VERPO:
Verified Evidence Regularized Policy Optimization

Anonymous authors
Paper under double-blind review

Abstract

Verifiable rewards improve language models through reliable task-level feedback, but methods based on Group Relative Policy Optimization (GRPO) apply a sequence-level advantage uniformly across all tokens. This coarse credit assignment reinforces or penalizes entire responses without identifying which local decisions to preserve, reinforce, or revise. Conversely, evidence-conditioned self-distillation provides denser token-level supervision, yet teacher imitation can transfer stylistic artifacts and miscalibrated confidence that destabilize training when misaligned with task success. We introduce VERPO, which converts evidence-conditioned guidance into reward-aligned token-level credit assignment while retaining the outcome objective. VERPO decomposes teacher guidance into an evidence-free reference term and signed, evidence-induced corrections at each token. A stopped controller combines selective acceptance, token-wise localization, and cost-aware scaling by balancing alignment with the local GRPO update direction against Fisher movement cost. Furthermore, we introduce Fisher Evidence Contrast (FEC), which attenuates nuisance shifts along an estimated evidence-presence direction through a regularized projection. Across five scientific reasoning and tool-use tasks, VERPO prevents optimization collapse and consistently achieves the highest multi-task average across model backbones, yielding marked improvements particularly on smaller models over strong baselines. Qualitative diagnostics confirm that token acceptance selectively targets reasoning bottlenecks consistent with local reward alignment and Fisher movement cost.

1 Introduction

Outcome-based policy optimization with verifiable rewards has become a central approach for improving the reasoning behavior of language models (DeepSeek-AI, 2025; Yu et al., 2025). In reinforcement learning with verifiable rewards (RLVR), task success can be checked automatically, allowing policy updates to be grounded in reliable outcome-level feedback. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) derives group-relative advantages from multiple on-policy rollouts and uses them to reinforce or suppress generated responses. In Fig. 1(a), GRPO (2024) resembles trial-and-error exploration: the reward is reliable, but it is returned only after a full response is completed. Consequently, the same sequence-level advantage is applied uniformly across response tokens, leaving the model without explicit information about which local decisions should be preserved, reinforced, or revised (Hübotter et al., 2026).

Evidence-conditioned self-distillation offers a natural way to make this feedback denser (Zhao et al., 2026a; Hübotter et al., 2026). A self-teacher can replay sampled trajectories with privileged evidence, such as training-only feedback, reference information, or external signals, and then provide token-level guidance to a student that will not receive such evidence at deployment time. In Fig. 1(b), evidence can point to local correction opportunities along a rollout trajectory rather than waiting until the final answer is judged. However, we observe that relatively little work has examined whether an evidence-induced teacher change is actually a reward-aligned correction. A teacher conditioned on privileged evidence can induce shifts not only task-relevant decisions, but also formatting, verbosity, confidence, or reasoning style. Naively imitating such a teacher can therefore transfer evidence-induced artifacts and destabilize learning when these changes are not aligned with verifiable task success (Armandpour et al., 2026; Harne et al., 2026).

Table 1 summarizes this gap. GRPO (2024) preserves a direct reward objective, but lacks dense token-level guidance and cannot decide which local decisions should receive credit or correction.

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.

Self-distillation methods such as SDPO (Hübotter et al., 2026) provide dense token guidance, but do not by themselves anchor each local teacher preference to verifiable reward improvement. What is missing is not merely another weighted combination of an RL loss and a distillation loss. Rather, the central problem is how to convert evidence-conditioned guidance into reward-aligned token-level credit for outcome-based policy optimization, while accounting for the cost of changing the policy. This requires deciding which evidence-induced corrections should be accepted, where along the response they should act, and how strongly they should be scaled. This leads us to ask: *How can evidence-conditioned guidance be converted into reward-aligned token-level credit while accounting for the cost of changing the policy?*

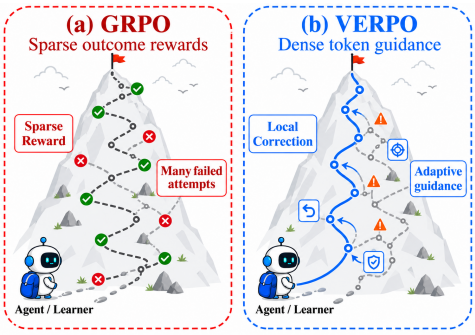


Figure 1: GRPO (2024) vs. VERPO (Ours). GRPO (2024) uses sparse sequence-level rewards for trial-and-error exploration in (a), whereas the proposed VERPO converts evidence-conditioned guidance into reward-aligned token-level credit assignment in (b).

Table 1: Token-level credit mechanisms. DRO: direct reward objective. DTG: dense token guidance/credit modulation. SA: Selective Acceptance via sample/token gating beyond evidence availability. TL: Token-wise Localization via explicit gating. CS: Cost-aware Scaling by local policy movement.

Method	DRO	DTG	SA	TL	CS
GRPO (2024)	✓	✗	✗	✗	✗
SDPO (2026)	✗	✓	✗	✗	✗
SRPO (2026)	✓	✓	✓	✗	✗
RLSD (2026a)	✓	✓	✗	✗	✗
RLCSD (2026)	✓	✓	✓	✓	✗
VERPO (Ours)	✓	✓	✓	✓	✓

Bearing this in mind, we introduce **VERPO**, a **V**erified **E**vidence-**R**egularized **P**olicy **O**ptimization framework. VERPO treats evidence as a proposal for policy correction rather than as a replacement for the outcome objective (Shao et al., 2024). It decomposes teacher guidance into two components: an evidence-free reference term that preserves the behavior available without privileged evidence, and a signed evidence-induced correction that proposes how the policy should move at each token. This separation allows the method to distinguish reference preservation from evidence-driven change, instead of collapsing both into a single imitation target.

To this end, VERPO uses a stopped token-wise ZPD controller, named after the zone of proximal development (Vygotsky, 1978). This concept distinguishes independent capabilities from those attainable with guidance. Analogously, the controller scales corrections by their alignment with the local GRPO (2024) update direction and Fisher movement cost. Guidance is strongest for reward-aligned corrections with modest policy movement, and weakest for costly or misaligned corrections. In Table 1, VERPO frames reward-aligned token-level credit assignment through three missing mechanisms: 1) selective acceptance, 2) token-wise localization, and 3) cost-aware scaling of evidence-induced corrections. To construct evidence directions, VERPO supports multiple teacher comparisons. The Fix construction compares an evidence-conditioned teacher with an evidence-free teacher, while the contrastive construction compares teacher distributions under correct and incorrect evidence. To reduce nuisance shifts caused by the mere presence of privileged evidence, we introduce Fisher Evidence Contrast (FEC), which projects out the corresponding evidence-presence component. These correction directions can be used through two complementary update paths: they can weight the evidence-distillation loss, modulate the token-level policy advantage, or enter both paths through the same stopped controller. In summary, our contributions are threefold:

- We formulate evidence-conditioned self-distillation for RLVR as a reward-aligned token-level credit assignment problem. This formulation separates evidence-free reference preservation from signed evidence-induced policy correction, clarifying why dense teacher guidance should be treated as a local correction proposal rather than a direct imitation target.
- We present VERPO, which uses a stopped token-wise ZPD controller to accept and scale each correction according to local reward alignment and Fisher movement cost. It further supports multiple evidence-direction constructions, including Fix, contrastive evidence correction, and Fisher Evidence Contrast, which removes nuisance shifts caused by evidence presence.
- We conduct extensive experiments on a five-task suite spanning scientific reasoning and tool use (Feng et al., 2024; Tang et al., 2023) across three model backbones. Diagnostic analyses further reveal selective, task-dependent correction behavior, showing that evidence is most useful when it is locally reward-aligned and incurs a small Fisher movement cost.

2 Related Work

Outcome Rewards and Token-level Supervision. GRPO (Shao et al., 2024) learns from verifiable outcomes using sequence-level advantages shared across tokens. Process supervision, including Let's Verify Step by Step (Lightman et al., 2024) and Math-Shepherd (Wang et al., 2024), supplies finer-grained targets through step verification (Setlur et al., 2025). Math-Shepherd (Wang et al., 2024) constructs step labels automatically, reducing dependence on human annotations while retaining an explicit process supervision stage. This distinction separates the source of supervision from the granularity at which it is applied: a reliable final outcome does not itself identify which intermediate decisions deserve correction (Hübotter et al., 2026). VERPO retains outcome-based optimization and uses Teacher replay to propose local corrections without a learned process verifier.

Self-distillation and Reward Alignment. On-policy distillation (Agarwal et al., 2023; Gu et al., 2023) trains on student-generated trajectories with teacher feedback, addressing the mismatch between fixed training sequences and the student's own generations. Privileged self-distillation instead obtains an informative Teacher view by enriching the context available during replay (Lu et al., 2026). OPSD (Zhao et al., 2026a) and SDPO (Hübotter et al., 2026) obtain dense supervision from self-teachers conditioned on privileged solutions or feedback. SRPO (Li et al., 2026) routes samples between reward optimization and self-distillation, while RLSD (Yang et al., 2026a) modulates distillation with reward-directed signals. These mechanisms motivate selective guidance, but do not determine token-wise acceptance by jointly measuring local reward alignment and policy movement cost. VERPO uses a stopped ZPD weight to make this decision for each eligible token. It keeps the evidence-free reference term active when an evidence correction is rejected, separating preservation of the base behavior from acceptance of a proposed change.

Contrastive and Selective Privileged Supervision. Contrastive methods separate task evidence from contextual shifts. RLCSD (Pan et al., 2026) contrasts correct and incorrect hints to suppress style drift. CEPO (Heakl et al., 2026) uses correct-versus-wrong evidence for token credit, while OCSD (Yang et al., 2026b) contrasts matched replay views to discount shared scaffold effects. Selective methods instead restrict where supervision applies: EDGE-OPD (Lazaridis et al., 2026) masks tokens by positive evidence, and ROSD (Zhao et al., 2026b) localizes correction to erroneous spans. Distillation studies identify reference-trajectory bias in privileged Teachers (Harne et al., 2026) and weaker gradient alignment on correct than incorrect rollouts (Armandpour et al., 2026). SA-OPD (Jiang et al., 2026) filters signals that combine weak input grounding with extreme divergence. These approaches address evidence relevance, correction location, and signal reliability. VERPO additionally separates evidence-free reference restoration from signed correction, uses FEC to attenuate a measured evidence-presence component under the Fisher metric, and accepts token-wise corrections according to local reward alignment and movement cost. Appendix A provides the extended comparison.

3 VERPO: Verified Evidence Regularized Policy Optimization

3.1 Overview of VERPO

VERPO separates signed Teacher corrections from token-wise acceptance. ZPD scales each correction by local reward alignment and Fisher movement cost, while an evidence-free reference anchors the Actor's behavior. We use forward-KL geometry, denoted FKL. Appendix E.3 derives the reverse-KL counterpart, denoted RKL. Figure 2 summarizes four training stages:

- **Stage 1: Rollout Sampling and Outcome Signal.** The Actor samples response groups, and verifier rewards yield the GRPO (Shao et al., 2024) advantage $\widehat{A}_i$.
- **Stage 2: Token-Level Evidence Correction.** Teacher replay under different evidence contexts yields Fix (Hübotter et al., 2026), CTR (Pan et al., 2026), or FEC corrections, defined in Section 3.4.
- **Stage 3: ZPD Weight Estimation.** Local alignment and Fisher cost determine each eligible token's stopped weight $w^*_{i,t}$, derived in Section 3.3.
- **Stage 4: Two-Path Policy Update.** The weight either scales the evidence correction or modulates the GRPO (Shao et al., 2024) advantage. Section 3.5 defines both update paths.

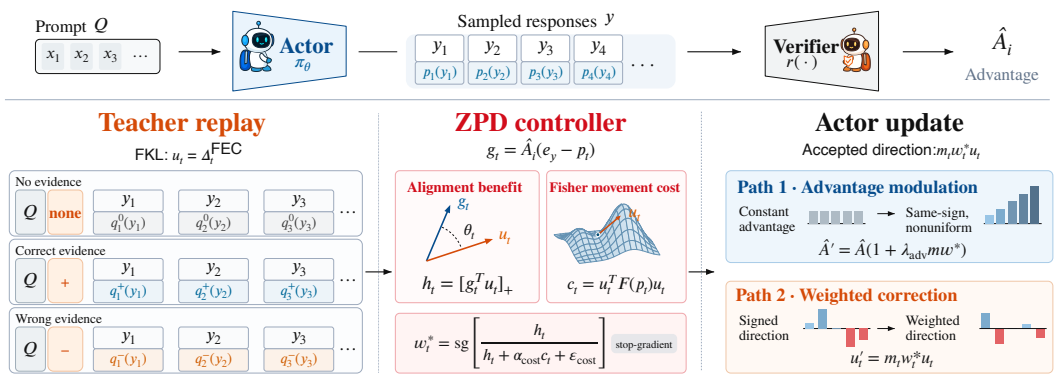


Figure 2: Overview of VERPO. Teacher replay constructs FEC correction directions. The ZPD controller balances alignment benefit against Fisher movement cost to determine token-wise acceptance. Path 1 modulates advantages without changing their signs; Path 2 applies weighted corrections.

3.2 FKL Evidence Direction and OPSD Decomposition

We separate reference restoration from evidence correction using base and enriched Teacher endpoints q_b, q_e, defined in Section 3.4. The Actor $p_{i,t} = \pi_\theta(\cdot \mid x_i, y_{i,<t})$ uses the original prompt and sampled prefix. For one Teacher comparison, define $\Delta = q_e - q_b$ and $q_w^{\text{FKL}} = q_b + w\Delta$ for $0 \le w \le 1$. Here Δ specifies the signed evidence correction, while w controls how far the Teacher target moves from q_b toward q_e. The path preserves normalization and nonnegativity. Holding the Teacher endpoints and w fixed, and assuming positive Actor probability on the Teacher support, we expand the forward KL as

$$\text{KL}(q_w^{\text{FKL}} \| p) = \text{KL}(q_b \| p) - w \sum_{v \in \mathcal{V}} \Delta(v) \log p(v) + C(q_w^{\text{FKL}}, q_b), \tag{1}$$

The first term restores q_b, the second contributes $w\Delta$ to the negative logit gradient, and C has no Actor gradient. Appendix E.2 derives both in Equations (E.5) and (E.6). We therefore restore an independent evidence-free reference $q_{i,t}^{\text{ref}}$ through

$$\mathcal{L}_{\text{ref}}^{\text{FKL}} = \frac{1}{N} \sum_{i,t} a_{i,t} \text{KL}(q_{i,t}^{\text{ref}} \| p_{i,t}), \tag{2}$$

Here $a_{i,t}$ selects valid response tokens and N counts them. This loss remains active independently of evidence mask $m_{i,t}$ and stopped acceptance $w_{i,t}$, which control the correction. Thus, rejecting an evidence-induced change does not disable reference restoration. The evidence loss is

$$\mathcal{L}_{\text{OPSD}}^{\text{FKL}}(w) = -\frac{1}{Z_{\text{evi}}} \sum_{i,t} m_{i,t} w_{i,t} \sum_{v \in \mathcal{V}} \Delta_{i,t}(v) \log p_{i,t}(v). \tag{3}$$

Z_{evi} counts eligible tokens with a floor of one, and zero weights remove corrections. These losses use the full vocabulary. Appendix H.4 details finite-support computation.

3.3 Token-Wise ZPD Controller

To derive acceptance, let θ_{base} denote a hypothetical Actor state after GRPO (Shao et al., 2024) and reference updates. Training instead uses a joint loss, with $\tilde{\theta}$ denoting the EMA Teacher. For logits $z_{i,t}$ and Jacobian $G_{i,t} = \partial z_{i,t} / \partial \theta$, the FKL direction $u_{i,t} = \Delta_{i,t}$ induces $d_{i,t} = G_{i,t}^\top u_{i,t}$. At step size $\eta > 0$, ideal acceptance is

$$w_{i,t}^{\text{ideal}} = \arg \max_{0 \le w \le 1} J_{\text{RL}}\left(\theta_{\text{base}} + \eta w d_{i,t}\right). \tag{4}$$

For twice-differentiable J_{RL}, a second-order expansion gives

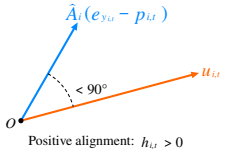
$$J_{\text{RL}}(\theta_{\text{base}} + \eta w d_{i,t}) - J_{\text{RL}}(\theta_{\text{base}}) \approx \eta w \nabla J_{\text{RL}}(\theta_{\text{base}})^\top d_{i,t} + \tfrac{1}{2} \eta^2 w^2 d_{i,t}^\top \nabla^2 J_{\text{RL}}(\theta_{\text{base}}) d_{i,t}. \tag{5}$$

The linear term measures reward alignment, but the quadratic term penalizes movement only under negative directional curvature. We use local alignment and nonnegative Fisher cost proxies instead.

With sampled-token one-hot vector $e_{y_{i,t}}$, $[x]_+ = \max(x, 0)$, and $F(p) = \mathrm{diag}(p) - pp^\top$, define

$$h_{i,t} = \left[\widehat{A}_i (e_{y_{i,t}} - p_{i,t})^\top u_{i,t}\right]_+, \qquad c_{i,t} = u_{i,t}^\top F(p_{i,t}) u_{i,t}. \tag{6}$$

Selective Acceptance. $h_{i,t}$ measures positive alignment with the unclipped local GRPO (Shao et al., 2024) score direction, not verified improvement. For nonzero $\widehat{A}_i(e_{y_{i,t}} - p_{i,t})$ and $u_{i,t}$, an angle below 90° gives a positive inner product and hence $h_{i,t} > 0$, as illustrated in Figure 3.3. Fisher cost still controls the acceptance magnitude. Equation (7) assigns zero weight to nonpositive alignment.



