequence log MC	0.011	0.016	0.062	52.509	1.000
7979	Local log	0.016	0.078	0.125	14.081	1.000
7979	Corrected log	-0.007	0.047	0.125	64.516	1.000
7979	Sequence log MC	-0.002	0.031	0.062	53.461	1.000

Table 61: Per-seed GSM8K training outcomes, method group 2.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
7201	Local power	0.016	0.047	0.125	1.406	0.867
7201	Corrected power	0.002	0.047	0.125	7.283	1.000
7201	Top-16 log	0.003	0.031	0.125	7.683	1.000
7307	Local power	0.019	0.094	0.188	1.317	0.750
7307	Corrected power	-0.006	0.109	0.188	6.006	0.992
7307	Top-16 log	0.006	0.062	0.125	7.600	1.000
7411	Local power	0.012	0.031	0.062	1.499	0.805
7411	Corrected power	-0.038	0.047	0.188	6.581	0.992
7411	Top-16 log	0.003	0.078	0.125	9.899	1.000
7523	Local power	0.012	0.031	0.125	1.254	0.695
7523	Corrected power	-0.033	0.125	0.438	6.263	0.992

Continued on next page

Table 61 (continued)

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
7523	Top-16 log	0.006	0.062	0.125	6.930	1.000
7643	Local power	0.028	0.047	0.125	1.439	0.844
7643	Corrected power	0.011	0.016	0.062	6.603	1.000
7643	Top-16 log	0.016	0.062	0.125	8.014	1.000
7753	Local power	0.011	0.078	0.125	1.395	0.836
7753	Corrected power	-0.018	0.094	0.250	6.601	1.000
7753	Top-16 log	-0.003	0.047	0.062	7.932	1.000
7867	Local power	0.021	0.062	0.125	1.308	0.781
7867	Corrected power	0.009	0	0	6.153	1.000
7867	Top-16 log	0.014	0.094	0.125	7.789	1.000
7979	Local power	0.017	0.062	0.125	1.366	0.805
7979	Corrected power	-0.008	0.047	0.125	5.818	1.000
7979	Top-16 log	0.006	0.094	0.125	7.931	1.000

Table 62: Paired method contrasts on GSM8K. Cells are seed means with 95% bootstrap intervals.

Contrast	Δ eval gap	Accuracy	Pass@k	Grad. norm
CL−LL	-0.028 [-0.039, -0.018]	0.016 [-0.027, 0.053]	0.102 [0.016, 0.188]	57.615 [54.095, 61.670]
MC−LL	-0.016 [-0.021, -0.010]	-0.014 [-0.047, 0.020]	0.031 [-0.047, 0.117]	39.350 [37.433, 41.295]
LP−LL	0.003 [0.002, 0.004]	-0.014 [-0.029, 0]	0 [0, 0]	-13.259 [-14.324, -12.487]
CP−LP	-0.027 [-0.037, -0.019]	0.004 [-0.023, 0.035]	0.047 [-0.031, 0.141]	5.041 [4.786, 5.329]
TK−LL	-0.007 [-0.010, -0.005]	-0.004 [-0.018, 0.014]	-0.008 [-0.031, 0.016]	-6.660 [-7.158, -6.286]

Table 63: Task outcomes relative to the restored initial adapter on GSM8K.

Method	Δ accuracy	Δ pass@k	Δ tokens
Local log	0.016 [0.002, 0.031]	0.016 [-0.016, 0.039]	0.008 [0, 0.023]
Corrected log	0.031 [-0.008, 0.068]	0.117 [0.031, 0.203]	-4.301 [-11.018, -0.125]
Sequence log MC	0.002 [-0.031, 0.035]	0.047 [-0.016, 0.109]	-1.051 [-2.641, -0.092]
Local power	0.002 [-0.016, 0.020]	0.016 [-0.016, 0.039]	0.008 [0, 0.023]
Corrected power	0.006 [-0.029, 0.035]	0.062 [-0.008, 0.133]	-3.584 [-9.836, -0.195]
Top-16 log	0.012 [-0.012, 0.039]	0.008 [-0.023, 0.039]	0.008 [0, 0.023]

K.11 HELLASWAG TRAINING OUTCOMES

Table 64: Per-seed HELLASWAG training outcomes, method group 1.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
10103	Local log	0.026	0.422	0.688	21.665	1.000
10103	Corrected log	-0.033	0.188	0.562	152.992	1.000
10103	Sequence log MC	-0.034	0.203	0.500	119.007	1.000
10211	Local log	0.025	0.312	0.562	21.423	1.000
10211	Corrected log	-0.009	0.250	0.625	150.405	1.000
10211	Sequence log MC	-0.011	0.266	0.688	121.928	1.000
10321	Local log	0.036	0.312	0.625	22.817	1.000
10321	Corrected log	-0.034	0.297	0.625	163.437	1.000
10321	Sequence log MC	-0.025	0.219	0.500	143.966	1.000
10433	Local log	0.035	0.391	0.625	22.969	1.000
10433	Corrected log	0.006	0.312	0.688	168.803	1.000
10433	Sequence log MC	0.013	0.297	0.625	136.794	1.000
10501	Local log	0.029	0.297	0.500	22.152	1.000
10501	Corrected log	-0.008	0.250	0.500	142.794	1.000
10501	Sequence log MC	-0.003	0.250	0.562	111.370	1.000
10613	Local log	0.030	0.266	0.562	23.360	1.000
10613	Corrected log	-0.014	0.203	0.500	162.582	1.000

