

Back to the Future: Forward-Only Commitor Projection for Trajectory-Level Collapse Attribution in RLVR Reasoning

Jaeun Jang* Hansle Lee Sangmin Kim

AI Research Team, Hanwha Systems, Seongnam-si, Gyeonggi-do, Republic of Korea
{wkdwodms0779, hssarah13, kimsangmin603}@Gmail.com

Abstract

GRPO-trained LLMs sharpen single-sample reasoning but collapse in policy entropy: 7B models lose to their base at Pass@64 on AIME. Existing mitigations score tokens from current-step statistics, missing that collapse is a trajectory-level event accumulating through coherent batch drift. We propose FRACT, a per-rollout score projecting a learned T -step committor’s gradient onto the entropy direction in the dual Fisher metric, with a forward-only closed form requiring no extra backward through the policy. As a sparse, sign-aware selective KL atop GRPO, FRACT improves Acc@16 over KL-Cov by $+0.037/ +0.034/ +0.045$ across Qwen2.5-Base 1.5B/3B/7B and simultaneously raises Pass@1 and Pass@64 on all four benchmarks at 7B—mitigating, not merely shifting, the accuracy–diversity trade-off. Random-matched-mass gains only $+0.009$ and Anti-FRACT underperforms GRPO by -0.007 —the score, not regularization mass, is the cause at $\sim 1.5\%$ structural overhead.

1 Introduction

Reinforcement learning from verifiable rewards (RLVR), exemplified by DeepSeek-R1 (Guo et al., 2025) and trained with GRPO, is now the dominant recipe for sharpening reasoning in large language models. Yet the same training that lifts single-sample accuracy also drives the policy toward entropy collapse. The signature is stark: on AIME24 and AIME25, a GRPO-trained 7B model is *overtaken by the un-trained base model at Pass@64*—consistent with evidence that RL fine-tuning reweights pre-existing reasoning paths rather than expanding the base model’s capability boundary (Yue et al., 2025).

Recent collapse-aware methods target individual tokens for intervention (Cui et al., 2025; Yu et al., 2025; Wang et al., 2025). They all share a

structural limitation: their selection criteria operate on *current-step* statistics, capturing local collapse pressure but unable to tell whether a particular rollout will participate in a future entropy basin several training steps later. Collapse, however, is a trajectory-level event: it accumulates over many steps through coherent drift across the batch.

We propose FRACT, a per-rollout score that targets this gap. FRACT projects the gradient of a learned finite-horizon *committor*—the probability of entering a low-entropy basin within T future training steps—onto the entropy direction in the dual Fisher metric of the simplex (Prop. 1). The projection yields a tractable entropy surrogate plus an explicit residual (Thm. 1), and the surrogate admits a closed-form score computable from forward statistics at the current parameters (Thm. 2). A complementary Fisher-coherence term captures the batch-level drift mechanism. The score drives a selective KL regularizer on top of GRPO, sign-aware: only collapse-pressuring rollouts are penalized.

Across Qwen2.5-Base 1.5B/3B/7B trained on MATH, FRACT improves Avg Acc@16 over the strongest baseline by $+0.037/ +0.034/ +0.045$, and improves the test-time scaling gain $RG_{1 \rightarrow 64}$ over GRPO on 11/12 (scale, benchmark) cells—mitigating, rather than shifting, the accuracy–diversity trade-off. Anti-FRACT (bottom-5% by score) *underperforms* GRPO, isolating the score itself as the cause. Structural overhead is $\sim 1.5\%$ at 7B ($\sim 6.2\%$ total with longer completions).

Contributions. (i) A *projection decomposition* of the committor gradient yielding a trajectory-alignment identity that converts an intractable trajectory gradient into a tractable entropy surrogate plus an explicit residual (Prop. 1, Thm. 1); (ii) A *forward-only, matrix-free closed form* for the resulting head score, computable from current-step statistics with no extra backward through the policy (Thm. 2); (iii) *Empirical validation* across three

*Corresponding author

scales showing simultaneous Pass@1 and Pass@ k improvement, with score-directionality controls.