Token-wise Localization. Despite the shared advantage $\widehat{A}_i$, prefix-dependent $p_{i,t}$ and $u_{i,t}$ give token-specific weights. The effective weight $m_{i,t} w^*_{i,t}$ combines evidence eligibility with local acceptance, independently of restoration.

Cost-aware Scaling. A small logit step $\delta u_{i,t}$ incurs second-order KL movement $\frac{1}{2}\delta^2 c_{i,t}$. Appendix E.6 derives this Fisher approximation. We combine movement cost with a positive floor as $\Lambda_{i,t} = \alpha_{\text{cost}} c_{i,t} + \epsilon_{\text{cost}} > 0$, where $\alpha_{\text{cost}} \geq 0$ scales cost and $\epsilon_{\text{cost}} > 0$ keeps the denominator positive. We then maximize the surrogate $f_{i,t}(w) = h_{i,t} \log w + \Lambda_{i,t} \log(1 - w)$ over $0 < w < 1$. For positive benefit, its first term discourages zero acceptance, while the second penalizes full acceptance. This surrogate balances the local proxies without asserting an exact reward improvement. For $h_{i,t} > 0$, strict concavity and $f'_{i,t}(w) = h_{i,t}/w - \Lambda_{i,t}/(1 - w) = 0$ give a unique maximizer, extended to zero at $h_{i,t} = 0$:

$$w^*_{i,t} = \mathrm{sg}\left[\frac{h_{i,t}}{h_{i,t} + \Lambda_{i,t}}\right] = \mathrm{sg}\left[\frac{h_{i,t}}{h_{i,t} + \alpha_{\text{cost}} c_{i,t} + \epsilon_{\text{cost}}}\right]. \tag{7}$$

The stopped weight lies in $[0, 1)$, increases with benefit at fixed cost, and decreases with cost when both benefit and α_{cost} are positive. Otherwise it is nonincreasing. Stop-gradient blocks controller backpropagation. Appendix E.7 proves these properties and the zero-benefit convention.

3.4 Fix, Contrastive, and Fisher Evidence Contrast

Fisher Evidence Contrast (FEC) subtracts the task contrast's regularized Fisher projection onto a measured evidence-presence direction. Figure 3 motivates this separation: incorrect and shuffled evidence also shift the Teacher, although this shift need not be entirely task-irrelevant.

Evidence Contexts. The frozen Teacher branches q^0, q^+, and q^- score the same sampled prefix with no evidence, the ground-truth answer, and a wrong answer, respectively.

Endpoint Comparisons. Fix (Hübotter et al., 2026) uses $\Delta^{\text{Fix}} = q^+ - q^0$. Contrastive correction, denoted CTR, follows RLCSD (Pan et al., 2026) with $\Delta^{\text{CTR}} = q^+ - q^-$.

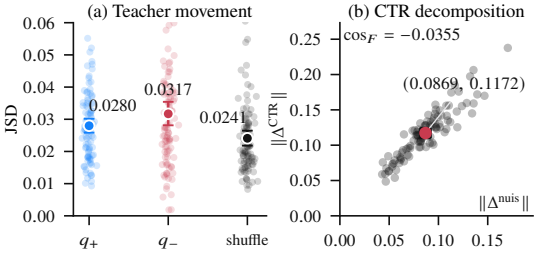


Figure 3: Evidence-presence shifts motivate FEC. (a) divergence under correct, incorrect, and shuffled evidence; (b) task-contrast versus evidence-presence displacement norms.

FEC Refinement. To attenuate the contrast along the measured evidence-presence direction, we average the correct and incorrect Teacher branches and compare this midpoint with the evidence-free branch. We then measure the task contrast's component along this direction under the Fisher metric with a positive stabilizer $\epsilon_{\text{proj}} > 0$:

$$\Delta^{\text{nuis}} = \tfrac{1}{2}\left(q^+ + q^-\right) - q^0, \qquad \alpha^{\text{FEC}} = \frac{(\Delta^{\text{CTR}})^\top F(p) \Delta^{\text{nuis}}}{(\Delta^{\text{nuis}})^\top F(p) \Delta^{\text{nuis}} + \epsilon_{\text{proj}}}. \tag{8}$$

The coefficient normalizes the Fisher inner product by the regularized squared length of the presence proxy. Subtracting this component gives

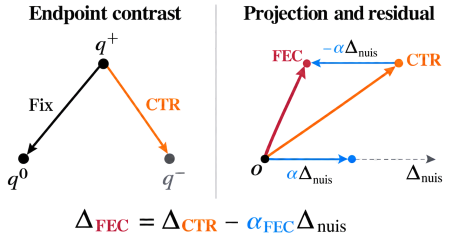
$$\Delta^{\text{FEC}} = \Delta^{\text{CTR}} - \alpha^{\text{FEC}} \Delta^{\text{nuis}}. \tag{9}$$

Regularization attenuates the residual's Fisher inner product with the nuisance proxy without guaranteeing exact orthogonality. We stop gradients through all of Δ^{FEC}, including its Actor-

dependent coefficient, and use it in the evidence loss and controller. Table 2 summarizes the variants alongside a geometric illustration of the endpoint comparisons and FEC refinement. Appendix E.4 derives the residual inner product and RKL construction.

Table 2: FKL evidence-direction constructions and FEC geometry. Endpoint arrows point toward the subtracted branch. FEC removes the regularized nuisance component from CTR.

Variant	Branches	FKL correction
Fix (2026)	$q^0 \to q^+$	$\Delta^{\text{Fix}} = q^+ - q^0$
CTR (2026)	$q^- \to q^+$	$\Delta^{\text{CTR}} = q^+ - q^-$
FEC(Ours)	q^0, q^-, q^+	$\Delta^{\text{FEC}} = \Delta^{\text{CTR}} - \alpha^{\text{FEC}}\Delta^{\text{nuis}}$



3.5 Two Actor-Update Paths

The stopped controller supports two paths, evaluated separately. Path 1 modulates the GRPO (Shao et al., 2024) advantage, while Path 2 scales the signed correction. For Path 1, the masked weight $\widetilde{w}_{i,t} = m_{i,t} w^*_{i,t}$ combines evidence eligibility with controller acceptance at each sampled token. For $\lambda_{\text{adv}} \geq 0$, the modulated advantage is

$$\widehat{A}'_{i,t} = \widehat{A}_i [1 + \chi_{\text{adv}} \lambda_{\text{adv}} \, \text{sg}(\widetilde{w}_{i,t})]. \tag{10}$$

The positive stopped multiplier preserves the advantage sign and clipping decision, equals one for zero or ineligible acceptance, and contributes no gradient. With switches $\chi_{\text{OPSD}}, \chi_{\text{adv}} \in \{0, 1\}$, the joint objective is

$$\boxed{\mathcal{L}^{\text{FKL}}_{\text{total}} = \mathcal{L}_{\text{GRPO}}(\widehat{A}') + \lambda_{\text{ref}} \mathcal{L}^{\text{FKL}}_{\text{ref}} + \chi_{\text{OPSD}} \lambda_{\text{evi}} \mathcal{L}^{\text{FKL}}_{\text{OPSD}}(w^*).} \tag{11}$$

The outcome loss supplies the task signal, while λ_{ref} and λ_{evi} scale reference restoration and accepted correction. Path 1 uses $(\chi_{\text{adv}}, \chi_{\text{OPSD}}) = (1, 0)$ and Path 2 uses $(0, 1)$. Algorithm 1 specifies masking, normalization, and the joint update.

Algorithm 1 VERPO Actor Update under FKL Geometry

Require: Actor π_θ, EMA Teacher $\pi_{\bar{\theta}}$, data $\mathcal{D}$, variant s, switches $\chi_{\text{OPSD}}, \chi_{\text{adv}}$

repeat
  Freeze $\theta_{\text{old}} \leftarrow \theta$; sample response groups from $\pi_{\theta_{\text{old}}}$ for prompts in $\mathcal{D}$.
  Obtain verifier rewards R_i; set $\widehat{A}_i \leftarrow R_i - |\mathcal{G}(i)|^{-1} \sum_{j \in \mathcal{G}(i)} R_j$. ▷ Outcome anchor
  Score Actor prefixes as $p_{i,t}$; replay them with stopped Teacher branches $q^{\text{ref}}_{i,t}, q^0_{i,t}, q^+_{i,t}, q^-_{i,t}$.
  Select $\Delta_{i,t}$: $q^+_{i,t} - q^0_{i,t}$ (Fix), $q^+_{i,t} - q^-_{i,t}$ (CTR), or Eq. (9) (FEC). ▷ Signed correction
  Set $u_{i,t} \leftarrow \text{sg}[\Delta_{i,t}]$; compute $h_{i,t}, c_{i,t}$ via Eq. (6). ▷ Alignment and cost
  Set $w^*_{i,t} \leftarrow \text{sg}[h_{i,t}/(h_{i,t} + \alpha_{\text{cost}} c_{i,t} + \epsilon_{\text{cost}})]$. ▷ Token-wise ZPD gate
  Form masks $a_{i,t}$ (valid) and $m_{i,t} \leq a_{i,t}$ (evidence); set $\widetilde{w}_{i,t} \leftarrow m_{i,t} w^*_{i,t}$.
  Set $N \leftarrow \sum_{i,t} a_{i,t}$ and $Z_{\text{evi}} \leftarrow \max(1, \sum_{i,t} a_{i,t} m_{i,t})$.
  Form $\widehat{A}'_{i,t}$ via Eq. (10); compute normalized losses via Eqs. (2) and (3). ▷ Two-path update
  Update $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}^{\text{FKL}}_{\text{total}}$ via Eq. (11); update the EMA Teacher via Eq. (D.7).
until stopping criterion is reached
return π_θ

4 Experiments

4.1 Experimental Setup

We use instruction-tuned Qwen3-4B, Qwen3-8B (Yang et al., 2025), and Llama-3.2-1B-Instruct (Meta, 2024) as backbones. Following the evaluation setup of SDPO (Hübotter et al., 2026), we consider five benchmarks: Chemistry, Physics, Biology, Materials, and Tool Use. The four science benchmarks use the reasoning subsets of SciKnowEval (Feng et al., 2024) to evaluate undergraduate-level scientific

reasoning, while Tool Use uses ToolAlpaca (Tang et al., 2023) to evaluate tool-call generation given a user request and tool specifications. We compare against GRPO (Shao et al., 2024), which learns from group-relative outcome rewards; SDPO (Hübotter et al., 2026), which obtains dense token supervision from a feedback-conditioned teacher; SRPO (Li et al., 2026), which routes successful samples to the GRPO objective and unsuccessful samples to the SDPO objective; RLSD (Yang et al., 2026a), which uses verifier rewards to modulate self-distillation; and RLCSD (Pan et al., 2026), which contrasts teachers conditioned on correct and incorrect hints. We evaluate two VERPO variants: VERPO-LW applies the ZPD weights to the evidence loss, while VERPO-AM uses them to modulate the GRPO advantages. We reports each task's avg@16. Appendix H clarifies this protocol and its limitations.

4.2 Main Results

Table 3: Best-checkpoint avg@16 results. Base scores appear in the backbone rows. In the Path column, 1 denotes advantage modulation, which rescales token advantages, and 2 denotes weighted correction, which scales token-level evidence corrections. RH denotes reward hacking, blank entries denote unavailable scores, and † marks collapse or an average containing one. SDPO Physics on Llama-3.2-1B reports its pre-collapse peak.

Path	Method	Biology	Chemistry	Materials	Physics	Tool Use	Average
Qwen3-8B		28.12	40.86	58.44	60.85	58.27	49.31
1	RLSD (Yang et al., 2026a)	53.50	69.82	76.06	66.17	64.06	65.92
	RLCSD (Pan et al., 2026)	38.62	62.94	66.88	51.01	54.50	54.79
	VERPO-AM (Ours)	59.13	<u>78.46</u>	<u>76.79</u>	**69.09**	**69.42**	<u>70.58</u>
2	GRPO (Shao et al., 2024)	<u>61.75</u>	70.38	76.66	66.87	67.18	68.57
	SDPO (Hübotter et al., 2026)	60.75	77.94	74.93	64.06	67.09	68.95
	SRPO (Li et al., 2026)	60.00	73.51	73.27	68.51	66.36	68.33
	VERPO-LW (Ours)	**62.37**	**80.26**	**77.62**	<u>68.85</u>	<u>68.15</u>	**71.44**
Qwen3-4B		33.25	44.37	60.90	61.40	58.91	51.77
1	RLSD (Yang et al., 2026a)	51.50	69.13	77.65	68.51	60.75	65.51
	RLCSD (Pan et al., 2026)	47.25	60.13	67.15	50.93	58.91	56.87
	VERPO-AM (Ours)	55.37	**72.85**	75.44	**70.78**	59.09	66.71
2	GRPO (Shao et al., 2024)	52.25	70.00	**80.18**	67.42	60.56	66.08
	SDPO (Hübotter et al., 2026)	47.87	69.79	72.40	62.89	57.72	62.13
	SRPO (Li et al., 2026)	<u>60.12</u>	<u>72.47</u>	78.78	67.89	**62.04**	<u>68.26</u>
	VERPO-LW (Ours)	**62.00**	69.17	<u>79.72</u>	<u>70.55</u>	<u>61.40</u>	**68.57**
Llama-3.2-1B		10.62	7.41	5.59	5.55	1.47	6.13
1	RLSD (Yang et al., 2026a)	RH	RH	RH	RH	RH	RH
	RLCSD (Pan et al., 2026)	15.87	24.97†	30.71	20.85	33.45	25.17†
	VERPO-AM (Ours)	<u>55.50</u>	63.69	**55.06**	50.70	**51.01**	<u>55.19</u>
2	GRPO (Shao et al., 2024)	42.38	52.27	42.93	<u>52.47</u>	47.51	47.51
	SDPO (Hübotter et al., 2026)	31.50	30.47	29.38	40.15†	43.85	35.07†
	SRPO (Li et al., 2026)	31.87	**67.11**	31.91	44.45	42.15	43.50
	VERPO-LW (Ours)	**60.15**	<u>66.57</u>	<u>52.26</u>	**54.22**	<u>49.63</u>	**56.57**

Table 3 shows that VERPO-LW achieves the highest five-task best-checkpoint average on every backbone among methods with complete and noncollapsed results. On Llama-3.2-1B, RLSD (Yang et al., 2026a) exhibits reward hacking, while SDPO (Hübotter et al., 2026) and RLCSD (Pan et al., 2026) include collapsed runs and are excluded from average ranking. The SDPO Physics score is a pre-collapse peak, not its final performance. Its descriptive five-task average retains this peak and carries the same collapse marker. The preferred update path varies by task. On Qwen3-4B, VERPO-LW achieves an average of 68.57, exceeding the strongest baseline by 0.31 points. On Llama-3.2-1B, it reaches 56.57, improving over the strongest baseline with complete and noncollapsed results by 9.06 points. On Qwen3-8B, VERPO-LW exceeds the strongest baseline average by 2.49 points. The compute-matched comparison in Appendix I.2 also favors VERPO on Llama-3.2-1B under a matched GPU budget. Within the same training time, VERPO achieves an average accuracy of 50.26%, exceeding GRPO (Shao et al., 2024) by 2.35 percentage points.

Table 4: (a) Llama-3.2-1B component. Teacher, direction, and geometry comparisons. (b) Chemistry evidence-loss scale. avg@16 is the step-200 score. Late reports the mean and sample standard deviation over 11 evaluations at steps 150–200. (c) Chemistry uniform weights. L1B/Q4B/Q8B denote Llama-3.2-1B/Qwen3-4B/Qwen3-8B. RH denotes reward hacking, – unavailable scores, and † collapse. Marked scores in (c) are pre-collapse peaks.

(a) Component ablations

Setting	Biology	Chemistry	Materials	Physics	Tool Use	Average
Teacher source						
Fixed Teacher	28.50	46.93	20.07	23.82	41.62	32.19
EMA-0.99	49.75	62.41	38.03	46.32	48.80	49.06
EMA-0.95	60.15	66.57	52.26	54.22	49.63	56.57
Evidence direction						
Fixed (2026)	RH	RH	–	–	–	–
CTR (2026)	57.75	65.86	44.89	52.09	47.19	53.56
FEC	60.15	66.57	52.26	54.22	49.63	56.57
KL geometry						
FKL	60.15	66.57	52.26	54.22	49.63	56.57
RKL	20.87†	31.10†	–	–	–	–

(b) Evidence-loss scale

λ_{evi}	avg@16	Late
0.2	65.09	63.35 ± 1.41
0.4	65.00	63.51 ± 0.96
0.6	61.28	62.78 ± 1.92
0.8	62.11	60.63 ± 1.29
1.0*	66.58	65.52 ± 1.09

(c) Uniform weights

w	L1B	Q4B	Q8B
0.25	53.48	71.99	73.30
0.50	20.88†	68.89	63.06
0.75	30.47†	46.57†	63.51
1.00	20.53†	46.16†	37.76

4.3 Ablation Study

Component Ablations. Table 4(a) compares Teacher update rules, evidence directions, and KL geometries on Llama-3.2-1B. The more responsive EMA Teacher performs best among the tested update rules, while FEC consistently outperforms CTR (Pan et al., 2026) across all tasks. The Fixed direction exhibits reward hacking on Biology and Chemistry, with the remaining task results unavailable. FKL provides complete results across all tasks, whereas the evaluated RKL runs collapse on Biology and Chemistry and are unavailable elsewhere.

Evidence-Loss Scale. Across the tested configurations, increasing λ_{evi} does not yield monotonic improvements in step-200 accuracy. The default configuration also achieves the highest late-window mean accuracy of 65.52%, consistent with its step-200 ranking. Appendix F.4 details the protocol.

Uniform Evidence Weights. To test the role of token-wise acceptance, we replace ZPD with a constant w in VERPO-LW while retaining the remaining settings within each backbone. Table 4(c) shows two trends on Chemistry. At the same uniform weight, smaller evaluated backbones are more prone to collapse. Within each backbone, increasing the uniform weight generally worsens validation performance, although the trend is not strictly monotonic. These observations support the instability concern in the introduction: indiscriminate evidence correction can destabilize training, with smaller backbones showing greater vulnerability in the tested settings.