Continued on next page

Table 64 (continued)

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
10613	Sequence log MC	-0.013	0.203	0.500	132.573	1.000
10723	Local log	0.023	0.281	0.438	22.252	1.000
10723	Corrected log	-0.018	0.234	0.500	165.017	1.000
10723	Sequence log MC	-0.026	0.172	0.438	128.075	1.000
10831	Local log	0.021	0.375	0.750	23.188	1.000
10831	Corrected log	-0.028	0.312	0.625	172.366	1.000
10831	Sequence log MC	-0.020	0.281	0.688	138.714	1.000
10909	Local log	0.021	0.344	0.625	22.144	1.000
10909	Corrected log	-0.019	0.219	0.500	162.150	1.000
10909	Sequence log MC	-0.021	0.250	0.562	143.217	1.000
11027	Local log	0.021	0.406	0.750	22.032	1.000
11027	Corrected log	-0.026	0.328	0.625	162.234	1.000
11027	Sequence log MC	-0.027	0.266	0.500	126.725	1.000
11131	Local log	0.027	0.281	0.625	21.563	1.000
11131	Corrected log	-0.008	0.328	0.625	153.361	1.000
11131	Sequence log MC	-0.014	0.219	0.562	125.890	1.000
11243	Local log	0.034	0.250	0.562	23.364	1.000
11243	Corrected log	-0.020	0.203	0.375	153.978	1.000
11243	Sequence log MC	-0.025	0.172	0.312	121.048	1.000
11353	Local log	0.034	0.281	0.562	22.097	1.000
11353	Corrected log	-0.012	0.281	0.688	168.106	1.000
11353	Sequence log MC	-0.007	0.234	0.625	143.949	1.000
11467	Local log	0.020	0.406	0.812	25.015	1.000
11467	Corrected log	-0.036	0.203	0.438	159.552	1.000
11467	Sequence log MC	-0.043	0.297	0.688	136.011	1.000
11579	Local log	0.029	0.344	0.750	23.505	1.000
11579	Corrected log	-0.046	0.141	0.438	167.335	1.000
11579	Sequence log MC	-0.028	0.234	0.625	129.415	1.000
11689	Local log	0.024	0.344	0.750	21.613	1.000
11689	Corrected log	-0.021	0.266	0.688	172.710	1.000
11689	Sequence log MC	-0.026	0.281	0.750	136.315	1.000

Table 65: Paired method contrasts on HELLASWAG. Cells are seed means with 95% bootstrap intervals.

Contrast	Δ eval gap	Accuracy	Pass@k	Grad. norm
CL−LL	-0.048 [-0.054, -0.042]	-0.081 [-0.118, -0.047]	-0.074 [-0.141, -0.012]	138.542 [134.460, 142.366]
MC−LL	-0.047 [-0.052, -0.041]	-0.092 [-0.114, -0.073]	-0.066 [-0.117, -0.016]	108.365 [103.776, 112.763]
LP−LL	0.006 [0.004, 0.007]	-0.005 [-0.032, 0.022]	0 [-0.051, 0.051]	-20.857 [-21.340, -20.421]
CP−LP	-0.053 [-0.059, -0.046]	-0.084 [-0.110, -0.059]	-0.074 [-0.117, -0.031]	10.480 [10.217, 10.733]
TK−LL	0.016 [0.012, 0.019]	0.007 [-0.027, 0.039]	0.008 [-0.051, 0.062]	-14.254 [-14.714, -13.854]

Table 66: Per-seed HELLASWAG training outcomes, method group 2.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
10103	Local power	0.034	0.297	0.688	1.763	0.961
10103	Corrected power	-0.026	0.250	0.562	11.413	0.969
10103	Top-16 log	0.055	0.266	0.500	8.178	1.000
10211	Local power	0.031	0.344	0.688	1.732	0.930
10211	Corrected power	-0.012	0.266	0.625	11.626	0.969
10211	Top-16 log	0.035	0.328	0.625	8.293	1.000
10321	Local power	0.036	0.250	0.500	1.739	0.977
10321	Corrected power	-0.030	0.219	0.500	12.680	0.977

Continued on next page

Table 66 (continued)