2 Related Work

Token-level collapse mitigation. Entropy collapse—the rapid per-token sharpening toward argmax under RL—was characterized by Yu et al. (2025) and Cui et al. (2025), the latter establishing an empirical performance–entropy law. **KL-Cov** (Cui et al., 2025) penalizes tokens whose log-probability and advantage covary most strongly; **DAPO** (Yu et al., 2025) decouples PPO clipping to prevent suppression of low-probability, positive-advantage tokens; Wang et al. (2025) restrict updates to the top $\sim 20\%$ high-entropy “forking tokens” driving reasoning diversity; and **CURE** (Li et al., 2025) regenerates rollouts at high-entropy critical tokens, training jointly on original and branched trajectories. All of these operate on instantaneous statistics of the *current* training step. Unlike all of these, FRACT scores trajectory-level collapse risk via a forward-only projection of a learned committor gradient onto the entropy direction, with a complementary Fisher-coherence term capturing batch-level drift.

Global entropy regularization. GRPO (Shao et al., 2024) extends policy-gradient methods (Sutton et al., 1999; Schulman et al., 2017) to LLM reasoning but inherits their exploitation pressure: Kirk et al. (2024) and Padmakumar and He (2024) document that RLHF-style fine-tuning systematically reduces output diversity, and Park et al. (2025) further show the PPO/GRPO clipping mechanism itself induces an entropy bias independent of reward. Global entropy bonuses and reference-model KL penalties mitigate collapse uniformly across rollouts, paying regularization cost on tokens that do not drive collapse. FRACT is complementary: a selective KL signal atop GRPO, focused on the small subset of collapse-pressuring rollouts.

Fisher-metric methods. The dual Fisher metric on the simplex underlying our projection step connects to natural policy gradient methods (Kakade, 2001) and the broader trust-region line (Schulman et al., 2015, 2017). NPG and TRPO use Fisher/KL geometry to precondition or constrain parameter-space updates, while PPO follows the proximal-update principle through a clipped first-order surrogate; FRACT instead uses this geometry for attribution rather than optimization. Specifi-

cally, FRACT performs a simplex-side projection of one cotangent gradient (the committor’s) onto another (entropy’s), yielding the entropy surrogate of Thm. 1 and the forward-only head score of Thm. 2. The construction is matrix-free and practical at LM vocabulary scale—each anchor contributes one $O(|V|)$ Fisher multiplication.

3 Method: FRACT

We propose FRACT (**F**orward-only **p**Rojection-**A**ligned Collapse-risk **a**Ttribution), a forward-only per-rollout score that approximates each rollout’s contribution to a future trajectory-level entropy-collapse risk—in contrast to token-level scores that aggregate per-rollout statistics of the *current* step alone (Cui et al., 2025; Yu et al., 2025). The score is then used as a sparse, signed-aware selector for a complementary KL regularizer on top of GRPO.

The method is built around three results: a projection decomposition of the committor gradient (Proposition 1), a trajectory-alignment identity that turns an intractable trajectory gradient into a tractable entropy surrogate plus an explicit residual (Theorem 1), and a closed-form head score that removes the remaining backward through the policy (Theorem 2). A complementary Fisher-coherent drift augmentation captures the batch-level mechanism by which collapse actually accumulates.

3.1 Setup and Notation

Let $\mathcal{V}$ be the vocabulary. The policy $\pi_\theta(\cdot | c) = \text{softmax}(z_{c,\cdot}(\theta))$ has parameters $\theta = (\theta_{\text{body}}, \theta_{\text{out}})$, with logit $z_{c,v} = w_v^\top h_c + b_v$, weight stack $W := [w_1 | \dots | w_{|V|}] \in \mathbb{R}^{d_h \times |V|}$, and hidden state $h_c \in \mathbb{R}^{d_h}$. The softmax–Fisher identity reads

$$J_\pi(c) = F(\pi_c) J_z(c), \quad F(\pi) := \text{diag}(\pi) - \pi\pi^\top, \quad (1)$$

where F is the categorical Fisher information.