4.4 Further Analysis

Training Dynamics. To assess how performance develops during training, Figure 4 tracks the 200-step Physics runs on Llama-3.2-1B. VERPO finishes with higher validation accuracy than GRPO (Shao et al., 2024), although their accuracy curves cross during training. The SDPO (Hübotter et al., 2026) run reaches zero validation accuracy at the displayed checkpoints from step 160 onward, after the pre-collapse peak reported in Table 3. VERPO retains longer responses late in training than the plotted baselines, while its Actor entropy remains low.

FEC Diagnostics. Figure 5 reports task-wise effective-weight coverage, signed pre-clamp benefit, Fisher cost, and residual nuisance covariance. Near the final displayed step, Materials shows broader coverage than Tool Use, but this reflects the extent of evidence acceptance rather than relative task benefit, and cannot be reconstructed from aggregate benefit and cost alone. The near-zero mean residual covariance supports the regularized projection only in aggregate and may conceal token-level cancellation. Definitions and the individual-direction identity are given in Appendices G and E.4.

Token-Level FEC Weight Analysis. Figure 6 examines where FEC and ZPD accept evidence in matched Materials responses. Correct rollouts show broader acceptance, while both trajectory classes concentrate most weight in intermediate reasoning rather than final-answer tokens, reflecting both region length and higher per-token weights. Token-type rankings illustrate lexical allocation patterns

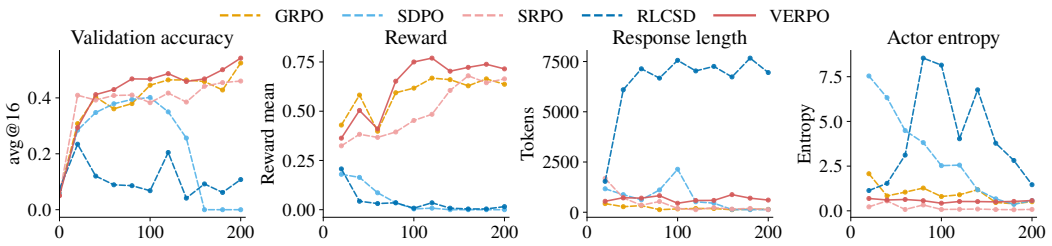


Figure 4: Physics training on Llama-3.2-1B for 200 steps: validation avg@16 ↑, reward, response length, and Actor entropy. Methods are GRPO (Shao et al., 2024), SDPO (Hübotter et al., 2026), SRPO (Li et al., 2026), RLCSD (Pan et al., 2026), and VERPO.

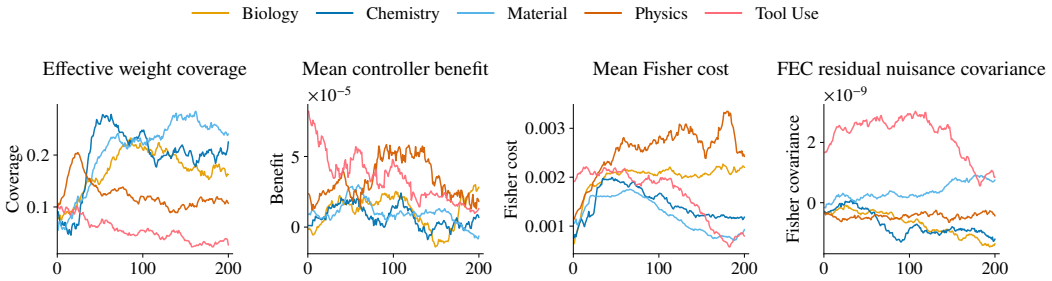
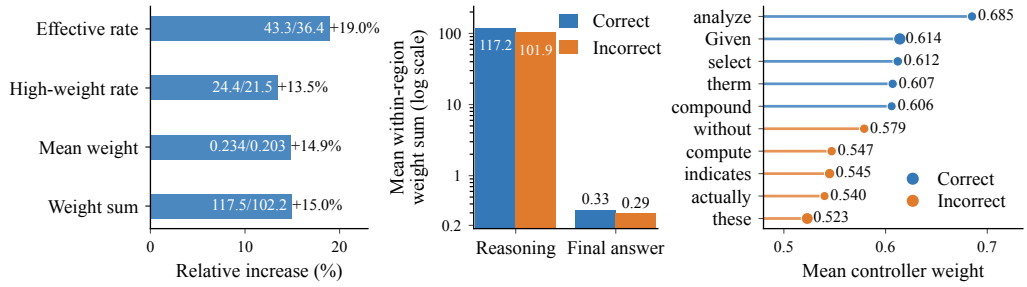


Figure 5: Llama-3.2-1B diagnostics across five tasks: effective-weight coverage ($w_{i,t} > 10^{-3}$), signed benefit, Fisher cost, and residual–nuisance covariance. Appendix G defines the metrics. Near-zero mean covariance does not imply token-wise orthogonality.



(a) Trajectory-level weight comparison (b) Region-level weight allocation (c) Token-level weight ranking

Figure 6: FEC and ZPD acceptance in matched Materials responses: (a) trajectory classes, (b) regional weight sums on a log scale, and (c) token-type mean weights with frequency-scaled markers. Effective/high weights satisfy $w > 10^{-3}/w \geq 0.5$. Token types occur at least 20 times.

but characterize acceptance rather than causal error contributions. The mean per-token weight in the reasoning region is approximately six times that in the final-answer region for both trajectory classes. In the offline replay, 36.4% of tokens in incorrect rollouts receive effective weights, showing that the controller also accepts candidate evidence corrections within unsuccessful responses. Appendix F.3 reports the complete statistics, and Appendix J provides a rollout-level example.

5 Conclusion

We introduce VERPO, which combines verifiable outcome rewards with dense evidence-conditioned supervision by separating evidence-free reference restoration from signed evidence-induced correction. FEC removes the measured evidence-presence component, while a token-wise ZPD controller accepts corrections based on local reward alignment and Fisher movement costs. Across five benchmarks and three model backbones, VERPO achieves the best average performance within each model block. Ablations show that training stability depends on the Teacher update rule, evidence construction, and KL geometry. Fine-grained analysis further shows that accepted corrections concentrate mainly in intermediate reasoning rather than final-answer tokens.

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. *arXiv preprint arXiv:2306.13649*, 2023.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M. Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why. *arXiv preprint arXiv:2605.10889*, 2026.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. SciKnowEval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

Wei Feng, Will Constable, and Yifan Mao. Getting started with fully sharded data parallel (FSDP2). PyTorch Tutorials, 2025. URL `https://docs.pytorch.org/tutorials/intermediate/FSDP_tutorial.html`. Updated September 2, 2025. Accessed September 5, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Sarthak Harne, Chinmay Karkar, Yash Pandya, Ahmed Awadallah, and Akshay Nambi. Privileged, but biased: How PI-conditioned teachers break self-distillation. *arXiv preprint arXiv:2608.04794*, 2026. doi: 10.48550/arXiv.2608.04794.

Ahmed Heakl, Abdelrahman M. Shaker, Youssef Mohamed, Rania Elbadry, Omar Fetouh, Fahad Shahbaz Khan, and Salman Khan. CEPO: RLVR self-distillation using contrastive evidence policy optimization. *arXiv preprint arXiv:2605.19436*, 2026.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Yinuo Jiang, Yongjie Ye, Zhou Tao, Xiang Zhuang, Qiang Zhang, Huajun Chen, and Tiankai Li. When teachers mislead: Spurious-signal-aware on-policy distillation. *arXiv preprint arXiv:2608.03632*, 2026.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Aristotelis Lazaridis, Dylan Bates, Aman Sharma, Brian King, Vincent Lu, and Jack FitzGerald. EDGE-OPD: Internalizing privileged context with evidence guided on-policy distillation. *arXiv preprint arXiv:2605.23493*, 2026.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019. doi: 10.48550/arXiv.1711.05101.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026.

Meta. Llama-3.2-1B-Instruct model card. Hugging Face, 2024. URL `https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`. Released September 25, 2024. Accessed September 5, 2026.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 60808–60838, 2025.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 2025.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. ToolAlpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

Lev S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Jianan Xie, Xin Sun, Zhongqi Chen, Xing Zheng, Shu Wu, Bowen Song, and Liang Wang. EviSD: Evidence-conditioned self-distillation for search-augmented agents. *arXiv preprint arXiv:2608.01359*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026a.

Yi Yang, Cong Qin, Xiaodan Liu, Chishui Chen, Qing Dong, Yan Zhang, Cao Liu, Zhao Yang, Lu Pan, Jiaye Lin, and Yi Feng. Agentic reinforcement learning with observation-calibrated self-distillation. *arXiv preprint arXiv:2608.04788*, 2026b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Haibin Lin, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026a.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. ROSD: Reflective on-policy self-distillation for language model reasoning across domains. *arXiv preprint arXiv:2605.28014*, 2026b.

AI Use Statement

We used generative AI tools solely to assist with manuscript drafting and English-language editing. The authors critically reviewed and revised all AI-assisted content and made all final decisions regarding the content and presentation of the manuscript. The authors take full responsibility for the final version of this work.

Reproducibility Statement

We document the VERPO objective in Section 3 and provide the detailed training procedure in Algorithm 2. Appendix E presents the mathematical derivations and their assumptions. Appendix H specifies the benchmark sources and data splits, model configurations, training and evaluation settings, hyperparameters, and prompt templates. Appendix I reports the hardware settings and computational costs.

Ethics Statement

This work studies language-model post-training on existing scientific reasoning and tool-use benchmarks. Improved benchmark accuracy does not establish factual reliability or safe deployment. In particular, alignment with verifiable task rewards does not guarantee the correctness of intermediate reasoning or the safety of generated tool actions. Potential misuse includes generating misleading scientific explanations and automating harmful tool interactions. Deployment beyond the evaluated settings therefore requires independent validation, appropriate tool-access controls, and human oversight.

Appendix Contents

A Extended Related Work . . . 14

B Notation . . . 15

C Algorithm . . . 16

D Preliminaries . . . 16
- D.1 Autoregressive Policy . . . 16
- D.2 RLVR and Group-Relative Policy Optimization . . . 17
- D.3 Feedback-Conditioned Teacher Replay . . . 17
- D.4 Divergence and Local Information Geometry . . . 18

E Theoretical Analysis . . . 18
- E.1 Teacher Paths . . . 18
- E.2 Exact Forward-KL Decomposition . . . 19
- E.3 Reverse-KL Counterpart . . . 19
- E.4 Fixed, Contrastive, and Fisher Evidence Contrast . . . 20
- E.5 Ideal Token-Wise Controller . . . 21
- E.6 Benefit and Local-KL Cost Proxies . . . 22
- E.7 Surrogate Acceptance Optimization and Properties . . . 23
- E.8 Multiplicative Advantage Modulation . . . 23
- E.9 Complete Objective and Exact Logit Gradients . . . 24

F Additional Results . . . 24
- F.1 Joint Loss Weighting and Advantage Modulation . . . 25
- F.2 Trajectory-Level Correction Scope . . . 25
- F.3 Token-Level FEC and ZPD Weight Analysis . . . 26
- F.4 Hyperparameter Sensitivity . . . 28
- F.5 Probability Mass Retention under Top-K Truncation . . . 28

G Diagnostic Metrics . . . 29
- G.1 ZPD Effective Weight Coverage . . . 29
- G.2 Mean Signed Benefit and Fisher Cost . . . 30
- G.3 FEC Residual Nuisance Fisher Covariance . . . 30

H Experimental Details . . . 30
- H.1 Datasets and Data Processing . . . 30
- H.2 Models, Baselines, and Runtime . . . 31
- H.3 Shared Training Setup . . . 31
- H.4 Method-Specific Hyperparameters . . . 32
- H.5 Prompt Templates . . . 33

I Computational Analysis . . . 35
- I.1 Per-Step Cost . . . 35
- I.2 Compute-Matched Performance . . . 35

J Case Study . . . 36

A Extended Related Work

This appendix expands the discussion in Section 2, covering outcome-based credit assignment, self-distillation, and selective privileged supervision in greater detail.

RLVR and Fine-grained Credit Assignment. Reinforcement learning with verifiable rewards (RLVR) optimizes language models using automatically checkable outcome rewards. Critic-free methods such as GRPO (Shao et al., 2024) replace traditional actor–critic value estimates with group-relative advantages computed from multiple on-policy responses. DeepSeek-R1 (DeepSeek-AI, 2025) and DAPO (Yu et al., 2025) use related group-relative updates for scalable RLVR. These sequence-level advantages are shared across response tokens, obscuring which local decisions caused the outcome. Process supervision obtains finer-grained signals through human-annotated step verification in Let's Verify Step by Step (Lightman et al., 2024), automatically constructed labels in Math-Shepherd (Wang et al., 2024), and outcome-derived progress estimates in Rewarding Progress (Setlur et al., 2025). Such approaches require human annotations or automatically generated process targets together with a learned verifier, whose local preferences need not match the verifiable sequence-level objective. VERPO retains the outcome-derived GRPO advantage as the task-level anchor and uses privileged Teacher replay to propose token-level corrections, accepting them only when they align with the local reward direction at justified Fisher movement cost.

On-policy Distillation and Self-distillation. Knowledge distillation transfers a teacher's predictive distribution to a student (Hinton et al., 2015). On-policy distillation reduces off-policy distribution shift by supervising student-generated trajectories with an external teacher (Agarwal et al., 2023). MiniLLM develops a related distribution-matching formulation for language-model distillation (Gu et al., 2023). OPSD removes the external teacher by conditioning the same model on privileged solutions (Zhao et al., 2026a), and SDPO constructs a feedback-conditioned self-teacher for dense token-level supervision (Hübotter et al., 2026). SRPO anchors self-distillation to verifiable outcomes through sample-level routing (Li et al., 2026). RLSD modulates the magnitude of self-distillation with a reward-directed signal (Yang et al., 2026a), and EviSD applies analogous reward-anchored modulation to evidence-conditioned search actions (Xie et al., 2026). RLCSD uses contrastive privileged contexts to shape its self-distillation signal (Pan et al., 2026). These methods do not jointly separate evidence-free reference retention from signed evidence-induced change, and they do not decide whether an individual correction is worth applying at its local policy cost. VERPO addresses this control gap with a stopped token-wise ZPD controller that accepts and scales each correction according to its alignment with the local GRPO direction and its Fisher movement cost.

Selective and Contrastive Privileged Supervision. Empirical studies show that privileged supervision is not uniformly reliable across rollouts or tokens (Armandpour et al., 2026; Harne et al., 2026; Lazaridis et al., 2026; Zhao et al., 2026b; Kaur et al., 2026; Jiang et al., 2026). Teacher guidance is better aligned on incorrect than on already-correct trajectories (Armandpour et al., 2026), while privileged conditioning can bias the Teacher toward a particular reference trajectory rather than toward correctness in general (Harne et al., 2026), induce behavior unrelated to the intended correction (Lazaridis et al., 2026), overwrite valid reasoning prefixes (Zhao et al., 2026b), suppress verification and backtracking in thinking models (Kaur et al., 2026), or amplify weakly input-grounded signals (Jiang et al., 2026). RLCSD addresses privilege-induced style drift by contrasting self-teachers conditioned on correct and incorrect hints (Pan et al., 2026). CEPO similarly exploits correct-wrong evidence for contrastive token credit (Heakl et al., 2026), while OCSD uses structurally matched privileged and ablated replay views to remove replay-scaffold effects in agentic RL (Yang et al., 2026b). These contrastive constructions reduce shared effects between privileged branches, but a Teacher difference can still contain a component caused by the mere presence of privileged evidence. Such a component can leak evidence-specific style, confidence, or formatting into the Student without improving task success. VERPO's Fisher Evidence Contrast (FEC) estimates this evidence-presence direction and projects it out of the task contrast under the local Fisher metric, reducing privileged-context leakage before the resulting correction is applied.

B Notation

Table 5 aligns the notation in Section 3, Algorithm 2 and Algorithm 1 , and the derivations. Symbols are grouped by their operational role. Algebraic dummy variables are defined locally.

Table 5: Notation used throughout the method and appendices.