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
10321	Top-16 log	0.049	0.375	0.688	8.466	1.000
10433	Local power	0.040	0.391	0.625	1.703	0.953
10433	Corrected power	0.013	0.234	0.562	12.645	1.000
10433	Top-16 log	0.034	0.359	0.688	8.510	1.000
10501	Local power	0.032	0.281	0.625	1.670	0.938
10501	Corrected power	-0.005	0.297	0.562	11.207	0.977
10501	Top-16 log	0.041	0.438	0.688	8.197	1.000
10613	Local power	0.037	0.266	0.500	1.726	0.945
10613	Corrected power	-0.013	0.219	0.562	12.033	0.977
10613	Top-16 log	0.039	0.312	0.562	8.296	1.000
10723	Local power	0.031	0.328	0.562	1.654	0.961
10723	Corrected power	-0.020	0.250	0.562	11.796	0.992
10723	Top-16 log	0.038	0.266	0.500	8.175	1.000
10831	Local power	0.027	0.312	0.625	1.700	0.969
10831	Corrected power	-0.029	0.250	0.562	12.772	0.969
10831	Top-16 log	0.040	0.297	0.562	8.127	1.000
10909	Local power	0.027	0.359	0.750	1.652	0.922
10909	Corrected power	-0.012	0.266	0.562	12.649	0.984
10909	Top-16 log	0.035	0.359	0.750	8.285	1.000
11027	Local power	0.028	0.344	0.625	1.747	0.961
11027	Corrected power	-0.038	0.250	0.562	12.389	1.000
11027	Top-16 log	0.041	0.328	0.625	8.098	1.000
11131	Local power	0.032	0.328	0.688	1.661	0.961
11131	Corrected power	-0.010	0.234	0.438	11.933	0.992
11131	Top-16 log	0.056	0.312	0.750	8.036	1.000
11243	Local power	0.038	0.359	0.688	1.694	0.922
11243	Corrected power	-0.030	0.219	0.500	11.505	0.984
11243	Top-16 log	0.049	0.266	0.625	8.511	1.000
11353	Local power	0.042	0.312	0.625	1.813	0.984
11353	Corrected power	-0.011	0.266	0.625	12.432	0.984
11353	Top-16 log	0.052	0.344	0.688	8.511	1.000
11467	Local power	0.025	0.359	0.688	1.683	0.961
11467	Corrected power	-0.032	0.219	0.562	13.081	0.984
11467	Top-16 log	0.039	0.406	0.812	8.329	1.000
11579	Local power	0.034	0.391	0.688	1.741	0.945
11579	Corrected power	-0.041	0.188	0.562	12.359	0.977
11579	Top-16 log	0.035	0.375	0.625	8.671	1.000
11689	Local power	0.034	0.312	0.625	1.770	0.930
11689	Corrected power	-0.023	0.266	0.688	12.603	0.984
11689	Top-16 log	0.047	0.391	0.625	8.411	1.000

Table 67: Task outcomes relative to the restored initial adapter on HELLASWAG.

Method	Δ accuracy	Δ pass@k	Δ tokens
Local log	0.033 [-0.004, 0.068]	0.074 [0.016, 0.133]	-0.435 [-0.772, -0.161]
Corrected log	-0.048 [-0.085, -0.012]	0 [-0.066, 0.074]	-14.074 [-16.880, -11.668]
Sequence log MC	-0.059 [-0.086, -0.029]	0.008 [-0.062, 0.078]	-12.974 [-16.142, -10.127]
Local power	0.028 [-0.008, 0.062]	0.074 [0.012, 0.133]	0.030 [-0.367, 0.352]
Corrected power	-0.056 [-0.084, -0.031]	0 [-0.047, 0.051]	-16.979 [-20.399, -13.975]
Top-16 log	0.040 [0.012, 0.069]	0.082 [0.012, 0.148]	0.438 [0.236, 0.670]

K.12 Model-Scale Training Outcomes

Table 68: Per-seed 4B-to-9B ARC training outcomes, method group 1.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
9643	Local log	0.031	0.156	0.250	7.149	0.984
9643	Corrected log	0.006	0.250	0.375	29.493	1.000
9643	Sequence log MC	0.010	0.250	0.500	28.797	0.992
9751	Local log	0.015	0.188	0.375	6.746	1.000
9751	Corrected log	0.003	0.250	0.375	30.626	1.000
9751	Sequence log MC	0.002	0.188	0.375	27.917	1.000
9863	Local log	0.033	0.125	0.125	7.364	1.000
9863	Corrected log	0.011	0.219	0.375	29.094	1.000
9863	Sequence log MC	0.013	0.188	0.375	26.679	1.000
9973	Local log	0.028	0.219	0.375	7.912	1.000
9973	Corrected log	0.013	0.156	0.375	32.680	1.000
9973	Sequence log MC	0.016	0.125	0.250	29.120	1.000

Table 69: Per-seed 4B-to-9B ARC training outcomes, method group 2.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
9643	Local power	0.036	0.156	0.250	0.980	0.469
9643	Corrected power	0.009	0.250	0.500	4.149	0.977
9643	Top-16 log	0.044	0.156	0.250	4.270	1.000
9751	Local power	0.022	0.250	0.375	1.050	0.438
9751	Corrected power	0.009	0.188	0.375	4.431	0.992
9751	Top-16 log	0.024	0.281	0.375	4.427	1.000
9863	Local power	0.033	0.219	0.375	1.057	0.484
9863	Corrected power	0.017	0.188	0.375	4.384	0.984
9863	Top-16 log	0.036	0.156	0.250	4.480	1.000
9973	Local power	0.031	0.188	0.375	1.115	0.539
9973	Corrected power	0.017	0.125	0.250	4.560	0.984
9973	Top-16 log	0.034	0.250	0.250	4.528	1.000

Table 70: Paired method contrasts on 4B-to-9B ARC. Cells are seed means with 95% bootstrap intervals.

Contrast	Δ eval gap	Accuracy	Pass@k	Grad. norm
CL−LL	-0.018 [-0.023, -0.013]	0.047 [-0.023, 0.094]	0.094 [0, 0.188]	23.180 [22.037, 24.324]
MC−LL	-0.016 [-0.020, -0.013]	0.016 [-0.055, 0.078]	0.094 [-0.062, 0.250]	20.836 [19.789, 21.529]
LP−LL	0.004 [0.001, 0.006]	0.031 [-0.016, 0.078]	0.062 [0, 0.188]	-6.242 [-6.640, -5.849]
CP−LP	-0.017 [-0.024, -0.014]	-0.016 [-0.062, 0.055]	0.031 [-0.094, 0.188]	3.330 [3.221, 3.416]
TK−LL	0.008 [0.005, 0.012]	0.039 [0.008, 0.078]	0 [-0.094, 0.094]	-2.866 [-3.258, -2.460]

Table 71: Task outcomes relative to the restored initial adapter on 4B-to-9B ARC.