Per-context quantities. Define entropy, centered log-score, weighted score, and entropy gradient:

$$\begin{aligned} H_c &= -\sum_v \pi_c(v) \log \pi_c(v), \\ s_v^{(c)} &:= \log \pi_c(v) + H_c \quad (\text{mean zero under } \pi_c), \\ \phi_v^{(c)} &:= \pi_c(v) s_v^{(c)}, \\ (\nabla_p H)_v &= -\log p(v) - 1. \end{aligned} \quad (2)$$

GRPO step (Shao et al., 2024).

$$\begin{aligned}\Delta\theta^{(t)} &= \frac{\eta}{N} \sum_i \alpha_i^{(t)} g_i^{(t)}, \\ g_i^{(t)} &:= \nabla_{\theta} \log \pi_{\theta^{(t)}}(y_i \mid c_i),\end{aligned}\quad (3)$$

with group-relative advantage $\alpha_i^{(t)}$ and step size η .

Assumption SGD. $\|\Delta\theta\|_2 \leq \eta_0$ and $z_{\theta} \in C^2$ with $\sup \|\nabla_{\theta}^2 z\|_{\text{op}} \leq L_J^{(2)} < \infty$, used in the appendix to bound logit-linearization remainders.

3.2 Anchor Committor and Collapse Risk

Finite-horizon committor. Single-step entropy scores cannot distinguish rollouts driving *future* collapse: a token appears harmless while contributing to a trajectory entering a collapsed entropy basin after T steps. To capture this, we define a trajectory-level target as the first-passage probability of a probe anchor distribution into a collapse region.

Fix an anchor set $\mathcal{A} = \{a_1, \dots, a_M\}$ of M probe contexts. Let $p_a^{(t)} := \pi_{\theta^{(t)}}(\cdot \mid a)$. Define the entropy-collapse set and first-passage time

$$\begin{aligned}\mathcal{C} &:= \{p \in \Delta^{|V|-1} : H(p) \leq H_{\min}\}, \\ \tau_a^{(t)} &:= \inf\{s \geq 0 : p_a^{(t+s)} \in \mathcal{C}\}.\end{aligned}\quad (4)$$

The T -horizon *committor* from parameter θ (default $T=20$, sensitivity in Appendix N) is

$$\chi_T(p_a(\theta)) := \Pr[\tau_a^{(0)} \leq T \mid \theta^{(0)} = \theta], \quad (5)$$

with associated binary trajectory labels

$$Y_a^{(t,T)} := \mathbb{1}[\tau_a^{(t)} \leq T] \quad (6)$$

used to train a small committor estimator $\hat{\chi}_T$ via the replay-update block of Algorithm 1. The anchor-averaged collapse risk and its gradient are

$$\Psi_T(\theta) := \frac{1}{M} \sum_{a \in \mathcal{A}} \chi_T(p_a(\theta)), \quad (7)$$

$$\nabla_{\theta} \Psi_T = \frac{1}{M} \sum_a J_{\pi}(a; \theta)^{\top} \nabla_p \chi_T(p_a). \quad (8)$$

The latter is intractable *per training step*: $\nabla_p \chi_T$ requires expectation over future T -step rollouts and $J_{\pi}(a; \theta)$ a full backward through the policy at every anchor. The estimator $\hat{\chi}_T$ amortizes the future-rollout cost across many steps via asynchronous training; FRACT removes the per-step backward through the policy via Theorem 2 (closed form).

Threshold. We set $H_{\min} = \gamma H_a^{(0)}$ per anchor (default $\gamma = 0.5$, sensitivity in Appendix I.4); the anchor-relative form prevents conflating collapse-prone anchors with intrinsically low-entropy ones.

Committor estimator. The default $\hat{\chi}_T$ is a logistic regression on three anchor-relative features $(\tau_H - H(p), m(p) - \tau_m, g_{12}(p) - \tau_g)$ with $m(p) := \max_v p(v)$ and $g_{12}(p)$ the top-two gap (training details and richer fallbacks in Appendix K.3, K, L).