Symbol	Definition
Data, policies, and rollout quantities	
$x_i, y_i, y_{i,<t}, t, v, \mathcal{V}$	Prompt, sampled response, response prefix, token position, vocabulary item, and vocabulary.
$\mathcal{G}(i)$	Rollout group containing responses for the same prompt as i.
$R_i, \widehat{A}_i$	Scalar outcome reward and group-relative sequence advantage.
$\theta, \theta_{\text{old}}, \bar{\theta}$	Current Actor, frozen behavior Actor, and exponential-moving-average Teacher parameters.
$p_{i,t}, p_{i,t}^{\text{old}}, z_{i,t}$	Current/behavior token distributions and current Actor logits at prefix $y_{i,<t}$.
$\rho_{i,t}(\theta)$	Token-level importance ratio in GRPO (Shao et al., 2024).
Teacher branches and evidence paths	
$q_{i,t}^{(c)}, c$	Frozen Teacher distribution and its replay/context condition.
$q_{i,t}^{\text{ref}}$	Evidence-free reference Teacher used for policy restoration.
$q_{i,t}^{0}, q_{i,t}^{e}$	Teacher distributions without and with task evidence.
$q_{i,t}^{+}, q_{i,t}^{-}$	Correct-evidence Teacher and incorrect-evidence Teachers.
$q_{b,i,t}, q_{e,i,t}$	Base and target endpoints of a selected evidence comparison.
$q_{w,t}^{\text{FKL}}, q_{w,t}^{\text{RKL}}, Z_t(w)$	Arithmetic/geometric Teacher path and its RKL partition function.
$\Delta_{i,t}, r_{i,t}$	Probability displacement $q_{e,i,t} - q_{b,i,t}$ and log-density contrast $\log q_{e,i,t} - \log q_{b,i,t}$.
$F(p_{i,t}), G_{i,t}$	Categorical Fisher matrix $\text{diag}(p_{i,t}) - p_{i,t}p_{i,t}^{\top}$ and logit Jacobian $\partial z_{i,t}/\partial\theta$.
$u_{i,t}$	Evidence logit direction: $\Delta_{i,t}$ for FKL and $F(p_{i,t})r_{i,t}$ for RKL.
Direction constructions and FEC	
$\Delta_t^{\text{task}}, \Delta_t^{\text{nuis}}, \Delta_t^{\text{FEC}}$	Zero-sum FKL task, nuisance, and residual directions. The task direction equals Δ_t^{CTR}.
$r_t^{\text{task}}, r_t^{\text{nuis}}, r_t^{\text{FEC}}$	RKL task, nuisance, and Fisher-residual log-density contrasts.
$u_t^{\text{task}}, u_t^{\text{nuis}}, u_t^{\text{FEC}}$	Fisher-mapped task, nuisance, and residual logit directions.
$q_t^{\text{pres,RKL}}, \alpha_t^{\text{FKL}}, \alpha_t^{\text{RKL}}$	RKL evidence-presence midpoint and geometry-specific FEC coefficients.
ϵ_{proj}	Ridge used in the Fisher projection denominator.
Masks, objectives, and normalization	
$a_{i,t}, m_{i,t}$	Binary valid-response-token and admissible evidence masks, with $m_{i,t} \le a_{i,t}$.
N, Z_{evi}	Response-token count and masked evidence-term normalizer.
$\mathcal{L}_{\text{GRPO}}, \mathcal{L}_{\text{ref}}, \mathcal{L}_{\text{evi}}, \mathcal{L}_{\text{VERPO}}$	Outcome, reference, evidence, and complete losses.
$\chi_{\text{OPSD}}, \chi_{\text{adv}}$	Binary switches for signed evidence correction and advantage modulation.
$\lambda_{\text{ref}}, \lambda_{\text{evi}}$	Coefficients for reference restoration and evidence correction.
FKL, RKL, FEC	Forward-KL geometry, reverse-KL geometry, and Fisher Evidence Contrast.
ZPD controller and derivation quantities	
$d_{i,t}, \eta, J_{\text{RL}}$	Parameter-space evidence direction, local step size, and RL objective used by controller.
θ_{base}	Hypothetical Actor state after the outcome and reference updates.
$B_{i,t}, C_{i,t}, w_{i,t}^{\text{ideal}}$	Directional benefit, curvature cost, and ideal token-wise acceptance.
$h_{i,t}, c_{i,t}, w_{i,t}$	Alignment proxy, local-KL movement proxy, and stopped acceptance weight.
$\alpha_{\text{cost}}, \epsilon_{\text{cost}}$	Fisher-cost coefficient and positive denominator floor used by the ZPD controller.

C Algorithm

Algorithm 2 VERPO Actor Update under FKL Geometry (Detailed)

Require: Current Actor π_θ, EMA Teacher $\pi_{\tilde{\theta}}$, dataset $\mathcal{D}$, variant $s \in \{\text{Fix}, \text{CTR}, \text{FEC}\}$, switches $\chi_{\text{OPSD}}, \chi_{\text{adv}}$
Ensure: Updated Actor π_θ

repeat
  Stage 1: Rollout Sampling and Outcome Signal
  Sample a prompt batch from $\mathcal{D}$; freeze $\theta_{\text{old}} \leftarrow \theta$.
  $\{y_i\}_{i \in \mathcal{G}(x)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ for each sampled prompt x.
  Obtain verifier rewards R_i; set $\widehat{A}_i \leftarrow R_i - |\mathcal{G}(i)|^{-1} \sum_{j \in \mathcal{G}(i)} R_j$. ▷ Outcome anchor
  Stage 2: Token-Level Evidence Correction
  Score sampled prefixes with the current Actor under the original prompt.
  Replay the same prefixes under selected evidence contexts with the Teacher held fixed.
  $p_{i,t} \leftarrow \pi_\theta(\cdot \mid x_i, y_{i,<t})$. ▷ Original-prompt distribution
  Obtain stopped Teacher distributions $\{q^{\text{ref}}_{i,t}, q^{0}_{i,t}, q^{+}_{i,t}, q^{-}_{i,t}\}$ via Eq. (D.6). ▷ Same-prefix replay
  $\Delta_{i,t} \leftarrow \begin{cases} q^{+}_{i,t} - q^{0}_{i,t}, & s = \text{Fix}, \\ q^{+}_{i,t} - q^{-}_{i,t}, & s = \text{CTR}, \text{ with } \Delta^{\text{FEC}}_{i,t} \text{ from Eq. (9).} \\ \Delta^{\text{FEC}}_{i,t}, & s = \text{FEC}, \end{cases}$ ▷ Signed correction
  Stage 3: ZPD Weight Estimation
  $\Delta_{i,t} \leftarrow \text{sg}[\Delta_{i,t}], \quad u_{i,t} \leftarrow \Delta_{i,t}, \quad h_{i,t} \leftarrow [\widehat{A}_i (e_{y_{i,t}} - p_{i,t})^\top u_{i,t}]_+$. ▷ GRPO alignment
  $c_{i,t} \leftarrow u_{i,t}^\top F(p_{i,t}) u_{i,t}$. ▷ Fisher movement cost
  Form valid-token mask $a_{i,t}$ and evidence mask $m_{i,t} \le a_{i,t}$.
  $w^*_{i,t} \leftarrow \text{sg}\left[\frac{h_{i,t}}{h_{i,t} + \alpha_{\text{cost}} c_{i,t} + \epsilon_{\text{cost}}}\right], \quad \widetilde{w}_{i,t} \leftarrow m_{i,t} w^*_{i,t}$. ▷ Token-wise ZPD gate
  Stage 4: Two-Path Policy Update
  Set $\widehat{A}'_{i,t} \leftarrow \widehat{A}_i [1 + \chi_{\text{adv}} \lambda_{\text{adv}} \text{sg}(\widetilde{w}_{i,t})]$. ▷ Advantage modulation
  Set $N \leftarrow \sum_{i,t} a_{i,t}$ and $Z_{\text{evi}} \leftarrow \max(1, \sum_{i,t} a_{i,t} m_{i,t})$. ▷ Separate normalizers
  Compute reference and evidence losses via Eqs. (2) and (3).
  $\mathcal{L}^{\text{FKL}}_{\text{total}} \leftarrow \mathcal{L}_{\text{GRPO}}(\widehat{A}') + \lambda_{\text{ref}} \mathcal{L}^{\text{FKL}}_{\text{ref}} + \chi_{\text{OPSD}} \lambda_{\text{evi}} \mathcal{L}^{\text{FKL}}_{\text{OPSD}}(w^*)$. ▷ Evidence-loss path
  $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}^{\text{FKL}}_{\text{total}}$.
  Update the EMA Teacher via Eq. (D.7). ▷ After the Actor update
until stopping criterion is reached
return π_θ

D Preliminaries

To specify the quantities used by Algorithm 2, this appendix defines rollout normalization, the outcome objective, frozen Teacher replay, and local KL geometry. These conventions support the loss decompositions and controller approximations in Appendix E.

D.1 Autoregressive Policy

To compare Actor and Teacher predictions at the same positions, we index both distributions by the sampled response prefix. For each prompt, the frozen behavior Actor $\pi_{\theta_{\text{old}}}$ samples a group of responses. Index i ranges over responses, and $\mathcal{G}(i)$ denotes the responses generated for the same prompt as response i. The current Actor factorizes as

$$\pi_\theta(y_i \mid x_i) = \prod_{t=1}^{T_i} p_{i,t}(y_{i,t}), \qquad p_{i,t} = \pi_\theta(\cdot \mid x_i, y_{i,<t}). \tag{D.1}$$

Each factor scores one observed token rather than a new Teacher completion. When padded tensors are used, $a_{i,t} \in \{0, 1\}$ is the valid-response-token mask. Analytic sums over $t \le T_i$ are equivalently restricted to $a_{i,t} = 1$. The binary evidence mask satisfies $m_{i,t} \le a_{i,t}$, and the normalizers are

$$N = \sum_{i,t} a_{i,t} = \sum_i T_i, \qquad Z_{\text{evi}} = \max\left(1, \sum_{i,t} a_{i,t} m_{i,t}\right). \tag{D.2}$$

We assume a batch contains at least one valid response token, so $N > 0$. The reference and outcome losses average over these tokens. The evidence loss averages over eligible tokens without renormalizing by acceptance weight. If none are eligible, its numerator vanishes and the floor keeps its normalizer defined.

To measure reward alignment, we use the unclipped local GRPO (Shao et al., 2024) score direction. For Actor logits $z_{i,t}$ and sampled-token one-hot vector $e_{y_{i,t}}$, this direction is

$$g^{\mathrm{RL}}_{i,t} := \widehat{A}_i \nabla_{z_{i,t}} \log p_{i,t}(y_{i,t}) = \widehat{A}_i (e_{y_{i,t}} - p_{i,t}). \tag{D.3}$$

This vector converts the sequence advantage into a token-local comparison direction for ZPD. It is not the gradient of the full clipped training loss, which also depends on importance ratios and clipping.

D.2 RLVR and Group-Relative Policy Optimization

To obtain the outcome signal used in Section 3.1, we evaluate each sampled response with a scalar verifier reward R_i. Our GRPO (Shao et al., 2024) variant centers this reward within its prompt group without dividing by the within-group standard deviation:

$$\widehat{A}_i = R_i - \frac{1}{|\mathcal{G}(i)|} \sum_{j \in \mathcal{G}(i)} R_j. \tag{D.4}$$

The advantage measures performance relative to responses to the same prompt and is broadcast across the response tokens. A group with identical rewards has zero advantage for every response.

To account for the difference between the current and behavior policies, define the sampled-token importance ratio as

$$\rho_{i,t}(\theta) = \frac{p_{i,t}(y_{i,t})}{p^{\mathrm{old}}_{i,t}(y_{i,t})}$$

The denominator remains fixed during the Actor update. With asymmetric clipping radii $\epsilon_{\mathrm{low}}, \epsilon_{\mathrm{high}} > 0$, the outcome loss is

$$\mathcal{L}_{\mathrm{GRPO}} = -\frac{1}{N} \sum_{i,t} a_{i,t} \min\Big\{ \rho_{i,t}(\theta)\widehat{A}_i,\ \mathrm{clip}\big(\rho_{i,t}(\theta), 1 - \epsilon_{\mathrm{low}}, 1 + \epsilon_{\mathrm{high}}\big)\, \widehat{A}_i \Big\}. \tag{D.5}$$

The minimum limits the incentive to move a sampled-token ratio farther in the advantage-favored direction after crossing its clipping threshold. The valid-token mask excludes padding, and N fixes the averaging scale. Section 3.5 changes the token advantage in this loss while retaining the same ratio and clipping rule.

D.3 Feedback-Conditioned Teacher Replay

To construct evidence comparisons without changing the sampled trajectory, we replay its prefixes under different Teacher contexts. SDPO (Hübotter et al., 2026) and SRPO (Li et al., 2026) also use feedback-conditioned replay for dense supervision. VERPO represents each context condition c through the stopped replay operator

$$q^{(c)}_{i,t} = \mathrm{sg}\big[\pi_{\bar{\theta}}\big(\cdot \mid \mathrm{reprompt}(x_i, c), y_{i,<t}\big)\big]. \tag{D.6}$$

The reprompt changes the Teacher's evidence context, while the Actor prefix $y_{i,<t}$ stays fixed. Teacher forcing therefore scores the original trajectory without generating a replacement completion. The stop-gradient operator prevents this supervision branch from receiving Actor-loss gradients.

The Teacher parameters $\bar{\theta}$ are an exponential-moving-average copy of the Actor across optimizer steps,

$$\bar{\theta} \leftarrow \gamma\bar{\theta} + (1 - \gamma)\theta, \qquad 0 < \gamma < 1, \tag{D.7}$$

Here γ retains the previous Teacher parameters and $1 - \gamma$ weights the newly updated Actor. The Teacher changes between optimizer steps but remains frozen within each Actor update. Its parameters $\bar{\theta}$ are distinct from the hypothetical Actor state θ_{base} used only in the controller derivation.

Each replay output is a full vocabulary distribution, rather than only the probability of the sampled token. Comparing these distributions at a shared prefix isolates the change in Teacher predictions induced by the chosen context condition.

D.4 Divergence and Local Information Geometry

To distinguish the two loss geometries, we fix the argument order of the KL divergence (Kullback & Leibler, 1951). For positive, normalized distributions over the vocabulary, our conventions are

$$\mathrm{KL}(q\|p) = \sum_{v\in\mathcal{V}} q(v)\log\frac{q(v)}{p(v)}, \qquad D_{\mathrm{FKL}}(q,p) = \mathrm{KL}(q\|p), \qquad D_{\mathrm{RKL}}(p,q) = \mathrm{KL}(p\|q). \tag{D.8}$$

FKL averages the log-density ratio under the Teacher, whereas RKL averages it under the Actor. We require positive support wherever these logarithms are evaluated. Finite-logit softmax distributions satisfy this assumption in exact arithmetic. The finite-support implementation is a separate approximation described in Appendix H.4.

Let $z_{i,t} \in \mathbb{R}^{|\mathcal{V}|}$ be the Actor logits and $p_{i,t} = \mathrm{softmax}(z_{i,t})$. For vocabulary items $u, v \in \mathcal{V}$, the softmax score identity is

$$\frac{\partial \log p_{i,t}(u)}{\partial z_{i,t}(v)} = \mathbf{1}[u = v] - p_{i,t}(v). \tag{D.9}$$

The Jacobian of the probabilities $p_{i,t}$ with respect to the logits, not the Jacobian of their logarithms, gives the categorical Fisher matrix. For an arbitrary vocabulary vector b,

$$F(p_{i,t}) = \mathrm{diag}(p_{i,t}) - p_{i,t}p_{i,t}^{\top}, \qquad F(p_{i,t})\mathbf{1} = 0, \qquad b^{\top}F(p_{i,t})b = \mathrm{Var}_{V\sim p_{i,t}}[b(V)] \geq 0. \tag{D.10}$$

Thus constant shifts of a log-density contrast do not change its induced logit direction. A small logit displacement $z_{i,t} \mapsto z_{i,t} + \epsilon b$ induces the local policy movement

$$\mathrm{KL}\big(p_{i,t}\|\,\mathrm{softmax}(z_{i,t} + \epsilon b)\big) = \tfrac{1}{2}\epsilon^2 b^{\top}F(p_{i,t})b + O(\epsilon^3). \tag{D.11}$$

For fixed finite logits and direction b, the quadratic term measures movement of the Actor distribution to second order as ϵ tends to zero. It supplies the controller's nonnegative movement proxy, not a guarantee of reward improvement for a finite optimizer step.

E Theoretical Analysis

This appendix provides the complete derivations behind the signed evidence directions and token-wise controller in Section 3, using the RLVR, Teacher-replay, and local-geometric conventions collected in Appendix D. We first establish exact loss identities, then distinguish the local approximations and chosen surrogate used to construct the controller. Teacher distributions, evidence masks, acceptance weights, and constructed FKL directions or RKL contrasts are stopped during Actor differentiation. This includes the Actor-dependent FEC projection coefficients.

E.1 Teacher Paths

To vary evidence strength while preserving a distribution, we connect two Teacher endpoints at the same rollout prefix $y_{<t}$. Only the evidence condition changes. We assume normalized distributions on a finite vocabulary and positive Actor probability on the Teacher support. For FKL, normalized endpoints imply $\mathbf{1}^{\top}\Delta_t = 0$ for $\Delta_t = q_{e,t} - q_{b,t}$. The arithmetic path is

$$q_{w,t}^{\mathrm{FKL}} = (1-w)q_{b,t} + wq_{e,t} = q_{b,t} + w\Delta_t, \qquad 0 \leq w \leq 1. \tag{E.1}$$

The weight interpolates from the base endpoint to the enriched endpoint. Convex mixing preserves normalization and nonnegativity for $0 \leq w \leq 1$, while Δ_t records a signed probability displacement.

For RKL, assume positive support for both endpoints and define

$$r_t(v) = \log\frac{q_{e,t}(v)}{q_{b,t}(v)}. \tag{E.2}$$

This contrast measures relative density rather than an additive probability change. To interpolate in these log-density coordinates, define

$$q_{w,t}^{\mathrm{RKL}}(v) = \frac{q_{b,t}(v)^{1-w}q_{e,t}(v)^{w}}{Z_t(w)} = \frac{q_{b,t}(v)e^{wr_t(v)}}{Z_t(w)}, \quad Z_t(w) = \sum_{u\in\mathcal{V}} q_{b,t}(u)e^{wr_t(u)}. \tag{E.3}$$

The partition function $Z_t(w)$ normalizes the exponentially tilted base. Both endpoints are recovered at $w = 0$ and $w = 1$. Unlike the arithmetic path, the intermediate distributions mix log densities before normalization.