Method	Δ accuracy	Δ pass@k	Δ tokens
Local log	-0.039 [-0.094, 0.016]	-0.125 [-0.250, 0]	0.680 [-0.094, 1.648]
Corrected log	0.008 [-0.023, 0.047]	-0.031 [-0.094, 0]	-0.391 [-0.836, -0.109]
Sequence log MC	-0.023 [-0.047, 0]	-0.031 [-0.094, 0]	0.055 [-0.188, 0.305]
Local power	-0.008 [-0.070, 0.047]	-0.062 [-0.188, 0]	0.477 [-0.266, 1.227]
Corrected power	-0.023 [-0.047, 0]	-0.031 [-0.094, 0]	0.344 [-0.656, 1.594]
Top-16 log	0 [-0.078, 0.078]	-0.125 [-0.219, -0.031]	0.812 [0.281, 1.453]

Table 72: Per-seed 4B-to-9B GSM8K training outcomes, method group 1.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
9203	Local log	0.009	0.031	0.125	3.531	0.906

Continued on next page

Table 72 (continued)

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
9203	Corrected log	0.006	0.062	0.125	9.136	1.000
9203	Sequence log MC	0.008	0.062	0.125	8.337	0.945
9311	Local log	0.011	0.094	0.125	4.244	0.922
9311	Corrected log	0.010	0.094	0.125	11.346	1.000
9311	Sequence log MC	0.011	0.094	0.125	9.861	0.945
9421	Local log	0.011	0.031	0.125	3.896	0.906
9421	Corrected log	0.012	0.031	0.125	10.164	1.000
9421	Sequence log MC	0.008	0.031	0.125	9.881	0.969
9533	Local log	0.011	0.031	0.125	4.160	0.914
9533	Corrected log	0.010	0.031	0.125	9.473	1.000
9533	Sequence log MC	0.008	0.062	0.125	8.438	0.945

Table 73: Per-seed 4B-to-9B GSM8K training outcomes, method group 2.

Seed	Method	Δ eval gap	Accuracy	Pass@k	Grad. norm	Clip fraction
9203	Local power	0.009	0.031	0.125	0.709	0.148
9203	Corrected power	0.008	0.031	0.125	1.558	0.625
9203	Top-16 log	0.012	0.031	0.125	2.854	1.000
9311	Local power	0.010	0.094	0.125	0.714	0.148
9311	Corrected power	0.011	0.094	0.125	1.846	0.742
9311	Top-16 log	0.014	0.094	0.125	3.175	0.992
9421	Local power	0.011	0.031	0.125	0.702	0.164
9421	Corrected power	0.011	0.031	0.125	1.675	0.672
9421	Top-16 log	0.013	0.031	0.125	2.920	0.984
9533	Local power	0.009	0.031	0.125	0.718	0.070
9533	Corrected power	0.010	0.031	0.125	1.505	0.742
9533	Top-16 log	0.014	0.062	0.125	2.874	0.984

Table 74: Paired method contrasts on 4B-to-9B GSM8K. Cells are seed means with 95% bootstrap intervals.

Contrast	Δ eval gap	Accuracy	Pass@k	Grad. norm
CL−LL	-0.001 [-0.003, 0.001]	0.008 [0, 0.023]	0 [0, 0]	6.072 [5.459, 6.728]
MC−LL	-0.002 [-0.003, -3.10×10^{-4}]	0.016 [0, 0.031]	0 [0, 0]	5.172 [4.542, 5.801]
LP−LL	-0.001 [-0.001, -1.89×10^{-4}]	0 [0, 0]	0 [0, 0]	-3.247 [-3.486, -2.977]
CP−LP	2.86×10^{-4} [-4.52×10^{-4}, 0.001]	0 [0, 0]	0 [0, 0]	0.935 [0.818, 1.061]
TK−LL	0.003 [0.002, 0.003]	0.008 [0, 0.023]	0 [0, 0]	-1.002 [-1.208, -0.775]

Table 75: Task outcomes relative to the restored initial adapter on 4B-to-9B GSM8K.

Method	Δ accuracy	Δ pass@k	Δ tokens
Local log	0.008 [0, 0.023]	0.031 [0, 0.094]	0 [0, 0]
Corrected log	0.016 [0, 0.047]	0.031 [0, 0.094]	0 [0, 0]
Sequence log MC	0.023 [0, 0.047]	0.031 [0, 0.094]	0 [0, 0]
Local power	0.008 [0, 0.023]	0.031 [0, 0.094]	0 [0, 0]
Corrected power	0.008 [0, 0.023]	0.031 [0, 0.094]	0 [0, 0]
Top-16 log	0.016 [0, 0.031]	0.031 [0, 0.094]	0 [0, 0]

K.13 Fresh Sequence Evaluation

Table 76: Fresh on-policy sequence objectives after ARC training.