Supervision and the T -step lag. Trajectory pairs with $s \leq t - T$ supply causal supervision, and a held-out pilot pool prevents temporal leakage (Appendix O.1, O.2). The T -step lag is a property of the supervision signal, not a leakage or staleness defect: $\rho_a^{(t)}$ uses $\nabla_p \hat{\chi}_T$ evaluated at the *current* $\pi_a^{(t)}$, with weights trained on past completions—a horizon predictor, not an instantaneous quantity.

Anchor set. $\mathcal{A}$ is fixed; the default uses $M = 8$ math-format probes selected by highest pre-training entropy decay rate (Appendix M.1, M.2). Proposition 1 and Theorem 1 hold for any $\mathcal{A}$; the strategy enters only through the empirical quality of ρ_a and $\hat{\chi}_T$, quantified by η_{proj} and $\text{CV}(\rho_a)$ (§3.9).

3.3 Clipped Fisher Information

To handle simplex boundary behaviour we clip the policy and form the clipped Fisher information

$$\begin{aligned}\tilde{p}(v) &:= \max(p(v), \epsilon) / \sum_u \max(p(u), \epsilon), \\ F_{\epsilon}(p) &:= \text{diag}(\tilde{p}) - \tilde{p} \tilde{p}^{\top}.\end{aligned}\quad (9)$$

We use F_{ϵ} as the inner product for projecting gradients of scalar functions on the simplex; policy drifts (tangent vectors) instead use F_{ϵ}^{+} .

Primal-dual structure. On the simplex, gradients of scalar functions are cotangent vectors, requiring the dual Fisher metric F_{ϵ} ; policy drifts are tangent vectors, using the primal F_{ϵ}^{+} . The softmax-Fisher identity (1) converts between them, a fact we exploit in §3.7 to express batch-drift coherence via cotangent-side logit directions q_i (full information-geometric justification in Appendix A).

Computational note. F_{ϵ} acts on a vector as

$$F_{\epsilon} v = \tilde{p} \odot v - \tilde{p} (\tilde{p}^{\top} v), \quad (10)$$

which is $O(|V|)$ per multiplication and avoids ever forming F_{ϵ}^{+} (whose entries scale as $1/\tilde{p}(v)$ and would amplify tail noise near collapse).

3.4 Projection Decomposition

We learn the estimator $\hat{\chi}_T$ from trajectory pairs $(p_a^{(t)}, Y_a^{(t,T)})$ (replay-update block, Algorithm 1) and project its gradient onto the entropy direction.

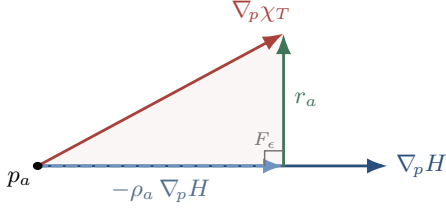


Figure 1: Projection decomposition (Proposition 1) in the cotangent space $T_{p_a}^* \Delta$. The committor gradient $\nabla_p \chi_T$ decomposes into an entropy-aligned component $-\rho_a \nabla_p H$ and a residual r_a , orthogonal in the dual Fisher inner product $\langle u, v \rangle_{F_\epsilon} := u^\top F_\epsilon(p_a) v$.

Proposition 1 (Projection decomposition). *For each anchor a with $\nabla_p H(p_a) \neq 0$, set*

$$\rho_a := -\frac{\nabla_p \chi_T(p_a)^\top F_\epsilon(p_a) \nabla_p H(p_a)}{\nabla_p H(p_a)^\top F_\epsilon(p_a) \nabla_p H(p_a)}. \quad (11)$$

Then $\nabla_p \chi_T(p_a)$ admits the decomposition

$$\nabla_p \chi_T(p_a) = -\rho_a \nabla_p H(p_a) + r_a, \quad (12)$$

with exact dual-Fisher orthogonality

$$r_a^\top F_\epsilon(p_a) \nabla_p H(p_a) = 0. \quad (13)$$

Proof in Appendix B.1.