E.2 Exact Forward-KL Decomposition

To recover the separation in Equation (1), we expand the FKL objective with the Teacher endpoints and w held fixed. We suppress the token index and assume positive Actor probability wherever the Teacher assigns mass. The loss for this single prefix is

$$\mathcal{L}_{\mathrm{FKL}}(\theta; w) = \mathrm{KL}(q_w^{\mathrm{FKL}} \| p_\theta). \tag{E.4}$$

Only its cross-entropy term depends on the Actor. Substituting $q_w^{\mathrm{FKL}} = q_b + w\Delta$ into that term gives

$$\begin{aligned}\mathcal{L}_{\mathrm{FKL}}(\theta; w) &= \sum_{v \in \mathcal{V}} q_w^{\mathrm{FKL}}(v) \log q_w^{\mathrm{FKL}}(v) - \sum_{v \in \mathcal{V}} q_w^{\mathrm{FKL}}(v) \log p_\theta(v) \\ &= \mathrm{KL}(q_b \| p_\theta) - w \sum_{v \in \mathcal{V}} \Delta(v) \log p_\theta(v) + C(q_w^{\mathrm{FKL}}, q_b),\end{aligned}$$

where

$$C(q_w^{\mathrm{FKL}}, q_b) = \sum_v q_w^{\mathrm{FKL}}(v) \log q_w^{\mathrm{FKL}}(v) - \sum_v q_b(v) \log q_b(v).$$

Thus the exact decomposition is

$$\boxed{\mathrm{KL}(q_w^{\mathrm{FKL}} \| p) = \mathrm{KL}(q_b \| p) - w \sum_{v \in \mathcal{V}} \Delta(v) \log p(v) + C(q_w^{\mathrm{FKL}}, q_b).} \tag{E.5}$$

The first term restores the base distribution, and the second isolates the signed evidence correction. The remainder C contains only Teacher entropy terms and has zero Actor gradient under the stated stop-gradient convention. This is an exact identity, without a small-step approximation.

To verify the correction sign, differentiate with respect to the logits of $p = \mathrm{softmax}(z)$ using

$$\frac{\partial \log p(u)}{\partial z(v)} = \mathbf{1}[u = v] - p(v).$$

Therefore

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathrm{FKL}}}{\partial z(v)} &= -\sum_{u \in \mathcal{V}} q_w^{\mathrm{FKL}}(u) \big(\mathbf{1}[u = v] - p(v)\big) \\ &= p(v) - q_w^{\mathrm{FKL}}(v),\end{aligned}$$

and the exact logit descent direction decomposes as

$$\boxed{-\nabla_z \mathcal{L}_{\mathrm{FKL}} = \underbrace{q_b - p}_{\text{base restoration}} + \underbrace{w\Delta}_{\text{evidence correction}}.} \tag{E.6}$$

The base contribution points from the Actor toward q_b, while $w\Delta$ adds the accepted Teacher displacement. Retaining only this second channel gives the signed evidence loss derived from OPSD (Zhao et al., 2026a):

$$\mathcal{L}_{\mathrm{evi}}^{\mathrm{FKL}}(w) = -w \sum_{v \in \mathcal{V}} \Delta(v) \log p(v), \qquad -\nabla_z \mathcal{L}_{\mathrm{evi}}^{\mathrm{FKL}} = w\Delta. \tag{E.7}$$

The zero-sum property of Δ cancels the softmax normalization term in the gradient. A separate reference loss can therefore use q^{ref} even when the evidence comparison starts at a different base q_b.

E.3 Reverse-KL Counterpart

To obtain the counterpart of the FKL separation, we substitute the geometric Teacher path into RKL. Assume positive Actor and Teacher distributions on the vocabulary, with r and w stopped. Here the Actor supplies the averaging weights, so a fixed Teacher contrast still produces an Actor-dependent correction. This differs from the FKL evidence gradient, whose direction is the fixed probability displacement between the endpoints. Stopping the Teacher contrast therefore leaves gradients through the Actor's expectation intact, while preventing differentiation through the evidence construction itself.

Expanding the normalized path gives

$$\begin{aligned}\mathcal{L}_{\mathrm{RKL}}(\theta; w) &= \mathrm{KL}(p_\theta \| q_w^{\mathrm{RKL}}) \\ &= \sum_{v \in \mathcal{V}} p(v)\,[\log p(v) - \log q_b(v) - w r(v) + \log Z(w)] \\ &= \mathrm{KL}(p\|q_b) - w\mathbb{E}_{v\sim p}[r(v)] + \log Z(w).\end{aligned}$$

Hence

$$\boxed{\mathrm{KL}(p\|q_w^{\mathrm{RKL}}) = \mathrm{KL}(p\|q_b) - w\mathbb{E}_{v\sim p}[r(v)] + \log Z(w).} \tag{E.8}$$

The base term retains the original Teacher anchor. The contrast term changes the Actor's expected log-density preference, and $\log Z(w)$ completes the loss identity without contributing an Actor gradient.

To identify the resulting logit direction, apply the probability Jacobian $F(p)$ to the derivative of RKL. For a fixed positive Teacher q,

$$\begin{aligned}\nabla_z \mathrm{KL}(p\|q) &= F(p)\big(\log p - \log q + \mathbf{1}\big) \\ &= F(p)(\log p - \log q),\end{aligned}$$

because $F(p)\mathbf{1} = 0$. Substituting $\log q_w^{\mathrm{RKL}} = \log q_b + wr - \log Z(w)$ yields

$$\boxed{-\nabla_z \mathcal{L}_{\mathrm{RKL}} = \underbrace{F(p)(\log q_b - \log p)}_{\text{base restoration}} + \underbrace{w\xi}_{\text{evidence correction}}, \qquad \xi = F(p)r.} \tag{E.9}$$

Here ξ denotes the RKL evidence direction, written as u in the unified controller. Unlike the FKL displacement, it depends on the Actor through $F(p)$. Writing $\bar{r} = \mathbb{E}_{v\sim p}[r(v)]$ exposes this dependence:

$$\xi(v) = p(v)\big(r(v) - \bar{r}\big), \qquad \mathbf{1}^\top \xi = 0. \tag{E.10}$$

The Fisher map centers the log contrast under the Actor and weights each coordinate by its current probability. Thus ξ sums to zero but is not the probability displacement $q_e - q_b$. Its signed loss is

$$\mathcal{L}_{\mathrm{evi}}^{\mathrm{RKL}}(w) = \log Z(w) - w\mathbb{E}_{v\sim p}[r(v)], \qquad -\nabla_z \mathcal{L}_{\mathrm{evi}}^{\mathrm{RKL}} = wF(p)r. \tag{E.11}$$

Differentiation acts through the expectation under p, not through the stopped contrast. The resulting $F(p)r$ supplies the direction used by the RKL controller and the complete gradient in Appendix E.9.

E.4 Fixed, Contrastive, and Fisher Evidence Contrast

To instantiate the evidence channel in Section 3.4, we specify the endpoints and the FEC residual in both geometries. All branches replay the same prefix. The RKL constructions require positive Teacher support, and every projection is computed at the current Actor distribution before its coefficient is stopped.

Fixed Evidence. The Fixed construction uses the evidence-conditioned replay of SDPO (Hübotter et al., 2026) to define the endpoints $q_b = q^0$ and $q_e = q^+$:

$$\Delta_t^{\mathrm{Fix}} = q_t^+ - q_t^0, \qquad r_t^{\mathrm{Fix}} = \log q_t^+ - \log q_t^0, \qquad u_t^{\mathrm{Fix,RKL}} = F(p_t) r_t^{\mathrm{Fix}}. \tag{E.12}$$

This comparison includes both the change associated with correct evidence and the change associated with presenting evidence at all. The observed reward hacking for this tested direction is reported with Table 4(a), not implied by this identity.

Contrastive Evidence. The endpoint comparison follows RLCSD (Pan et al., 2026). For $K \geq 1$ incorrect-evidence Teachers, define $q_t^- = K^{-1}\sum_{k=1}^K q_t^{-,k}$. The construction uses $q_b = q^-$ and $q_e = q^+$, so

$$\Delta_t^{\mathrm{CTR}} = q_t^+ - q_t^-, \qquad r_t^{\mathrm{CTR}} = \log q_t^+ - \log q_t^-, \qquad u_t^{\mathrm{CTR,RKL}} = F(p_t) r_t^{\mathrm{CTR}}. \tag{E.13}$$

The incorrect branches are averaged in probability space before taking a logarithm for RKL. Subtraction cancels additive shifts shared by the two endpoints, but it retains context effects that differ between them.

Forward-KL FEC. For normalized Teacher branches and a fixed Actor Fisher matrix, define the task contrast and evidence-presence proxy. The target claim concerns the residual Fisher inner product,

without assuming that the proxy contains only task-irrelevant information. In probability coordinates, the directions are

$$\Delta_t^{\text{task}} = q_t^+ - q_t^-, \qquad \Delta_t^{\text{nuis}} = \tfrac{1}{2}(q_t^+ + q_t^-) - q_t^0.$$

Both vectors sum to zero. The task vector equals Δ_t^{CTR} in the main text, while the midpoint comparison measures evidence presence. For the Fisher inner product $\langle a, b\rangle_{F(p_t)} = a^\top F(p_t) b$, the projection coefficient and residual are

$$\boxed{\alpha_t^{\text{FKL}} = \frac{\langle \Delta_t^{\text{task}}, \Delta_t^{\text{nuis}}\rangle_{F(p_t)}}{\langle \Delta_t^{\text{nuis}}, \Delta_t^{\text{nuis}}\rangle_{F(p_t)} + \epsilon_{\text{proj}}}, \qquad \Delta_t^{\text{FEC}} = \Delta_t^{\text{task}} - \alpha_t^{\text{FKL}} \Delta_t^{\text{nuis}}.} \tag{E.14}$$

The coefficient α_t^{FKL} is the FKL coefficient denoted α^{FEC} in Section 3.4. It subtracts a component along the measured nuisance direction. To determine what regularization leaves behind, substitute this coefficient into the residual inner product:

$$\left\langle \Delta_t^{\text{FEC}}, \Delta_t^{\text{nuis}} \right\rangle_{F(p_t)} = \frac{\epsilon_{\text{proj}}}{\langle \Delta_t^{\text{nuis}}, \Delta_t^{\text{nuis}}\rangle_{F(p_t)} + \epsilon_{\text{proj}}} \left\langle \Delta_t^{\text{task}}, \Delta_t^{\text{nuis}} \right\rangle_{F(p_t)}. \tag{E.15}$$

For $\epsilon_{\text{proj}} > 0$, the remaining inner product equals the original one times a factor in $(0, 1]$. Exact orthogonality follows in the unregularized case only when the denominator is nonzero. The residual still sums to zero, but adding it to a base distribution need not preserve nonnegativity. We therefore use the stopped residual directly in the signed loss:

$$\mathcal{L}_{\text{FEC},t}^{\text{FKL}} = -w_t \sum_{v \in \mathcal{V}} \Delta_t^{\text{FEC}}(v) \log p_t(v), \qquad -\nabla_{z_t} \mathcal{L}_{\text{FEC},t}^{\text{FKL}} = w_t \Delta_t^{\text{FEC}}. \tag{E.16}$$

The stated gradient requires stopping the entire residual, including its Actor-dependent coefficient. It recovers the direction in Equation (9) without claiming a valid arithmetic Teacher endpoint for FEC.

Reverse-KL FEC. To apply the same projection idea in RKL geometry, first measure evidence presence using positive Teacher branches and the normalized geometric midpoint

$$q_t^{\text{pres,RKL}}(v) = \frac{\sqrt{q_t^+(v) q_t^-(v)}}{\sum_{u \in \mathcal{V}} \sqrt{q_t^+(u) q_t^-(u)}}$$

to measure evidence presence in log-density coordinates. Define

$$r_t^{\text{task}} = \log q_t^+ - \log q_t^-, \qquad u_t^{\text{task}} = F(p_t) r_t^{\text{task}},$$
$$r_t^{\text{nuis}} = \log q_t^{\text{pres,RKL}} - \log q_t^0, \qquad u_t^{\text{nuis}} = F(p_t) r_t^{\text{nuis}}.$$

These log contrasts are mapped into logit directions before projection. The midpoint normalization does not change u_t^{nuis} because $F(p_t)\mathbf{1} = 0$. Projecting in the same Fisher inner product gives

$$\boxed{\alpha_t^{\text{RKL}} = \frac{(u_t^{\text{task}})^\top F(p_t) u_t^{\text{nuis}}}{(u_t^{\text{nuis}})^\top F(p_t) u_t^{\text{nuis}} + \epsilon_{\text{proj}}}, \quad r_t^{\text{FEC}} = r_t^{\text{task}} - \alpha_t^{\text{RKL}} r_t^{\text{nuis}}, \quad u_t^{\text{FEC}} = F(p_t) r_t^{\text{FEC}}.} \tag{E.17}$$

Linearity of $F(p_t)$ gives $u_t^{\text{FEC}} = u_t^{\text{task}} - \alpha_t^{\text{RKL}} u_t^{\text{nuis}}$. Its residual Fisher inner product has the same ridge-dependent form as Equation (E.15), with u replacing Δ. We stop the constructed contrast r_t^{FEC}, including α_t^{RKL}, before Actor differentiation. Using q_t^- as the base then defines

$$q_{w,t}^{\text{RKL,FEC}}(v) = \frac{q_t^-(v) e^{w_t r_t^{\text{FEC}}(v)}}{Z_t^{\text{FEC}}(w_t)}, \qquad Z_t^{\text{FEC}}(w) = \sum_{u \in \mathcal{V}} q_t^-(u) e^{w r_t^{\text{FEC}}(u)}. \tag{E.18}$$

Exponential tilting preserves a positive normalized distribution for finite contrasts. Its endpoint at $w = 1$ need not equal q_t^+ after projection. The corresponding signed correction is

$$\mathcal{L}_{\text{FEC},t}^{\text{RKL}} = \log Z_t^{\text{FEC}}(w_t) - w_t \mathbb{E}_{v \sim p_t}[r_t^{\text{FEC}}(v)], \qquad -\nabla_{z_t} \mathcal{L}_{\text{FEC},t}^{\text{RKL}} = w_t u_t^{\text{FEC}}. \tag{E.19}$$

The Fisher map in this gradient comes from differentiating the Actor expectation. No gradient flows through the projection construction. For controller evaluation, the resulting direction and acceptance weight are treated as fixed inputs to the Actor update.

E.5 Ideal Token-Wise Controller

To connect acceptance to local reward improvement, we analyze one fixed evidence direction at a time. Let $G_{i,t} = \partial z_{i,t} / \partial \theta$ and $d_{i,t} = G_{i,t}^\top u_{i,t}$ be its parameter-space descent direction. The hypothetical Actor state θ_{base} absorbs the outcome and reference updates for this local analysis. It differs from the

EMA Teacher state $\bar{\theta}$ and does not require separate optimizer steps. At a fixed local step size $\eta > 0$, applying the correction with strength w produces

$$\theta^+(w) = \theta_{\text{base}} + \eta w d_{i,t}. \tag{E.20}$$

This defines a one-dimensional family of candidate Actor states rather than an additional training loop. The ideal weight selects the state with the largest RL objective:

$$\boxed{w_{i,t}^{\text{ideal}} = \arg\max_{0 \le w \le 1} J_{\text{RL}}(\theta_{\text{base}} + \eta w d_{i,t}).} \tag{E.21}$$

Evaluating this objective for each token is not the deployed controller. To expose its local benefit and cost, assume that J_{RL} has locally bounded third derivatives along the fixed direction. Taylor expansion for a sufficiently small step gives

$$\begin{aligned} J_{\text{RL}}(\theta_{\text{base}} + \eta w d_{i,t}) - J_{\text{RL}}(\theta_{\text{base}}) &= \eta w \nabla J_{\text{RL}}(\theta_{\text{base}})^\top d_{i,t} \\ &\quad + \tfrac{1}{2}\eta^2 w^2 d_{i,t}^\top \nabla^2 J_{\text{RL}}(\theta_{\text{base}}) d_{i,t} + O((\eta w)^3). \end{aligned}$$

Define the directional benefit and curvature cost

$$B_{i,t} = \nabla J_{\text{RL}}(\theta_{\text{base}})^\top d_{i,t}, \qquad C_{i,t} = -d_{i,t}^\top \nabla^2 J_{\text{RL}}(\theta_{\text{base}}) d_{i,t}. \tag{E.22}$$

The sign of $B_{i,t}$ determines first-order improvement, while positive $C_{i,t}$ denotes negative reward curvature along the direction. Under this additional condition, the quadratic model is $\eta w B_{i,t} - \frac{1}{2}\eta^2 w^2 C_{i,t}$. Its constrained optimum is

$$\boxed{w_{i,t}^{\text{quad}} = \text{clip}\left(\frac{[B_{i,t}]_+}{\eta C_{i,t}}, 0, 1\right).} \tag{E.23}$$

This maximizes the truncated quadratic, not the original RL objective. When $C_{i,t} \le 0$, the diminishing-return argument does not apply. The practical controller therefore replaces both quantities with computable proxies and uses a separately chosen bounded surrogate.

E.6 Benefit and Local-KL Cost Proxies

To avoid computing the true reward gradient and Hessian, we construct the proxies in Equation (6). At sampled token $y_{i,t}$, the unclipped local GRPO (Shao et al., 2024) score gives the estimator

$$\nabla_\theta J_{i,t} \approx G_{i,t}^\top \widehat{A}_i (e_{y_{i,t}} - p_{i,t}).$$

This is the parameter-space pullback of the local logit direction in Equation (D.3). It supplies a comparison direction without replacing the clipped training objective. Substituting $d_{i,t} = G_{i,t}^\top u_{i,t}$ into the exact directional benefit gives

$$\begin{aligned} B_{i,t} &\approx \widehat{A}_i (e_{y_{i,t}} - p_{i,t})^\top G_{i,t} G_{i,t}^\top u_{i,t} \\ &\approx \kappa_{i,t} \widehat{A}_i (e_{y_{i,t}} - p_{i,t})^\top u_{i,t}, \qquad \kappa_{i,t} > 0, \end{aligned}$$

where the second line uses the local isotropic approximation $G_{i,t} G_{i,t}^\top \approx \kappa_{i,t} I$. Absorbing the unknown positive scale and rejecting negative alignment yields

$$\boxed{h_{i,t} = \left[\widehat{A}_i (e_{y_{i,t}} - p_{i,t})^\top u_{i,t}\right]_+.} \tag{E.24}$$

The isotropic approximation discards anisotropy in the logit Jacobian and the score estimator discards interactions with other tokens. Consequently, $h_{i,t}$ measures local alignment rather than verified reward improvement. Its positive part rejects nonpositive alignment under this proxy.