Method	Sequence KL	Log ratio	Power	Top-K	Tokens
Local log	61.676 [60.437, 62.871]	45.067 [43.815, 46.597]	4.860 [4.713, 4.994]	42.826 [42.005, 43.651]	63.219 [62.947, 63.521]

Continued on next page

Table 76 (continued)

Method	Sequence KL	Log ratio	Power	Top-K	Tokens
Corrected log	61.675 [59.064, 63.823]	43.837 [41.771, 45.667]	4.510 [4.274, 4.754]	40.328 [38.408, 42.150]	57.783 [54.900, 60.537]
Sequence log MC	62.009 [59.548, 64.119]	44.944 [42.429, 47.058]	4.506 [4.304, 4.712]	40.742 [39.000, 42.376]	58.533 [55.580, 61.143]
Local power	61.565 [60.702, 62.365]	45.664 [44.059, 47.354]	4.841 [4.739, 4.952]	42.812 [42.226, 43.457]	63.393 [62.934, 63.744]
Corrected power	61.811 [59.353, 63.803]	44.157 [41.573, 46.201]	4.501 [4.280, 4.716]	40.413 [38.724, 42.066]	57.404 [54.412, 60.260]
Top-16 log	59.272 [58.558, 59.973]	44.985 [43.527, 46.526]	4.940 [4.870, 5.035]	42.454 [41.944, 43.072]	63.576 [63.393, 63.748]

Table 77: Paired fresh-sequence contrasts after ARC training.

Contrast	Δ seq. KL	Δ log ratio	Δ power	Δ top-K
CL−LL	-0.001 [-2.211, 1.916]	-1.230 [-3.101, 0.561]	-0.350 [-0.486, -0.197]	-2.498 [-3.864, -1.199]
MC−LL	0.333 [-1.720, 2.318]	-0.124 [-2.405, 2.088]	-0.353 [-0.481, -0.217]	-2.085 [-3.395, -0.857]
LP−LL	-0.111 [-0.832, 0.558]	0.597 [-0.310, 1.591]	-0.018 [-0.116, 0.073]	-0.014 [-0.554, 0.487]
CP−LP	0.246 [-2.031, 2.213]	-1.507 [-4.248, 1.046]	-0.341 [-0.492, -0.206]	-2.399 [-3.955, -1.095]
TK−LL	-2.404 [-3.596, -1.188]	-0.082 [-1.159, 0.925]	0.080 [-0.038, 0.203]	-0.372 [-1.160, 0.309]

Table 78: Fresh on-policy sequence objectives after GSM8K training.

Method	Sequence KL	Log ratio	Power	Top-K	Tokens
Local log	30.861 [29.837, 31.828]	23.214 [21.877, 24.388]	3.470 [3.379, 3.545]	26.639 [25.829, 27.364]	63.965 [63.895, 64.000]
Corrected log	30.638 [27.848, 32.625]	22.272 [20.486, 23.860]	3.165 [2.799, 3.431]	25.620 [22.958, 27.451]	59.164 [52.051, 63.826]
Sequence log MC	31.494 [30.429, 32.540]	22.818 [21.995, 23.631]	3.280 [3.133, 3.412]	26.515 [25.620, 27.381]	62.047 [58.996, 63.918]
Local power	30.549 [29.685, 31.336]	23.463 [22.694, 24.262]	3.442 [3.332, 3.550]	26.364 [25.638, 27.036]	63.965 [63.895, 64.000]
Corrected power	31.559 [28.965, 33.435]	22.377 [20.694, 23.949]	3.251 [2.876, 3.522]	26.392 [23.880, 28.146]	59.902 [53.262, 63.818]
Top-16 log	28.900 [28.328, 29.469]	24.108 [23.470, 24.739]	3.629 [3.569, 3.698]	25.641 [25.213, 26.086]	64.000 [64.000, 64.000]

Table 79: Paired fresh-sequence contrasts after GSM8K training.

Contrast	Δ seq. KL	Δ log ratio	Δ power	Δ top-K
CL−LL	-0.223 [-3.320, 1.801]	-0.942 [-2.897, 0.819]	-0.305 [-0.664, -0.075]	-1.019 [-3.870, 0.806]
MC−LL	0.633 [-0.393, 1.517]	-0.395 [-1.332, 0.590]	-0.191 [-0.312, -0.098]	-0.124 [-1.176, 0.689]
LP−LL	-0.311 [-0.813, 0.188]	0.250 [-0.297, 0.881]	-0.029 [-0.076, 0.026]	-0.275 [-0.690, 0.150]
CP−LP	1.010 [-1.719, 2.760]	-1.086 [-2.800, 0.434]	-0.191 [-0.528, 0.027]	0.029 [-2.595, 1.711]
TK−LL	-1.961 [-2.573, -1.387]	0.894 [-0.111, 1.968]	0.159 [0.094, 0.229]	-0.998 [-1.551, -0.479]

K.14 HELLASWAG FRESH SEQUENCE EVALUATION

Table 80: Fresh on-policy sequence objectives after HELLASWAG training.