In practice we substitute $\hat{\chi}_T$ for χ_T and add ε_{num} to the denominator of (11) for stability; gradient calibration of $\hat{\chi}_T$ is monitored separately from its value calibration via η_{proj} and $\text{CV}(\rho_a)$ (§3.9). Figure 1 illustrates the decomposition. The signed ρ_a is retained to preserve Theorem 1’s exact identity; $\rho_{\text{ratio}} := |\{a : \rho_a < 0\}|/M$ flags anchors where the committor varies against the entropy direction.

3.5 Trajectory-Alignment Theorem

Substituting (12) into $\nabla_\theta \Psi_T$ yields the central identity of FRACT. Define the risk-weighted entropy surrogate, treating each ρ_a as a detached scalar,

$$\Phi_\rho(\theta) := \frac{1}{M} \sum_a \rho_a \cdot H_a(\theta). \quad (14)$$

Theorem 1 (Trajectory alignment). *With coefficient ρ_a and residual r_a from Proposition 1,*

$$\nabla_\theta \Psi_T(\theta) = -\nabla_\theta \Phi_\rho(\theta) + \mathcal{E}_{\text{proj}}(\theta), \quad (15)$$

where the parameter-space projection residual is

$$\mathcal{E}_{\text{proj}}(\theta) := \frac{1}{M} \sum_a J_\pi(a; \theta)^\top r_a, \quad (16)$$

satisfying $\|\mathcal{E}_{\text{proj}}\|_2 \leq \frac{1}{M} \sum_a \|J_\pi(a)\|_{\text{op}} \|r_a\|_2$.

Proof in Appendix B.2.

The residual $\mathcal{E}_{\text{proj}}$ vanishes when the committor varies along the entropy direction at every anchor.

The flow score $S_i^{\text{flow}} := -\alpha_i \langle \nabla_\theta \Phi_\rho, g_i \rangle$ then satisfies the first-order expansion $\Psi_T(\theta + \Delta\theta) - \Psi_T(\theta) = \frac{\eta}{N} \sum_i S_i^{\text{flow}} + \langle \mathcal{E}_{\text{proj}}, \Delta\theta \rangle + O(\|\Delta\theta\|^2)$.

3.6 Forward-Only Head Score

The flow score S_i^{flow} requires a backward through π_θ . The body-head factorization $\theta = (\theta_{\text{body}}, \theta_{\text{out}})$ yields $S_i^{\text{flow}} = S_i^{\text{head}} + S_i^{\text{body}}$ with the head admitting a forward-only closed form; the body’s analogous form (Appendix C.1) is unused, head-only sufficiency assessed empirically via r_i^{body} (§3.9).

$$S_i^{\text{head}} := -\alpha_i \langle \nabla_{\theta_{\text{out}}} \Phi_\rho, g_i^{\text{out}} \rangle, \quad (17)$$

with the context-anchor head coupling

$$\kappa(a, c_i) := h_a^\top h_{c_i} + \mathbf{1}_{\text{bias}}, \quad (18)$$

where $\mathbf{1}_{\text{bias}} \in \{0, 1\}$ is 1 if head bias is trainable, 0 otherwise. Wherever $\mathbf{1}_{\text{bias}}$ appears in proofs as a coefficient of a bias-gradient term, it acts as this scalar (broadcast over the relevant logit index).

Theorem 2 (Head closed form). *The head score is*

$$S_i^{\text{head}} = \frac{\alpha_i}{M} \sum_{a \in \mathcal{A}} \rho_a \cdot \kappa(a, c_i) \cdot (\phi_a(y_i) - \mathbb{E}_{v \sim \pi_{c_i}}[\phi_a(v)]). \quad (19)$$

Proof in Appendix B.3.

The body contribution S_i^{body} has a closed form involving the Gram matrix $G_{\text{body}}(a, c)$ (Theorem 3 in Appendix C.1). We do not compute S_i^{body} at every step; instead, the relative residual $r_i^{\text{body}} := |S_i^{\text{body}}|/(|S_i^{\text{head}}| + \varepsilon_{\text{num}})$, sampled on a subset of rollouts, tracks head-only sufficiency (§4).