The reward Hessian is generally indefinite, so $C_{i,t} > 0$ need not hold for the exact quadratic. Replace the negative Hessian along this update by a positive-semidefinite Fisher surrogate, using the local information geometry underlying natural-gradient methods (Amari, 1998):

$$-\nabla_\theta^2 J_{\text{RL}}(\theta_{\text{base}}) \approx \beta_{\text{curv}} F_\theta, \qquad F_\theta = G_{i,t}^\top F(p_{i,t}) G_{i,t}, \qquad \beta_{\text{curv}} > 0. \tag{E.25}$$

The positive scale β_{curv} belongs only to this curvature approximation and differs from the EMA decay γ. This replacement is a modeling choice, not an identity for the reward Hessian. Reusing the isotropic approximation gives

$$\begin{aligned} C_{i,t} &\approx \beta_{\text{curv}} d_{i,t}^\top F_\theta d_{i,t} \\ &= \beta_{\text{curv}} (G_{i,t} d_{i,t})^\top F(p_{i,t}) (G_{i,t} d_{i,t}) \\ &\approx \beta_{\text{curv}} \kappa_{i,t}^2 u_{i,t}^\top F(p_{i,t}) u_{i,t}. \end{aligned}$$

This quadratic form is the leading policy movement by Equation (D.11) with $b = u_{i,t}$. Define the Fisher movement term and absorb the unknown positive curvature scale into the active cost coefficient:

$$\boxed{c_{i,t} = u_{i,t}^\top F(p_{i,t}) u_{i,t}, \qquad u_{i,t}^\top F(p_{i,t}) u_{i,t} = \mathrm{Var}_{V \sim p_{i,t}}[u_{i,t}(V)] \geq 0.} \tag{E.26}$$

The variance identity establishes nonnegativity without assumptions on the reward Hessian. For a small logit step $\delta u_{i,t}$, the leading KL movement is $\frac{1}{2}\delta^2 c_{i,t}$. The coefficient α_{cost} controls this proxy's influence in the deployed weight, without estimating the omitted curvature scale.

E.7 Surrogate Acceptance Optimization and Properties

To turn the two proxies into acceptance, we use the surrogate introduced in Section 3.3. Suppress token indices and assume finite $h \geq 0$, $c \geq 0$, $\alpha_{\text{cost}} \geq 0$, and $\epsilon_{\text{cost}} > 0$. With $\Lambda = \alpha_{\text{cost}} c + \epsilon_{\text{cost}}$, define

$$f(w) = h \log w + \Lambda \log(1 - w), \qquad w \in (0, 1). \tag{E.27}$$

For positive benefit, the first term penalizes zero acceptance and the second penalizes full acceptance. This choice retains a benefit-versus-cost tradeoff but is not derived as an exact reward objective. For $h > 0$, $f''(w) = -h/w^2 - \Lambda/(1 - w)^2 < 0$, so f is strictly concave. The first-order condition is

$$\frac{h}{w} - \frac{\Lambda}{1 - w} = 0,$$

whose unique global maximizer is

$$\boxed{w^\star = \frac{h}{h + \Lambda} = \frac{h}{h + \alpha_{\text{cost}} c + \epsilon_{\text{cost}}} \in [0, 1).} \tag{E.28}$$

For $h = 0$, the function reduces to $\Lambda \log(1 - w)$ and has no maximizer in the open interval. Its supremum occurs as w tends to zero, so we extend the solution continuously to $w^\star = 0$. The deployed controller sets $w_{i,t} = \mathrm{sg}[w^\star_{i,t}]$, using $w^*_{i,t}$ for the same stopped weight in the main text. These bounds require the default positive-part benefit. They do not apply to an ablation that inserts an unclamped signed benefit into the same fraction.

For $h > 0$, the acceptance odds provide another interpretation of this solution:

$$\frac{w}{1 - w} = \frac{h}{\alpha_{\text{cost}} c + \epsilon_{\text{cost}}}, \qquad \mathrm{logit}(w) = \log h - \log(\alpha_{\text{cost}} c + \epsilon_{\text{cost}}). \tag{E.29}$$

The log-odds expression requires positive benefit, whereas the rational rule also covers the zero-benefit boundary. To verify monotonicity before stopping the weight, differentiate the rational expression:

$$\frac{\partial w^\star}{\partial h} = \frac{\Lambda}{(h + \Lambda)^2} > 0, \qquad \frac{\partial w^\star}{\partial c} = -\frac{\alpha_{\text{cost}} h}{(h + \Lambda)^2} \leq 0, \qquad \frac{\partial w^\star}{\partial \alpha_{\text{cost}}} = -\frac{hc}{(h + \Lambda)^2} \leq 0.$$

Cost strictly reduces acceptance when $h > 0$ and $\alpha_{\text{cost}} > 0$. At $h = 0$, the controller rejects the direction regardless of cost. The positive floor keeps acceptance below one for finite inputs. PPO (Schulman et al., 2017) clips a policy probability ratio, whereas ZPD computes a stopped scale from evidence alignment and movement cost. Neither these monotonicity properties nor the local proxies guarantee a reward increase. Reference restoration remains ungated in both update paths.

E.8 Multiplicative Advantage Modulation

To verify Path 1 in Section 3.5, we insert acceptance into the outcome advantage instead of adding an evidence gradient. With the path enabled, define

$$\widetilde{w}_{i,t} = m_{i,t} w_{i,t}, \qquad s_{i,t} = 1 + \lambda_{\text{adv}}\, \mathrm{sg}\big(\widetilde{w}_{i,t}\big), \qquad \lambda_{\text{adv}} \geq 0, \tag{E.30}$$

and define the token-level advantage override

$$\boxed{\widehat{A}'_{i,t} = s_{i,t} \widehat{A}_i = \widehat{A}_i \left[1 + \lambda_{\text{adv}}\, \mathrm{sg}\big(\widetilde{w}_{i,t}\big)\right].} \tag{E.31}$$

Here $\widehat{A}_i$ is the original group-relative sequence advantage, repeated over valid response tokens, while $s_{i,t}$ is token dependent. The stop-gradient prevents the current loss from differentiating through the controller to change this scale. The scale can still change when it is recomputed at a later Actor state.

To check its interaction with clipping, write the token-level loss inherited from PPO (Schulman et al., 2017) using the same importance ratio and asymmetric radii as Equation (D.5):

$$\ell^{\text{PPO}}_{i,t}(A) = -\min\big(\rho_{i,t}(\theta) A, \mathrm{clip}(\rho_{i,t}(\theta), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}) A\big). \tag{E.32}$$

For the default nonnegative controller, $s_{i,t} > 0$ is fixed during the update, so positive scaling factors out of both branches:

$$\ell_{i,t}^{\mathrm{PPO}}(\widehat{A}'_{i,t}) = s_{i,t}\ell_{i,t}^{\mathrm{PPO}}(\widehat{A}_i), \qquad \nabla_\theta \ell_{i,t}^{\mathrm{PPO}}(\widehat{A}'_{i,t}) = s_{i,t}\nabla_\theta \ell_{i,t}^{\mathrm{PPO}}(\widehat{A}_i). \tag{E.33}$$

Thus positive scaling preserves the advantage sign and the selected clipping branch, while scaling its gradient wherever the loss is differentiable. At a clipping boundary, the same relation holds for a consistently chosen subgradient. Ineligible tokens and zero acceptance both give $s_{i,t} = 1$. The mask therefore limits modulation without changing the reference loss.

E.9 Complete Objective and Exact Logit Gradients

To assemble the training update, we combine the two channels with the normalizers in Equation (D.2). Assume $N > 0$, binary $m_{i,t} \le a_{i,t}$, and stopped Teacher distributions, constructed directions, masks, and acceptance weights. Here $\mathcal{L}_{\mathrm{evi}}$ denotes the signed evidence loss written as $\mathcal{L}_{\mathrm{OPSD}}$ in the main text. The FKL terms are

$$\begin{aligned}
\mathcal{L}_{\mathrm{ref}}^{\mathrm{FKL}} &= \frac{1}{N}\sum_{i,t} a_{i,t}\mathrm{KL}(q_{i,t}^{\mathrm{ref}} \| p_{i,t}), \\
\mathcal{L}_{\mathrm{evi}}^{\mathrm{FKL}} &= -\frac{1}{Z_{\mathrm{evi}}}\sum_{i,t} m_{i,t}w_{i,t}\sum_{v\in\mathcal{V}} \Delta_{i,t}(v)\log p_{i,t}(v).
\end{aligned} \tag{E.34}$$

The reference channel averages over valid response tokens, while the evidence channel averages over the eligible subset. Because $m_{i,t} \le a_{i,t}$, an additional factor $a_{i,t}$ in the evidence numerator would be redundant. For RKL, assume positive support and use the selected stopped log contrast $r_{i,t}$, including the residual contrast when FEC is enabled:

$$\begin{aligned}
\mathcal{L}_{\mathrm{ref}}^{\mathrm{RKL}} &= \frac{1}{N}\sum_{i,t} a_{i,t}\mathrm{KL}(p_{i,t} \| q_{i,t}^{\mathrm{ref}}), \\
\mathcal{L}_{\mathrm{evi}}^{\mathrm{RKL}} &= \frac{1}{Z_{\mathrm{evi}}}\sum_{i,t} m_{i,t}\left[\log Z_{i,t}(w_{i,t}) - w_{i,t}\mathbb{E}_{v\sim p_{i,t}}[r_{i,t}(v)]\right].
\end{aligned} \tag{E.35}$$

The RKL partition term preserves the loss identity but contributes no Actor gradient. The reference Teacher remains evidence-free even when the selected contrast uses q^- as its base. For binary path switches $\chi_{\mathrm{OPSD}}, \chi_{\mathrm{adv}}$, define $\widetilde{w}_{i,t} = m_{i,t}w_{i,t}$ and $\widehat{A}'_{i,t} = \widehat{A}_i[1 + \chi_{\mathrm{adv}}\lambda_{\mathrm{adv}}\,\mathrm{sg}(\widetilde{w}_{i,t})]$. Combining these terms with GRPO (Shao et al., 2024) gives Equation (11) under FKL and its geometry-matched RKL counterpart. Enabling only χ_{adv} gives Path 1, enabling only χ_{OPSD} gives Path 2, and enabling both gives the joint setting. Write $\mathcal{L}_{\mathrm{VERPO}}$ for the complete loss denoted $\mathcal{L}_{\mathrm{total}}$ in the main text. Differentiating the three terms separately gives the following token-logit descent directions, with the outcome loss evaluated using the stopped modulated advantage. For compactness, write $\lambda_{\mathrm{evi}}^\star = \chi_{\mathrm{OPSD}}\lambda_{\mathrm{evi}}$:

$$\begin{aligned}
-\nabla_{z_{i,t}}\mathcal{L}_{\mathrm{VERPO}}^{\mathrm{FKL}} &= -\nabla_{z_{i,t}}\mathcal{L}_{\mathrm{GRPO}}(\widehat{A}') + \frac{\lambda_{\mathrm{ref}}a_{i,t}}{N}(q_{i,t}^{\mathrm{ref}} - p_{i,t}) + \frac{\lambda_{\mathrm{evi}}^\star}{Z_{\mathrm{evi}}}m_{i,t}w_{i,t}\Delta_{i,t}, \\
-\nabla_{z_{i,t}}\mathcal{L}_{\mathrm{VERPO}}^{\mathrm{RKL}} &= -\nabla_{z_{i,t}}\mathcal{L}_{\mathrm{GRPO}}(\widehat{A}') + \frac{\lambda_{\mathrm{ref}}a_{i,t}}{N}F(p_{i,t})(\log q_{i,t}^{\mathrm{ref}} - \log p_{i,t}) + \frac{\lambda_{\mathrm{evi}}^\star}{Z_{\mathrm{evi}}}m_{i,t}w_{i,t}F(p_{i,t})r_{i,t}.
\end{aligned} \tag{E.36}$$

These exact logit gradients recover the main-text FKL update and its RKL counterpart under the stated stop-gradient and support assumptions. The reference term contains neither acceptance weights nor path switches. ZPD scales the outcome gradient through the advantage in Path 1 and the signed evidence direction directly in Path 2. The two routes therefore share acceptance without sharing a gradient channel.

F Additional Results

To examine the scope of the main results, this appendix reports joint-path performance, trajectory routing, offline token weights, and evidence-loss sensitivity, followed by a support-mass audit of vocabulary truncation. The comparisons extend Table 3 and the controller analysis in Section 4.4. The training comparisons examine how the choice of update paths and corrected trajectories affects task performance. Offline replay complements these comparisons by characterizing evidence acceptance across trajectory classes, response regions, and token types without updating model parameters. The

support-mass audit quantifies the probability mass retained by the truncated vocabulary support, providing context for the approximation used in the implementation.

F.1 Joint Loss Weighting and Advantage Modulation

Joint Activation. To test additive gains, we enabled both switches in Section 3.5 as VERPO-LW+AM. Under the main evaluation protocol, Table 6 shows that joint activation falls below the stronger single-path setting across the evaluated backbones. These endpoint results show no additive average gains but do not explain the decrease.

Table 6: Acc/mean@16 of VERPO-LW+AM across three model backbones.

Model	Biology	Chemistry	Materials	Physics	Tool Use	Average
Qwen3-4B	60.16	70.17	77.26	68.98	62.56	67.83
Qwen3-8B	61.78	75.62	74.62	68.82	67.15	69.60
Llama-3.2-1B	55.38	70.41	56.38	44.29	49.26	55.14

Annealing Protocol and Results. To test gradual switching, we trained Llama-3.2-1B-Instruct (Meta, 2024) for 200 trainer steps per task with EMA decay 0.95, FEC-FKL, and all-trajectory evidence eligibility. For trainer step s, the progress $a(s) = \mathrm{clip}((s - 50)/100, 0, 1)$ sets $\lambda_{\mathrm{evi}} = a(s)$ and $\lambda_{\mathrm{adv}} = 1 - a(s)$. This transfers weight from advantage modulation to evidence regularization over steps 50–150, retaining the stopped ZPD controller, $\lambda_{\mathrm{ref}} = 0.1$, and the base GRPO (Shao et al., 2024) loss throughout. Figure 7 shows the stage scores and full traces. Gradual switching does not consistently preserve earlier gains, and the training dynamics show changes in response length alongside performance.

(a) Annealing results

Task	50	150	200
Biology	37.50	15.12	19.12
Chemistry	47.29	62.41	62.62
Materials	35.17	19.02	27.26
Physics	40.08	38.05	36.48
Tool Use	44.49	45.68	48.25

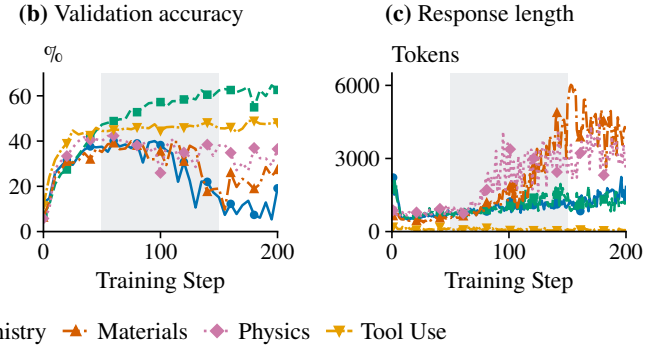


Figure 7: Path annealing on Llama-3.2-1B-Instruct (Meta, 2024). (a) Validation Acc/mean@16 percentages at trainer steps 50, 150, and 200. (b) Validation accuracy every five steps, with higher being better. (c) Mean training-rollout length in tokens at every step, with no preferred direction. Shared colors and line styles identify tasks. Shading marks the linear transition over steps 50–150.

Interpretation. Advantage modulation rescales sampled-action reinforcement, while evidence regularization changes probability allocation. These different update targets can favor different optimization trajectories and output behaviors. One possible explanation is that the two paths induce different effective optimization objectives and could therefore favor different optimal policy distributions. Switching between or jointly activating them could redirect optimization, making earlier gains difficult to preserve or combine.

F.2 Trajectory-Level Correction Scope

To examine the role of the evidence mask, Table 7 compares routing only incorrect trajectories with routing both outcome classes for each of the three update configurations. All six rows use Llama-3.2-1B (Meta, 2024) on the same five tasks. The comparison concerns trajectory eligibility, which precedes token-wise acceptance. An eligible token still receives no evidence weight when its local alignment is nonpositive.

Table 7: Acc/mean@16 by evidence-correction scope on Llama-3.2-1B. Wrong-only makes only incorrect trajectories eligible, whereas All makes both outcome classes eligible. All entries are percentages, with percentage symbols omitted. Average is the unweighted five-task mean. Bold marks the higher score within each update-path pair.

Scope	Biology	Chemistry	Materials	Physics	Tool Use	Average
VERPO-LW						
Wrong-only	56.32	64.40	44.01	48.20	45.49	51.68
All	**60.15**	**66.57**	**52.26**	**54.22**	**49.63**	**56.57**
VERPO-AM						
Wrong-only	**57.65**	63.35	**60.30**	48.67	46.37	**55.27**
All	55.50	**63.69**	55.06	**50.70**	**51.01**	55.19
VERPO-LW+AM						
Wrong-only	53.66	63.51	48.40	**45.54**	45.86	51.39
All	**55.38**	**70.41**	**56.38**	44.29	**49.26**	**55.14**

For VERPO-LW, All reaches 0.5657 compared with 0.5168 for Wrong-only and scores higher on every task. For VERPO-LW+AM, the corresponding averages are 0.5514 and 0.5139. All leads on four tasks, while Wrong-only leads on Physics. VERPO-AM instead gives a slightly higher average with Wrong-only, 0.5527 compared with 0.5519. Wrong-only leads on Biology and Materials, while All leads on Chemistry, Physics, and Tool Use. A possible explanation is that correct rollouts still contain local decisions that benefit from evidence-guided refinement, making their inclusion useful for loss weighting. For advantage modulation alone, the two scopes differ by only 0.08 percentage points on average, with mixed task-level preferences. These results suggest that the value of correcting successful trajectories depends on how evidence enters the update.