Method	Sequence KL	Log ratio	Power	Top-K	Tokens
Local log	63.799 [63.213, 64.377]	49.089 [48.455, 49.705]	5.210 [5.158, 5.264]	44.878 [44.390, 45.379]	63.326 [63.036, 63.596]
Corrected log	43.736 [39.204, 47.901]	33.689 [30.320, 36.870]	3.568 [3.254, 3.842]	30.108 [27.221, 32.721]	49.597 [46.994, 52.004]
Sequence log MC	45.627 [40.756, 50.291]	35.486 [32.017, 38.844]	3.694 [3.340, 4.029]	31.284 [28.176, 34.271]	50.587 [47.866, 53.122]
Local power	63.888 [63.419, 64.362]	49.084 [48.439, 49.716]	5.259 [5.212, 5.310]	45.125 [44.759, 45.497]	63.623 [63.403, 63.815]

Continued on next page

Table 80 (continued)

Method	Sequence KL	Log ratio	Power	Top-K	Tokens
Corrected power	42.011 [37.051, 46.623]	32.800 [29.081, 36.313]	3.379 [3.039, 3.700]	28.835 [25.651, 31.757]	46.854 [43.603, 49.870]
Top-16 log	61.133 [60.582, 61.662]	49.299 [48.607, 49.954]	5.370 [5.321, 5.420]	44.629 [44.196, 45.051]	63.869 [63.792, 63.941]

Table 81: Paired fresh-sequence contrasts after HELLASWAG training.

Contrast	Δ seq. KL	Δ log ratio	Δ power	Δ top-K
CL−LL	-20.063 [-24.452, -16.023]	-15.400 [-18.584, -12.352]	-1.642 [-1.942, -1.384]	-14.770 [-17.608, -12.299]
MC−LL	-18.172 [-23.010, -13.611]	-13.603 [-16.980, -10.191]	-1.517 [-1.847, -1.205]	-13.593 [-16.494, -10.825]
LP−LL	0.089 [-0.479, 0.739]	-0.005 [-0.743, 0.783]	0.049 [0.003, 0.096]	0.248 [-0.155, 0.639]
CP−LP	-21.877 [-26.763, -17.348]	-16.283 [-20.025, -12.609]	-1.880 [-2.223, -1.561]	-16.290 [-19.359, -13.330]
TK−LL	-2.666 [-3.338, -2.013]	0.210 [-0.481, 0.858]	0.160 [0.104, 0.214]	-0.248 [-0.840, 0.288]

K.15 Model-Scale Fresh Sequence Evaluation

Table 82: Fresh on-policy sequence objectives after 4B-to-9B ARC training.

Method	Sequence KL	Log ratio	Power	Top-K	Tokens
Local log	20.464 [19.988, 20.912]	18.230 [17.871, 18.814]	3.306 [3.197, 3.416]	19.097 [18.639, 19.519]	63.797 [63.656, 63.938]
Corrected log	21.574 [21.016, 21.969]	19.208 [18.018, 20.334]	3.390 [3.300, 3.489]	20.090 [19.585, 20.500]	63.219 [62.297, 63.844]
Sequence log MC	21.450 [20.801, 22.098]	18.931 [17.816, 20.053]	3.387 [3.274, 3.500]	19.937 [19.326, 20.548]	62.797 [62.109, 63.484]
Local power	20.529 [20.163, 20.958]	18.018 [17.383, 18.773]	3.318 [3.244, 3.392]	19.168 [18.838, 19.628]	63.773 [63.562, 63.984]
Corrected power	21.053 [20.730, 21.376]	18.703 [18.062, 19.056]	3.317 [3.265, 3.361]	19.572 [19.337, 19.807]	63.211 [62.625, 63.781]
Top-16 log	20.246 [19.884, 20.556]	18.051 [16.673, 19.286]	3.362 [3.246, 3.450]	19.041 [18.738, 19.208]	64.000 [64.000, 64.000]

Table 83: Paired fresh-sequence contrasts after 4B-to-9B ARC training.

Contrast	Δ seq. KL	Δ log ratio	Δ power	Δ top-K
CL−LL	1.110 [0.571, 1.648]	0.978 [-0.118, 2.084]	0.084 [0.002, 0.159]	0.993 [0.484, 1.502]
MC−LL	0.986 [0.708, 1.264]	0.701 [-0.077, 1.479]	0.081 [0.034, 0.131]	0.840 [0.590, 1.124]
LP−LL	0.065 [-0.243, 0.284]	-0.212 [-0.484, -0.045]	0.012 [-0.024, 0.047]	0.071 [-0.232, 0.302]
CP−LP	0.524 [0.421, 0.620]	0.686 [0.210, 1.135]	-0.001 [-0.054, 0.063]	0.404 [0.173, 0.581]
TK−LL	-0.218 [-0.786, 0.349]	-0.179 [-1.629, 0.993]	0.056 [-0.004, 0.137]	-0.056 [-0.626, 0.513]

Table 84: Fresh on-policy sequence objectives after 4B-to-9B GSM8K training.

Method	Sequence KL	Log ratio	Power	Top-K	Tokens
Local log	5.094 [4.899, 5.268]	4.297 [4.087, 4.508]	0.975 [0.935, 1.015]	5.035 [4.850, 5.206]	64.000 [64.000, 64.000]
Corrected log	5.424 [5.118, 5.699]	4.632 [4.446, 4.819]	1.025 [0.934, 1.092]	5.350 [5.043, 5.620]	64.000 [64.000, 64.000]
Sequence log MC	5.286 [5.073, 5.601]	4.483 [4.048, 4.918]	0.998 [0.936, 1.068]	5.218 [5.011, 5.528]	64.000 [64.000, 64.000]
Local power	5.185 [4.940, 5.509]	4.433 [4.018, 4.936]	0.996 [0.945, 1.072]	5.124 [4.890, 5.439]	64.000 [64.000, 64.000]

Continued on next page

Table 84 (continued)

Method	Sequence KL	Log ratio	Power	Top-K	Tokens
Corrected power	5.322 [5.030, 5.615]	4.466 [4.184, 4.747]	1.013 [0.945, 1.081]	5.254 [4.967, 5.541]	64.000 [64.000, 64.000]
Top-16 log	4.883 [4.679, 5.190]	4.184 [3.851, 4.609]	0.985 [0.903, 1.075]	4.816 [4.611, 5.107]	64.000 [64.000, 64.000]

Table 85: Paired fresh-sequence contrasts after 4B-to-9B GSM8K training.