3.7 Fisher-Coherent Drift Augmentation

Single-token updates rarely cause collapse; *coherent drift across batch updates* does, with a small subset of high-impact tokens accumulating most of the entropy-reducing pressure (Wang et al., 2025)—for example, when many high-advantage tokens at distinct contexts simultaneously push π_a toward a low-entropy basin. We capture this batch-level mechanism with a second score derived from cumulative push-forward drifts at the anchor.

Let $q_i := e_{y_i} - \pi_{c_i}$, the logit-direction gradient of $\log \pi_{c_i}(y_i)$ (cotangent). The anchor-side push-forward policy drift is the tangent vector

$$\delta_i^a := F_\epsilon(\pi_a) q_i \in T_{\pi_a} \Delta. \quad (20)$$

Using (1) and $F F_\epsilon^\top F = F$,

$$\langle \delta_i^a, \delta_j^a \rangle_{F_\epsilon^\top} = q_i^\top F_\epsilon(\pi_a) q_j, \quad (21)$$

so $q^\top F_\epsilon q$ is the tangent Fisher coherence of policy drifts at the anchor, expressed via cotangent-side logit directions. Define the batch drift

$$u_a := \sum_j \alpha_j \kappa(a, c_j) q_j \quad (22)$$

and the risk-weighted Fisher coherence

$$S_i^{\text{FC}} := \frac{\alpha_i}{M} \sum_a \rho_a \kappa(a, c_i) \cdot q_i^\top F_\epsilon(\pi_a)(u_a - \alpha_i \kappa(a, c_i) q_i). \quad (23)$$

A positive S_i^{FC} indicates push-forward drift coherent with the batch in the tangent Fisher metric, weighted by ρ_a , signaling collapse pressure.

Two readings of S_i^{FC} . Mechanistically, S_i^{FC} is the risk-weighted tangent-Fisher coherence of token i 's push-forward drift with the batch, requiring no assumption beyond Proposition 1—our primary reading. Under a Fisher-isotropic curvature assumption, S_i^{FC} further approximates the per-rollout contribution to second-order trajectory curvature of χ_T , a *conditional* reading monitored via η_{FIC} (Appendix D; empirical diagnostics in Appendix J).

The total per-rollout FRACT attribution score is

$$S_i^* := S_i^{\text{head}} + \beta \cdot S_i^{\text{FC}}, \quad \text{with } \beta \geq 0. \quad (24)$$

3.8 Algorithm, Loss, and Complexity

S_i^* drives a sparse, signed-aware selective-KL regularizer. The mask selects the top $k\%$ of the collapse-pressuring rollouts $\mathcal{P}^{(t)} := \{i : S_i^{*,(t)} > 0\}$ (rollouts with negative scores reduce collapse pressure and are excluded), yielding the total loss:

$$m_i := \mathbb{1}[i \in \text{TopK}_{k\%}(\{S_j^{*,(t)}\}_{j \in \mathcal{P}^{(t)}})], \quad (25)$$

$$\mathcal{L} = \mathcal{L}_{\text{GRPO}} + \lambda \cdot \frac{\sum_i m_i D_{\text{KL},i}}{\sum_i m_i + \varepsilon_{\text{num}}}, \quad (26)$$

with $D_{\text{KL},i} := D_{\text{KL}}(\pi_\theta(\cdot|c_i) \parallel \pi_{\theta_{\text{ref}}}(\cdot|c_i))$.

Mass-normalized KL. Dividing by $\sum_i m_i$ rather than N keeps the effective per-rollout KL penalty invariant to the TopK budget k , leaving hyperparameters λ and k independent (Appendix I).

Procedure. Per GRPO step, FRACT (i) anchor pre-computes $\pi_a, F_\epsilon(\pi_a), \nabla_p \hat{\chi}_T, \nabla_p H, \rho_a, \eta_{\text{proj}}$; (ii) computes $S_i^{\text{head}}, S_i^{\text{FC}}, S_i^*$ for each rollout in the batch; (iii) forms $\mathcal{P} \cap \text{TopK}_{k\%}$ as the KL mask; (iv) updates θ with $\mathcal{L}_{\text{GRPO}} + \lambda \cdot$ masked KL; and (v) once $t \geq T$, appends $(p_a^{(t-T)}, Y_a^{(t-T,T)})$ to the $\hat{\chi}_T$ replay buffer and takes one AdamW step. Full pseudocode is in Algorithm 1 (Appendix H).