F.3 Token-Level FEC and ZPD Weight Analysis

To locate acceptance within a response, we analyze 100 matched Materials pairs, each containing one format-valid correct and incorrect rollout from the same prompt group. We replay all responses offline without parameter updates and report candidate-level acceptance over both outcome classes, covering 50,292 correct-side and 50,241 incorrect-side token events.

Trajectory-level Acceptance. Table 8 summarizes the aggregate weight statistics in Figure 6. Correct rollouts have an effective-weight fraction of 43.3% compared with 36.4% for incorrect rollouts, and their mean weights are 0.2337 and 0.2034. The mean therefore supports broader acceptance in this sample without implying that every correct rollout receives more weight.

Table 8: Token-level FEC and ZPD weight statistics for Llama-3.2-1B. Effective denotes $w > 10^{-3}$ and High-weight denotes $w \geq 0.5$. Rates and means include zero weights. The final column reports the mean trajectory-level sum of controller weights.

Trajectory	Effective %	High-weight %	Mean Weight	Mean Sum/Pair
Correct	43.3	24.4	0.2337	117.55
Incorrect	36.4	21.5	0.2034	102.21

Weight by Response Region. We partition each trajectory at its `<answer>` opening tag. Tokens before that tag form the reasoning region, and the remaining tokens form the final-answer region. Table 9 shows that the mean reasoning-region weight sums are 117.22 and 101.92 for the two classes, compared with 0.33 and 0.29 in their final answers. Region length contributes to these totals. The higher per-token means in reasoning show that the difference also reflects weight allocation, not only the number of tokens.

High-weight Token Types. To examine the lexical content of highly weighted positions, we aggregate weights by token identity. Token types are retained only when they occur at least 20 times, and their mean weights include zero-weight occurrences rather than conditioning on controller

Table 9: FEC and ZPD weight statistics by response region under the same matched replay protocol and thresholds as Table 8. The final column reports the mean within-region sum of controller weights.

Trajectory	Region	Effective %	High-weight %	Mean Weight	Mean Sum/Pair
Correct	Reasoning	43.5	24.7	0.2369	117.22
	Final answer	30.7	4.5	0.0406	0.33
Incorrect	Reasoning	36.3	21.8	0.2062	101.92
	Final answer	40.6	3.0	0.0359	0.29

acceptance. Table 10 reports high-ranking lexical or subword types for each trajectory class and provides the statistics summarized in Figure 6.

Table 10: High-ranking token types in the Materials replay sample. Rankings require at least 20 occurrences and include zero-weight occurrences.

Trajectory	Token	Occurrences	Effective %	High-weight %	Mean Weight
Correct	`analyze`	23	73.9	73.9	0.685
	`Given`	61	68.9	67.2	0.614
	`select`	25	68.0	64.0	0.612
	`therm`	23	65.2	65.2	0.607
	`compound`	20	80.0	65.0	0.606
	`greater`	30	66.7	63.3	0.597
	`one`	30	73.3	60.0	0.596
	`reference`	140	64.3	62.9	0.589
Incorrect	`without`	24	62.5	62.5	0.579
	`compute`	24	62.5	58.3	0.547
	`indicates`	35	60.0	57.1	0.545
	`actually`	20	60.0	60.0	0.540
	`these`	54	64.8	55.6	0.523
	`Therefore`	34	67.6	52.9	0.515
	`find`	49	59.2	57.1	0.515
	`will`	28	53.6	53.6	0.506

The correct-side ranking includes `analyze`, `select`, and `reference`, alongside domain-related types such as `therm`. The incorrect-side ranking includes `compute`, `without`, and `Therefore`. These examples describe lexical associations with acceptance. A token-type mean merges different contexts and does not identify a causal reasoning operation or an error location.

Matched Token Types. The previous ranking can partly reflect differences in which words are used by the two classes. To make the comparison more direct, we next restrict attention to token types occurring at least 20 times in *both* classes and compare their mean weights. Table 11 reports the largest differences in each direction.

Table 11: Largest differences in mean controller weight between correct and incorrect rollouts among token types occurring at least 20 times in both classes. For the left pair of columns, the gap is $\bar{w}_{\text{correct}} - \bar{w}_{\text{incorrect}}$. For the right pair, the gap is $\bar{w}_{\text{incorrect}} - \bar{w}_{\text{correct}}$.

Correct-favored			Incorrect-favored		
Token	Correct	Gap	Token	Incorrect	Gap
`therm`	0.607	+0.370	`without`	0.579	+0.234
`Given`	0.614	+0.360	`Correct`	0.492	+0.226
`greater`	0.597	+0.327	`P`	0.321	+0.217
`sqrt`	0.527	+0.305	`Therefore`	0.515	+0.216
`incorrect`	0.520	+0.290	`indicates`	0.545	+0.180
`element`	0.525	+0.281	`does`	0.466	+0.178
`reference`	0.589	+0.221	`properties`	0.459	+0.177

For example, `therm` has a correct-favored gap of 0.370, while `without` has an incorrect-favored gap of 0.234. Comparing the same token types removes differences in type identity but does not match their surrounding contexts.

F.4 Hyperparameter Sensitivity

To examine the evidence-loss scale in Equation (11), we extend Table 4(b) with the first 200 training steps of five Llama-3.2-1B Chemistry runs. The four settings $\lambda_{\text{evi}} \in \{0.2, 0.4, 0.6, 0.8\}$ used all-trajectory evidence correction. All five used FEC, FKL, EMA decay 0.95, $\lambda_{\text{ref}} = 0.1$, and loss weighting without advantage modulation.

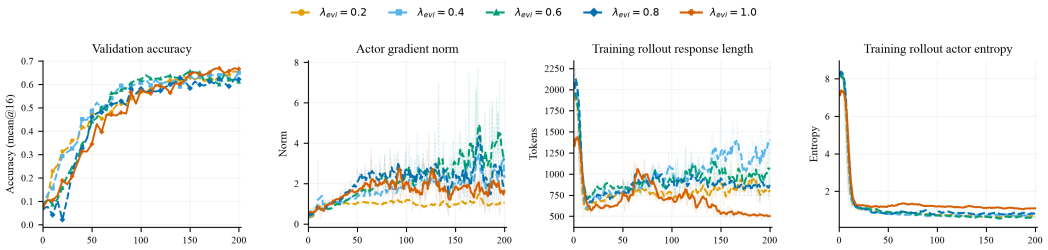


Figure 8: Sensitivity of Llama-3.2-1B to the evidence-loss coefficient λ_{evi} on Chemistry. From left to right, the panels report validation accuracy, Actor gradient norm, rollout response length, and Actor entropy. Higher accuracy indicates better task performance.

Figure 8 and Table 4(b) show no monotonic improvement with λ_{evi}. The 1.0 run reaches the highest step-200 score of 66.58%, but its code revision and planned horizon differ. This comparison cannot attribute that advantage to λ_{evi} alone or establish 1.0 as the best coefficient under a matched protocol.

F.5 Probability Mass Retention under Top-K Truncation

To assess the vocabulary approximation in Appendix H.4, we examine the first 200 steps of five Llama-3.2-1B-Instruct (Meta, 2024) FEC runs, one per task. All use FKL, joint loss weighting and advantage modulation, all-trajectory evidence routing, and EMA decay 0.95.

Support Construction and Measurement. The correction support S combines the top 128 candidates from each of q^+, q^-, and q^0 with the sampled token and removes duplicates. Thus $K = 128$ specifies candidates per branch, not the final support size. The reference channel uses a separate support S_{ref} formed from its own top 128 candidates and the sampled token. All selected probabilities retain the full softmax normalization, without renormalizing the selected coordinates or adding a tail bucket. For a distribution q, retained mass on S is $M = \sum_{v \in S} q(v)$. Figure 9 reports the logged token means of $1 - M$. The positive Teacher, negative Teacher, and Actor share S, whereas the reference mass uses S_{ref}.

Observed Retention. Each logged mass averages over routed valid response tokens before acceptance weighting. Across the five tasks and four branches, token-weighted retained mass ranges from 95.92% to 99.32%, while the step-200 means range from 99.62% to 99.88%. The task-specific token-weighted mean correction-support sizes range from 148.26 to 171.33 coordinates. These measurements support $K = 128$ as a compact support with high aggregate probability retention in the tested configuration. Retention is not uniformly high throughout training. The lowest valid step mean is 28.90% in the early Chemistry reference trace, and the plot retains these early low-coverage observations.

Exact Truncated Gradient. To identify what truncation changes, consider one token with $p = \text{softmax}(z)$ at unit temperature. Freeze S, the signed direction Δ on its selected coordinates, and the nonnegative weight w during differentiation. Define the zero-padded vector $\Delta_S(v) = \mathbf{1}[v \in S]\Delta(v)$ and the truncated loss $L_S = -w \sum_{v \in S} \Delta(v) \log p(v)$. The softmax derivative $\partial \log p(v)/\partial z_j = \mathbf{1}[v = j] - p(j)$ gives

$$s_S = \sum_{v \in S} \Delta(v), \qquad -\nabla_z L_S = w\big(\Delta_S - s_S p\big). \tag{F.1}$$

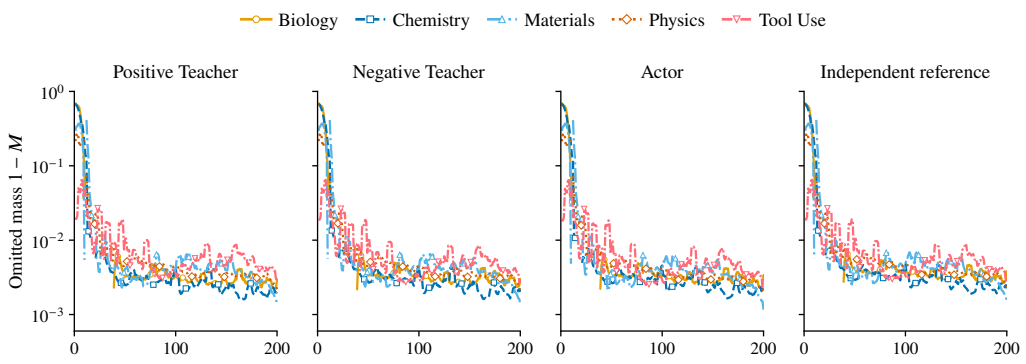


Figure 9: Omitted probability mass over the first 200 training steps for five FEC-LW+AM runs on Llama-3.2-1B-Instruct (Meta, 2024). The underlying observations are step means over routed valid response tokens. Lower values indicate greater probability retention on the selected support. The first three panels use the shared evidence-correction support, and the reference panel uses its independent support. The logarithmic axes show centered five-step moving averages within contiguous valid segments, with segment endpoints preserved. Colors, line styles, and markers distinguish tasks. Gaps mark zero-token steps and are not interpolated.

The extra term $-ws_S p$ comes from the live full-vocabulary normalizer. It acts on selected and omitted logits, ensuring that the logit update sums to zero even when the selected direction does not. Since $\|p\|_1 = 1$ and $w \geq 0$, its absolute size satisfies

$$\left\|-\nabla_z L_S - w\Delta_S\right\|_1 = w|s_S|. \tag{F.2}$$

Thus the simple direction $w\Delta_S$ recovers the negative loss gradient when $s_S = 0$, including a full-vocabulary difference of normalized endpoints. A small-vocabulary finite-difference check verifies both identities for nonzero mass defect, zero mass defect, full support, and zero weight, with maximum coordinate error below 3.2×10^{-12}.

G Diagnostic Metrics

To interpret the controller traces in Figure 5, this appendix defines their token sets, averaging rules, and inference limits. The four diagnostics describe acceptance and the measured FEC residual during training. They are not validation or test metrics. Let $\mathcal{I}_{\text{route}}$ denote the valid response-token positions included by the evidence-route diagnostic mask. The implementation first sums each statistic and its token count across microbatches and data-parallel workers and then divides. Every mean below is therefore a global token-weighted mean rather than a mean of per-batch means. For the displayed FEC runs, the FEC diagnostic set equals $\mathcal{I}_{\text{route}}$. The mean definitions below require a nonempty diagnostic set. They do not specify a logging convention for empty sets.

G.1 ZPD Effective Weight Coverage

To measure the breadth of acceptance, WeightEffectiveCoverage counts diagnostic tokens whose stopped weight exceeds a fixed monitoring threshold:

$$\text{WeightEffectiveCoverage} = \frac{1}{|\mathcal{I}_{\text{route}}|} \sum_{(i,t)\in\mathcal{I}_{\text{route}}} \mathbf{1}\left[w_{i,t} > \tau_{\text{eff}}\right], \qquad \tau_{\text{eff}} = 10^{-3}. \tag{G.1}$$

The denominator counts routed diagnostic tokens, not every token in the batch. Coverage therefore measures acceptance conditional on this set and does not measure how often the routing mask selects tokens. The threshold is used only for monitoring and does not enter the controller. This metric differs from the strict nonzero fraction $\mathbf{1}[w_{i,t} > 0]$, which would count arbitrarily small numerical weights as active. Tokens above the threshold contribute equally regardless of their weight magnitude. Thus identical coverage values can accompany different correction strengths, and higher coverage alone establishes neither a larger update nor improved task performance. Comparisons across training steps therefore require the accompanying benefit and cost diagnostics.

G.2 Mean Signed Benefit and Fisher Cost

To distinguish favorable and conflicting directions before rejection, we retain signed alignment in the benefit diagnostic. Equation (6) applies a positive part to this value for controller use. To expose rejected as well as accepted directions, the logged benefit diagnostic retains the value before this positive-part operation:

$$b_{i,t} = \widehat{A}_i (e_{y_{i,t}} - p_{i,t})^\top u_{i,t}, \qquad h_{i,t} = [b_{i,t}]_+. \tag{G.2}$$

The one-hot score direction compares the evidence move with the sampled token's outcome signal. Only $h_{i,t}$ enters the default acceptance rule. The logged metric BenefitMean, shown in the second panel of Figure 5, is

$$\text{BenefitMean} = \frac{1}{|\mathcal{I}_{\text{route}}|} \sum_{(i,t)\in\mathcal{I}_{\text{route}}} b_{i,t}. \tag{G.3}$$

It is a raw signed mean, not the mean of $h_{i,t}$ and not a conditional mean over positive-benefit tokens. A negative value means that the signed evidence directions oppose the local GRPO (Shao et al., 2024) score direction on average. The default controller assigns zero weight to each token with nonpositive benefit, but a negative mean does not imply that every token is rejected. It also does not measure the reward change after an optimizer step.

The logged metric FisherCostMean, shown in the third panel, is the mean nonnegative Fisher movement term from Equation (6):

$$\text{FisherCostMean} = \frac{1}{|\mathcal{I}_{\text{route}}|} \sum_{(i,t)\in\mathcal{I}_{\text{route}}} c_{i,t}, \qquad c_{i,t} = u_{i,t}^\top F(p_{i,t}) u_{i,t} \geq 0. \tag{G.4}$$

This nonnegative statistic measures local distribution movement, not observed reward loss. It is the unscaled dynamic term in the controller denominator. The full movement penalty is $\alpha_{\text{cost}} c_{i,t} + \epsilon_{\text{cost}}$, so the plotted quantity excludes both the scale and the positive floor.

G.3 FEC Residual Nuisance Fisher Covariance

To check the aggregate residual left by the projection, FECResidualCov averages the signed Fisher covariance between the retained FEC residual in Equation (9) and the measured nuisance direction in Equation (8):

$$\text{FECResidualCov} = \frac{1}{|\mathcal{I}_{\text{FEC}}|} \sum_{(i,t)\in\mathcal{I}_{\text{FEC}}} \left\langle \Delta_{i,t}^{\text{FEC}}, \Delta_{i,t}^{\text{nuis}} \right\rangle_{F(p_{i,t})}. \tag{G.5}$$

Here $\mathcal{I}_{\text{FEC}}$ contains the valid response-token positions on which the FEC branches and projection diagnostics are available. Equation (E.15) gives the remaining ridge-dependent covariance for each token under the coefficient in Equation (8).

A small mean signed covariance does not establish token-wise orthogonality. Positive and negative residuals can cancel in the average, and small vector norms also produce small covariance without a small angle. Figure 5 therefore provides an aggregate numerical check of the projection rather than a bound on every residual. Neither the sign nor the magnitude establishes a causal contribution to task accuracy.

H Experimental Details

To document the comparisons in Section 4, this appendix collects the data splits, runtime, shared training settings, method-specific coefficients, and supplied prompt templates. Hardware-specific timing runs are reported separately in Appendix I.

H.1 Datasets and Data Processing

We use the five-task benchmark release associated with SRPO (Li et al., 2026) as a fixed data source. Biology, Chemistry, Materials Science, and Physics are the Level-3 reasoning subsets of SciKnowEval (Feng et al., 2024). Tool Use is derived from ToolAlpaca (Tang et al., 2023). Table 12 reports the complete train and test splits used by every compared method.

Table 12: Five-task data splits shared by the compared methods. Counts denote examples rather than sampled responses.

Task	Source	Train	Test	Total
Biology	SciKnowEval (Feng et al., 2024) Level 3	450	50	500
Chemistry	SciKnowEval (Feng et al., 2024) Level 3	1,890	210	2,100
Materials Science	SciKnowEval (Feng et al., 2024) Level 3	841	94	935
Physics	SciKnowEval (Feng et al., 2024) Level 3	720	80	800
Tool Use	ToolAlpaca (Tang et al., 2023)	4,046	68	4,114
Total	–	7,947	502	8,449

The fixed splits prevent data membership from changing across methods. Science prompts request one of four answer options, while Tool Use prompts provide the available tool specifications. The literal input scaffolds are given in Appendix H.5. For a collapsed run, the reported maximum is its pre-collapse peak and is marked with †. Specifically, the Llama-3.2-1B SDPO (Hübotter et al., 2026) Physics score of 40.15 is the highest pre-collapse score from the run in Figure 4. The displayed curve peaks at step 100 and reaches zero at steps 160, 180, and 200. The score and its five-task average are marked with † and excluded from comparisons restricted to complete, noncollapsed runs. This average is descriptive and does not represent final performance. Table 4(b) instead reports fixed step-200 scores and late-window summaries. Marked entries in Table 4(c) report pre-collapse peaks.