Contrast	Δ seq. KL	Δ log ratio	Δ power	Δ top-K
CL−LL	0.330 [0.074, 0.551]	0.335 [0.088, 0.578]	0.049 [-0.009, 0.086]	0.315 [0.061, 0.533]
MC−LL	0.192 [0.041, 0.360]	0.185 [-0.039, 0.419]	0.022 [-0.005, 0.066]	0.182 [0.028, 0.354]
LP−LL	0.091 [-0.036, 0.268]	0.136 [-0.160, 0.432]	0.021 [-0.020, 0.070]	0.089 [-0.039, 0.265]
CP−LP	0.137 [0.045, 0.281]	0.032 [-0.265, 0.330]	0.016 [-0.021, 0.068]	0.130 [0.039, 0.274]
TK−LL	-0.211 [-0.362, -0.051]	-0.114 [-0.290, 0.110]	0.009 [-0.044, 0.062]	-0.219 [-0.381, -0.058]

K.16 Trained-Endpoint Geometry

Table 86: Local log-field geometry at trained ARC endpoints.

Training method	$\widetilde{\Phi}_8$	$\widetilde{\Phi}_{16}$	$\Delta\widetilde{\Phi}_{16}$	$\|c\|/\|u\|_{16}$	$\|C\|/\|u\|_{16}$
Corrected log	0.075 [0.011, 0.148]	-0.021 [-0.054, 0.012]	-0.046 [-0.066, -0.027]	6.352 [5.647, 7.129]	0.074 [0.065, 0.083]
Corrected power	0.065 [0.003, 0.135]	-0.023 [-0.055, 0.011]	-0.047 [-0.069, -0.026]	6.656 [5.644, 7.920]	0.076 [0.065, 0.087]
Local log	0.191 [0.130, 0.250]	0.032 [0.006, 0.058]	0.008 [4.63×10^{-4}, 0.016]	5.616 [5.031, 6.207]	0.077 [0.068, 0.085]
Local power	0.173 [0.125, 0.220]	0.041 [0.016, 0.063]	0.016 [0.003, 0.031]	5.555 [5.009, 6.074]	0.077 [0.069, 0.085]
Sequence log MC	0.075 [0.015, 0.145]	-0.015 [-0.048, 0.019]	-0.039 [-0.058, -0.021]	6.371 [5.513, 7.429]	0.075 [0.065, 0.085]
Top-16 log	0.178 [0.131, 0.227]	0.044 [0.022, 0.065]	0.020 [0.008, 0.034]	5.711 [5.070, 6.340]	0.085 [0.073, 0.094]

Table 87: Local log-field geometry at trained GSM8K endpoints.

Training method	$\widetilde{\Phi}_8$	$\widetilde{\Phi}_{16}$	$\Delta\widetilde{\Phi}_{16}$	$\|c\|/\|u\|_{16}$	$\|C\|/\|u\|_{16}$
Corrected log	0.314 [0.210, 0.413]	-0.031 [-0.104, 0.070]	-0.170 [-0.245, -0.087]	4.685 [3.688, 5.878]	0.080 [0.065, 0.097]
Corrected power	0.315 [0.225, 0.406]	-0.039 [-0.117, 0.060]	-0.178 [-0.255, -0.092]	4.426 [3.532, 5.558]	0.076 [0.065, 0.089]
Local log	0.442 [0.370, 0.531]	0.177 [0.104, 0.251]	0.038 [0.009, 0.074]	4.049 [2.985, 5.283]	0.080 [0.071, 0.089]
Local power	0.434 [0.364, 0.530]	0.151 [0.077, 0.232]	0.013 [-0.017, 0.044]	4.131 [3.028, 5.455]	0.082 [0.073, 0.091]
Sequence log MC	0.400 [0.311, 0.487]	0.012 [-0.097, 0.144]	-0.127 [-0.226, -0.030]	4.327 [3.234, 5.562]	0.078 [0.064, 0.093]
Top-16 log	0.434 [0.329, 0.547]	0.172 [0.111, 0.239]	0.033 [0.006, 0.057]	4.813 [3.601, 6.249]	0.100 [0.084, 0.116]

K.17 HellaSwag Trained-Endpoint Geometry

Table 88: Local log-field geometry at trained HELLASWAG endpoints.