Component	Risk monitored	Diagnostic
Proposition 1	gradient projection	η_{proj}
$\hat{\chi}_T$ estimation	gradient calibration	$\text{CV}(\rho_a)$
Entropy direction	shape consistency	ρ_{ratio}^-
Theorem 2	head sufficiency	r_i^{body}
Overall	trajectory alignment	AUC / AUPRC

Table 1: FRACT components, their monitored risks, and matched diagnostics; the first four are per-step monitors, the last is an end-of-run evaluation. All reported in §4.

Complexity. Per-step quantities in Algorithm 1 are forward statistics at $\theta^{(t)}$ —no extra backward through π_θ . Head and FC sums dominate at $O(MN|V|)$ each; anchor pre-compute ((10)) is $O(M|V|)$, logistic $\hat{\chi}_T$ negligible. Empirical structural overhead: $\sim 1.5\%$ of GRPO wall-clock at 7B (§4.6; FLOP/wall-clock breakdown, Appendix V).

3.9 Diagnostics

FRACT pairs each component with a lightweight diagnostic: four per-step monitors and one end-of-run validation (Table 1). The *projection residual*

$$\eta_{\text{proj}}(p_a) := \frac{\sqrt{r_a^\top F_\epsilon(p_a) r_a}}{\sqrt{\nabla_p \hat{\chi}_T^\top F_\epsilon \nabla_p \hat{\chi}_T + \varepsilon_{\text{num}}}} \quad (27)$$

measures alignment of $\nabla_p \hat{\chi}_T$ with the entropy direction. Three further monitors track components: *weight stability* $\text{CV}(\rho_a) := \sqrt{\text{Var}_t[\rho_a]} / (|\mathbb{E}_t[\rho_a]| + \varepsilon_{\text{num}})$ for gradient calibration of $\hat{\chi}_T$; *negative-weight ratio* $\rho_{\text{ratio}}^- := |\{a : \rho_a < 0\}| / M$ flagging when entropy fails as a collapse direction (motivating the multi-feature committor, Appendix E); and *head sufficiency* r_i^{body} validating head-only approximation against full closed form (Appendix C.1). End-of-run AUC/AUPRC against entropy-collapse labels $Y^{(t,T)}$ assess trajectory alignment (§4).

4 Experiments

We evaluate FRACT along four questions. **(Q1)** Does the closed-form score match the projection-aligned analytical form? **(Q2)** Does it predict and intervene on future trajectory-level collapse? **(Q3)** Which components—projection weighting, multi-anchor aggregation, and Fisher coherence—contribute? **(Q4)** Does it improve reasoning accuracy and preserve test-time scaling as a selective KL signal? We do not claim FRACT is a perfect collapse oracle, only that it provides a forward-computable, sparse collapse-pressure signal improving the accuracy–diversity trade-off under RL

post-training—unlike token-level scores (Cui et al., 2025; Yu et al., 2025; Wang et al., 2025) based on current-step statistics, missing the trajectory-level coupling FRACT captures via projection-aligned weighting against a finite-horizon committor.

4.1 Setup

Qwen2.5-Base 1.5B/3B/7B with GRPO on MATH training set (Hendrycks et al., 2021) (7,500 problems) via Open-R1 (Hugging Face, 2025) on 4×H200 NVL GPUs; defaults $M = 8$ MATH-format anchors, $T = 20$, $k = 5\%$, $\lambda = 5 \times 10^{-2}$, $\beta = 0.5$, $\gamma = 0.5$ (full configuration in Appendix G). At 1.5B, Acc@16 is stable for $T = 10$ –40 (.435–.438), with $T = 20$ giving the best AUC (.74; Appendix N). Increasing anchors from $M = 8$ to 16 changes Acc@16 only from .438 to .440 (within one SD) while overhead rises from $1.062 \times$ to $1.091 \times$, so we retain $M = 8$ (Appendix M.2). A 100-step pilot GRPO run warm-starts $\hat{\chi}_T$ and is held out from main training (Appendix O.2); the pilot is one-time setup, not per-step overhead.