H.2 Models, Baselines, and Runtime

To separate model and runtime choices from method coefficients, we summarize the Actor initialization, hardware profile, and thinking-mode switches are summarized in Table 13. The two-GPU Llama cost study in Appendix I is a separate hardware setting and is not pooled with the eight-GPU timing comparison. The implementation is built on `verl` (Sheng et al., 2025), uses PyTorch FSDP2 (Feng et al., 2025) for distributed Actor training, and uses vLLM (Kwon et al., 2023) for batched rollout generation. We use AdamW (Loshchilov & Hutter, 2019) for optimization.

H.3 Shared Training Setup

To define the common sampling and optimization budget, Table 13 separates rollout settings from validation decoding. These settings describe matched benchmark runs, not the hardware-specific cost profiles. Table 14 then isolates the coefficients and support approximation used by VERPO.

Table 13: Common benchmark, runtime, and optimization settings used by matched experiments. Training and validation decoding are listed separately to distinguish rollout sampling from evaluation. Sequence-length limits and the optimizer-step budget specify the input and generation bounds and training duration for these runs.

Parameter	Value
Actor initialization	Qwen3-4B (Yang et al., 2025), Qwen3-8B (Yang et al., 2025), Llama-3.2-1B-Instruct (Meta, 2024)
Hardware	8× NVIDIA A800 GPUs per run
Actor / Teacher / validation thinking	False / False / False
Global prompt batch / rollouts per prompt	32 / 8
Mini-batch trajectories / update epochs	32 / 1
Maximum prompt / response length	2,048 / 8,192 tokens
Maximum reprompt / model length	10,240 / 18,944 tokens
Training decoding	temperature 1.0, top-p 1.0, top-k disabled
Validation decoding	16 rollouts, temperature 0.6, top-p 0.95, top-k disabled
Outcome clipping	upper radius $\epsilon_{\text{high}} = 0.28$
Optimizer	AdamW (Loshchilov & Hutter, 2019), weight decay 0.01, gradient clip 1.0
Schedule	constant after 10 warmup steps
Budget	200 optimizer steps

H.4 Method-Specific Hyperparameters

To distinguish algorithm settings from the shared budget, Table 15 lists the GRPO, SDPO, and SRPO configurations, while Table 16 reports the RLSD and RLCSD reproduction settings. The SDPO and SRPO entries use the symmetric Jensen-Shannon divergence (Lin, 1991), distinct from the FKL and RKL decompositions used for VERPO. Table 14 separates VERPO's reference and evidence coefficients from its controller parameters. Given the benefit and cost proxies, the acceptance fraction uses α_{cost} and ϵ_{cost}. The projection ridge instead changes the evidence direction before weighting.

Table 14: VERPO-specific hyperparameters for reference restoration, evidence correction, and token-wise acceptance. The table also specifies the FEC projection ridge, vocabulary computation settings, and EMA Teacher decay. These method-specific settings complement the shared training protocol in Table 13.

Parameter	Value
λ_{ref}	0.1
λ_{evi}	1.0
Direction construction	Fix (Hübotter et al., 2026), CTR (Pan et al., 2026), FEC
KL geometry	FKL, RKL
ZPD cost coefficient α_{cost}	0.0025
ZPD denominator floor ϵ_{cost}	2.5×10^{-5}
FEC ridge	10^{-8}
Teacher candidates per branch	Top-K = 128
Vocabulary chunk size	4,096
EMA Teacher decay	0.95

Table 15: Baseline-specific loss and optimizer settings. Dashes preserve entries not specified for the corresponding baseline. Rows separate outcome and distillation losses, dynamic weighting, and optimizer configuration. The side-by-side layout makes differences in loss coefficients and learning rates explicit under the shared benchmark setup.

Parameters	GRPO (Shao et al., 2024)	SDPO (Hübotter et al., 2026)	SRPO (Li et al., 2026)
Outcome loss			
Asymmetric upper clip ϵ_{high}	0.28	–	0.28
Rollout IS clip ρ	–	–	2.0
KL coefficient	0.0	–	0.0
Distillation loss			
Top-K distillation	–	100	100
Distillation divergence	–	Jensen-Shannon (Lin, 1991)	Jensen-Shannon (Lin, 1991)
EMA update rate	–	0.05	0.05
Rollout IS clip ρ	–	2.0	2.0
Dynamic weighting			
Dynamic-weighting coefficient β	–	–	1
Training			
Optimizer	AdamW (Loshchilov & Hutter, 2019)	AdamW (Loshchilov & Hutter, 2019)	AdamW (Loshchilov & Hutter, 2019)
Learning rate	5×10^{-6}	1×10^{-5}	5×10^{-6}
Warmup steps	10	10	10
Weight decay	0.01	0.01	0.01
Gradient clip norm	1.0	1.0	1.0

Table 16: RLSD and RLCSD reproduction hyperparameters in our unified PGR-Probe baseline implementation. The method references and official repositories are RLSD (Yang et al., 2026a) (code) and RLCSD (Pan et al., 2026) (code). Dashes preserve entries not used by the corresponding baseline.

Parameters	RLSD (Yang et al., 2026a)	RLCSD (Pan et al., 2026)
Outcome loss		
Asymmetric upper clip ϵ_{high}	0.28	0.28
Rollout IS clip ρ	2.0	2.0
KL coefficient	0.0	0.0
Teacher construction		
Teacher mode / EMA update rate	EMA / 0.05	EMA / 0.05
Teacher temperature	1.0	1.0
Maximum negative hints	–	4
Remove thinking from demonstration	True	True
Advantage modulation		
Modulation coefficient λ	0.5	0.5
Evidence-ratio clip	0.2	–
Linear decay steps	50	–
Contrast temperature τ	–	0.02
Contrast scale β	–	1.0
Selection threshold δ	–	0.02
Residual clip	–	$[-2, 2]$
Selected-path weight	–	1.0
Training		
Optimizer	AdamW (Loshchilov & Hutter, 2019)	AdamW (Loshchilov & Hutter, 2019)
Learning rate	5×10^{-6}	5×10^{-6}
Warmup steps	10	10
Weight decay	0.01	0.01
Gradient clip norm	1.0	1.0

To reduce vocabulary computation, the implementation forms a deduplicated union of the top-K candidates from the Teacher branches used by each correction and adds the sampled token. FKL FEC uses the three branches q^+, q^-, and q^0. The RKL correction additionally includes the Actor's top-K candidates. Thus $K = 128$ specifies candidates per branch, not the size of the final union.

Selected log probabilities retain their full-vocabulary softmax normalization. The omitted tail is neither renormalized into the selected coordinates nor represented by an extra bucket. Evidence losses, Fisher moments, and projection statistics on this support are therefore approximations to the full-vocabulary quantities in Appendix E. In particular, a truncated FKL displacement need not sum to zero on the selected coordinates, so the exact full-vocabulary logit identities cannot be assumed unchanged after truncation. Support selection, Teacher values, and projection coefficients are stopped during Actor differentiation, while the Actor softmax normalizer remains live. Appendix F.5 gives the exact truncated logit gradient and measures probability retention for $K = 128$ over 200 training steps.

H.5 Prompt Templates

To make the task formatting explicit, the following listings preserve the supplied prompt text, including literal placeholders, punctuation, and response tags. The science system message constrains the final output to an option letter, while its user message supplies the question and options. Together, these messages specify both the task input and the expected answer format. The listings retain their separate roles to clarify how each question is presented to the model. Reading the two messages together makes the formatting requirements explicit at the point of use. This also distinguishes the fixed instructions from the question-specific content supplied in each example.

Listing H.1: Science system message

```
Given a question and four options, please select the right answer.
    Respond in the following format:
<reasoning>
...
</reasoning>
<answer>
...
</answer>

For the answer, only output the letter corresponding to the correct
    option (A, B, C, or D), and nothing else. Do not restate the answer
     text. For example, if the answer is "A", just output:
<answer>
A
</answer>
```

Listing H.2: Science user message

```
[record-specific question and four answer options]
Please reason step by step.
```

Listing H.3: Tool Use system message

```
You are a tool-use assistant. Solve each request by reasoning about
    the task and calling the provided tools when needed.
Use only the tools provided in the user message.
Follow the required response format exactly.
```

Listing H.4: Tool Use user-message scaffold

```
Your task is to answer the user's question using available tools.
You have access to the following tools:
[record-specific tool names, descriptions, API documentation,
 parameter schemas, and output schemas]

Use the following format:
Thought: you should always think about what to do
Action: the action to take, should be one of the tool names.
Action Input: the input to the action, must be in JSON format. All of
    the action input must be realistic and from the user.

Begin!
Question: [record-specific question]
```

The final listing shows the positive-evidence Teacher scaffold. It augments the task prompt with privileged context before replaying the sampled prefix as defined in Equation (D.6).

Listing H.5: Evidence-conditioned self-Teacher reprompt

```
[original task user prompt]
Correct solution:

[validated privileged context or selected sibling response]


Correctly solve the original question.
```

I Computational Analysis

To put the accuracy comparisons in computational context, this appendix reports logged step costs and the approximate compute-matched comparison discussed with Table 3. The Llama-3.2-1B measurements use two A800 GPUs, while the Qwen3-4B Chemistry measurements use eight. These hardware groups are evaluated separately.

I.1 Per-Step Cost

To distinguish total step cost from the recorded component timers, Tables 17 and 18 retain the generation, Teacher/evidence, and Actor-update columns alongside the logged total. Each entry averages its corresponding logged statistic over steps 10 to 200.

Table 17: Logged Llama-3.2-1B computation costs on two A800 GPUs over steps 10 to 200. Gen denotes generation. Peak alloc. denotes logged peak allocated memory.

Method	Hardware	Gen s/step	Teacher/evidence s/step	Actor update s/step	Total step s/step	Throughput tok/s/GPU	Peak alloc. GB
Fixed (Hübotter et al., 2026)	2×A800	7.21	53.48	12.30	23.30	508.4	10.77
CTR (Pan et al., 2026)	2×A800	10.92	53.99	14.84	29.68	638.9	11.66
FEC	2×A800	15.53	54.20	15.58	35.20	1021.4	12.30

Table 18: Logged Qwen3-4B Chemistry computation costs on eight A800 GPUs over steps 10 to 200. N/A means that the corresponding Teacher/evidence timer was not available, not that the method performed no Teacher computation.

Method	Hardware	Gen s/step	Teacher/evidence s/step	Actor update s/step	Total step s/step	Throughput tok/s/GPU	Peak alloc. GB
GRPO (Shao et al., 2024)	8×A800	22.52	N/A	17.58	48.50	466.8	16.20
SDPO (Hübotter et al., 2026)	8×A800	3.32	N/A	28.84	40.80	278.8	35.07
SRPO (Li et al., 2026)	8×A800	5.98	N/A	27.51	42.20	309.2	23.07
VERPO	8×A800	31.09	176.34	42.14	82.41	281.5	18.81

In Table 17, Fixed (Hübotter et al., 2026) has a logged total of 23.30 seconds per step, compared with 29.68 for CTR (Pan et al., 2026) and 35.20 for FEC. This ranks the observed step costs but does not establish equal training quality. In particular, the Fixed-direction experiments exhibited reward hacking as reported with Table 4(a). In Table 18, VERPO records 82.41 seconds per step compared with 48.50 for GRPO (Shao et al., 2024). Its reported peak allocation is 18.81 GB, compared with 16.20 GB for that baseline and 35.07 GB for SDPO (Hübotter et al., 2026). These are measurements from the listed profiles. Throughput alone does not rank end-to-end efficiency because generated token counts and response lengths differ. The recorded peak-allocation field also does not establish whether worker aggregation uses a maximum or another reduction.

I.2 Compute-Matched Performance

To compare quality at an approximately shared budget, we evaluate VERPO, GRPO (Shao et al., 2024), SDPO (Hübotter et al., 2026), and SRPO (Li et al., 2026) on the same Llama-3.2-1B backbone and 2×A800 hardware in Table 19. For each run, we estimate the training compute based on the average wall-clock time per optimization step after the initial warm-up period. The budget for each task is defined by the compute required for the corresponding 200-step outcome-only baseline, and the nearest available five-step validation checkpoint is selected for the other methods. VERPO has the highest reported accuracy in each of the five task columns. Its average of 0.5026 exceeds the 0.4791 from GRPO (Shao et al., 2024) by 0.0235, at an average of 94 rather than 200 optimizer steps.

Table 19: Approximate compute-matched Llama-3.2-1B results. Each method cell reports matched step / validation accuracy. The budget row gives the task-specific target in A800 GPU-hours.

Method	Biology	Chemistry	Materials	Physics	Tool Use	Average
Budget, GPU-h	10.91	11.53	10.61	9.97	10.08	10.62
VERPO	105 / 51.00	105 / 56.61	85 / 45.70	85 / 49.14	90 / 48.85	94 / 50.26
GRPO (Shao et al., 2024)	200 / 49.88	200 / 52.47	200 / 43.48	200 / 48.67	200 / 45.04	200 / 47.91
SDPO (Hübotter et al., 2026)	150 / 8.62	150 / 0.18	145 / 11.17	125 / 39.84	110 / 31.07	136 / 18.18
SRPO (Li et al., 2026)	145 / 31.87	160 / 56.49	130 / 31.91	125 / 43.13	160 / 44.28	144 / 41.54

J Case Study

To illustrate the weight allocation summarized in Appendix F.3, Figure 10 shows one matched pair of correct and incorrect rollouts for a Materials question. Darker text groups mark greater mean acceptance under the local benefit and cost proxies, not an estimated causal contribution to the answer. Uncolored groups fall below the figure's coloring threshold, which does not establish that they are irrelevant to reasoning. This example illustrates nonuniform allocation within two responses rather than a general ordering of token importance. The contrast is specific to the displayed question and sampled responses, so it does not establish a consistent allocation pattern across tasks. Broader trends are assessed through the aggregate token-level diagnostics in Appendix F.3.

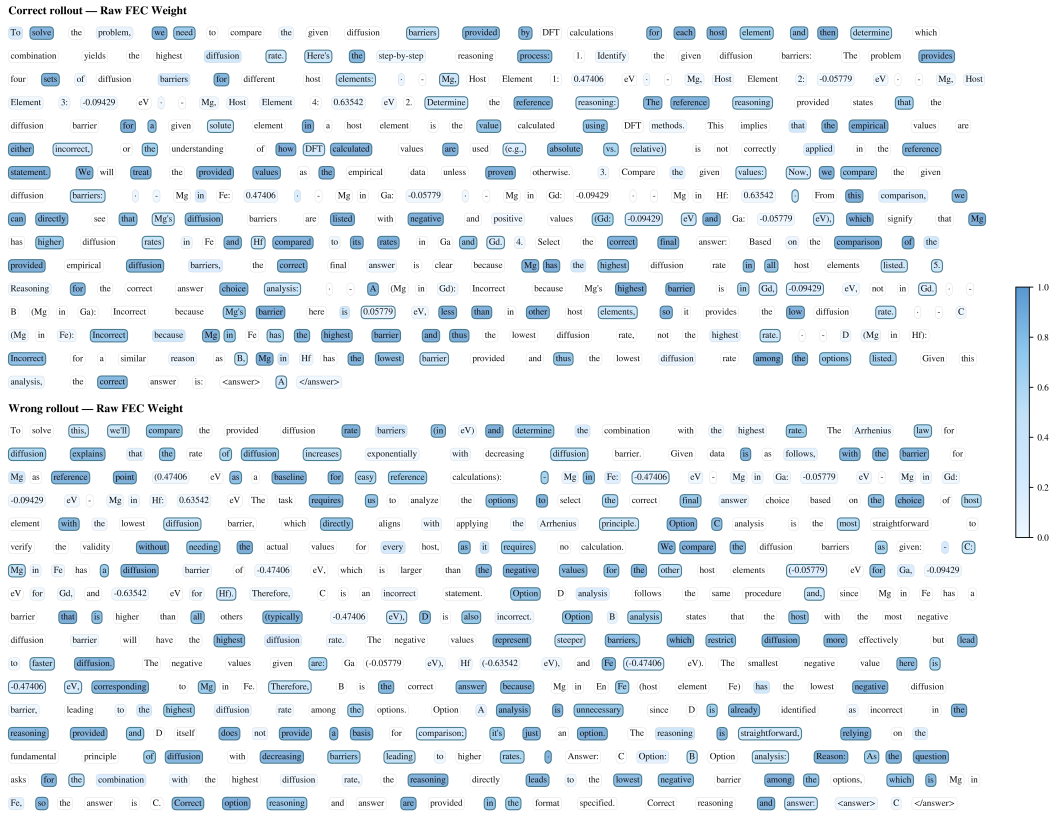


Figure 10: Token-level FEC weights for one matched pair of rollouts. The Materials question asks which host yields the highest diffusion rate for Mg, given diffusion barriers of 0.47406 eV in Fe, −0.05779 eV in Ga, −0.09429 eV in Gd, and 0.63542 eV in Hf. The ground-truth answer is A, corresponding to Mg in Gd. The upper correct rollout predicts A, whereas the lower incorrect rollout predicts C, corresponding to Mg in Fe. Adjacent subword pieces are grouped for readability. Fill color shows their mean offline scope-all raw FEC weight, with darker blue denoting a larger mean. A stronger outline marks a group containing a weight of at least 0.5. Groups whose maximum weight does not exceed 10^{-8} are left uncolored. Weights indicate controller acceptance, not causal contributions to correctness or the training-time routing frequency.