Training method	$\widetilde{\Phi}_8$	$\widetilde{\Phi}_{16}$	$\Delta\widetilde{\Phi}_{16}$	$\|c\|/\|u\|_{16}$	$\|C\|/\|u\|_{16}$
Corrected log	-0.116 [-0.159, -0.078]	-0.054 [-0.077, -0.030]	-0.034 [-0.074, 0.004]	9.003 [8.152, 9.850]	0.095 [0.086, 0.104]
Corrected power	-0.101 [-0.141, -0.060]	-0.045 [-0.068, -0.014]	-0.024 [-0.065, 0.017]	9.531 [8.633, 10.428]	0.100 [0.091, 0.110]

Continued on next page

Table 88 (continued)

Training method	$\widetilde{\Phi}_8$	$\widetilde{\Phi}_{16}$	$\Delta\widetilde{\Phi}_{16}$	$\|c\|/\|u\|_{16}$	$\|C\|/\|u\|_{16}$
Local log	-0.122 [-0.198, -0.045]	-0.070 [-0.116, -0.033]	-0.050 [-0.085, -0.020]	6.487 [5.840, 7.145]	0.085 [0.076, 0.094]
Local power	-0.110 [-0.179, -0.044]	-0.057 [-0.088, -0.029]	-0.036 [-0.076, -0.003]	7.074 [6.315, 7.815]	0.096 [0.085, 0.107]
Sequence log MC	-0.123 [-0.168, -0.082]	-0.055 [-0.080, -0.027]	-0.034 [-0.076, 0.003]	9.018 [8.132, 9.926]	0.096 [0.086, 0.105]
Top-16 log	0.041 [-0.020, 0.101]	0.001 [-0.019, 0.021]	0.021 [-0.006, 0.050]	8.424 [7.644, 9.272]	0.125 [0.112, 0.138]

K.18 Model-Scale Trained-Endpoint Geometry

Table 89: Local log-field geometry at trained 4B-to-9B ARC endpoints.

Training method	$\widetilde{\Phi}_8$	$\widetilde{\Phi}_{16}$	$\Delta\widetilde{\Phi}_{16}$	$\|c\|/\|u\|_{16}$	$\|C\|/\|u\|_{16}$
Corrected log	0.607 [0.508, 0.755]	0.301 [0.200, 0.402]	-0.011 [-0.040, 0.017]	2.395 [1.855, 2.845]	0.026 [0.023, 0.028]
Corrected power	0.583 [0.475, 0.738]	0.283 [0.197, 0.370]	-0.029 [-0.059, -0.008]	2.381 [1.925, 2.828]	0.025 [0.023, 0.027]
Local log	0.517 [0.434, 0.620]	0.319 [0.216, 0.422]	0.007 [-0.052, 0.067]	2.527 [2.163, 2.868]	0.027 [0.019, 0.035]
Local power	0.516 [0.427, 0.621]	0.327 [0.235, 0.419]	0.014 [-0.041, 0.061]	2.523 [2.117, 2.770]	0.027 [0.021, 0.035]
Sequence log MC	0.568 [0.458, 0.727]	0.303 [0.221, 0.389]	-0.009 [-0.029, 0.010]	2.469 [2.068, 2.870]	0.027 [0.024, 0.029]
Top-16 log	0.413 [0.233, 0.593]	0.256 [0.148, 0.411]	-0.057 [-0.142, 0.029]	2.567 [2.176, 2.782]	0.030 [0.025, 0.037]

Table 90: Local log-field geometry at trained 4B-to-9B GSM8K endpoints.

Training method	$\widetilde{\Phi}_8$	$\widetilde{\Phi}_{16}$	$\Delta\widetilde{\Phi}_{16}$	$\|c\|/\|u\|_{16}$	$\|C\|/\|u\|_{16}$
Corrected log	0.847 [0.759, 0.941]	0.564 [0.369, 0.704]	-8.84×10^{-5} [-0.096, 0.141]	1.226 [0.999, 1.505]	0.019 [0.014, 0.025]
Corrected power	0.867 [0.776, 0.957]	0.612 [0.394, 0.784]	0.048 [-0.061, 0.214]	1.158 [0.986, 1.421]	0.018 [0.014, 0.025]
Local log	0.815 [0.767, 0.861]	0.574 [0.378, 0.712]	0.009 [-0.068, 0.148]	1.178 [0.929, 1.537]	0.019 [0.015, 0.026]
Local power	0.838 [0.789, 0.887]	0.575 [0.372, 0.720]	0.011 [-0.072, 0.157]	1.138 [0.900, 1.455]	0.018 [0.014, 0.026]
Sequence log MC	0.874 [0.736, 1.013]	0.617 [0.387, 0.821]	0.053 [-0.056, 0.261]	1.184 [0.954, 1.454]	0.018 [0.015, 0.025]
Top-16 log	0.800 [0.699, 0.902]	0.502 [0.310, 0.641]	-0.062 [-0.139, 0.083]	1.220 [0.826, 1.653]	0.020 [0.015, 0.028]

K.19 Reading the endpoint comparisons

Tables 89 and 90 evaluate the same local-log field at adapters reached by different methods. The method label identifies the training path; the fidelity describes a common field at its endpoint. Positive scale-panel values agree with the longer-horizon boundary in Observation 3, without implying that the methods optimize identical sequence objectives. The paired horizon-16 change records relocation relative to each restored initial adapter under a fixed measurement protocol.

These diagnostics complement the preceding evaluations. Cached-prefix gaps hold evaluation prefixes fixed, whereas fresh-sequence objectives also reflect the endpoint's trajectory distribution. Accuracy and pass@k supply separate task criteria, so their rankings need not agree. The seed-level outcomes and paired intervals distinguish uncertainty across training runs from differences between the objectives being measured. Each comparison therefore retains its declared model pair, task, horizon, parameter subspace, and evaluation objective.