Baselines. We compare against GRPO, four prior baselines—Entropy Bonus (Mnih et al., 2016), R1-zero-Div (Yao et al., 2025), Pos-Adv-Reweight (Jin et al., 2026), KL-Cov (Cui et al., 2025)—and two FRACT-specific controls: *Random-matched-mass* (5% uniformly at random, matching FRACT’s regularization mass) and *Anti-FRACT* (bottom 5% by FRACT score, testing directionality). Predictive-validity and intervention experiments (§4.2) additionally compare against Random (chance floor), token-entropy mean, gradient norm, and KL-Cov.

Reporting. We report the 5-math-benchmark unweighted mean of Acc@16: component ablations (§4.3) use the *FRACT-head multi* base (Avg $\in$ 0.40–0.44 at 1.5B); E2E tables (§4.4) use full *GRPO* + *variant* pipelines (Avg $\in$ 0.26–0.34 at 1.5B). $\pm$ in E2E tables is pooled std across cells and seeds; tighter seed-only variability and paired per-seed results are in Appendix W (seed allocation in Table 13). Primary claims use 5 seeds, trend-level ablations 3.

Acc@k vs. Pass@k. **Acc@16** averages correctness over 16 generations—a Monte-Carlo estimate of Pass@1, *not* Pass@16. **Pass@k** (Chen et al., 2021) is the probability at least one of k samples is correct. We use Acc@16 (§4.3–§4.4) and Pass@k (§4.5). To compare scaling with different Pass@1,

we report the normalized residual test-time gain

$$\text{RG}_{1 \rightarrow 64} := \frac{\text{Pass@64} - \text{Pass@1}}{1 - \text{Pass@1}}, \quad (28)$$

measuring how much of the residual failure mass is recovered by additional samples (Appendix F).

4.2 Theory Consistency and Prediction

4.2.1 Implementation and

Trajectory-Gradient Consistency

We verify three links in the forward-only construction (Figure 2). First, the coded closed form matches a PyTorch autograd evaluation of the analytical expression at 0.5B over 270 measurement points (Spearman 0.997, $R^2 \geq 0.995$, median relative error $\sim 10^{-5}$). Second, a tractable small- K Monte-Carlo trajectory-gradient estimate converges toward the closed form: from $K = 4$ to 32, relative error falls from 0.38 to 0.16 and rank correlation rises from 0.55 to 0.78. Third, the projected residual is numerically dual-Fisher orthogonal to the entropy gradient at the 10^{-4} level. The finite- K trend is evidence rather than a high- K proof; large-horizon Monte-Carlo verification remains intractable (full numerics in Appendix Q, R).

As a temporal-leakage check, the causal committor trained only on matured pairs $s \leq t - T$ achieves AUC $0.74 \pm .04$, compared with $0.80 \pm .04$ for an intentionally non-causal oracle using future information. All reported FRACT results use the causal estimator (Appendix O.2). Theorem 1’s validity beyond its assumptions is monitored via the diagnostics of §4.3.4.

4.2.2 Predictive Validity

We assess whether FRACT scores at step t predict future entropy collapse in the window $(t, t + T]$. For per-step aggregate signal we define

$$P_\star(t) := \sum_i \max(S_i^{\star,(t)}, 0), \quad (29)$$

the total positive-pressure mass across rollouts at step t . The binary entropy-collapse label is

$$Y^{(t,T)} := \mathbb{1}[\exists \tau \in (t, t + T]: H(\pi_\tau^{(a)}) \leq H_{\min}], \quad (30)$$

computed per anchor and averaged. AUC/AUPRC are computed for $P_\star(t) \rightarrow Y^{(t,T)}$ across training steps. We also assess under a non-entropy *Pass@I-drop* label to defend against trivial self-correlation with entropy. Table 2 reports both labels.

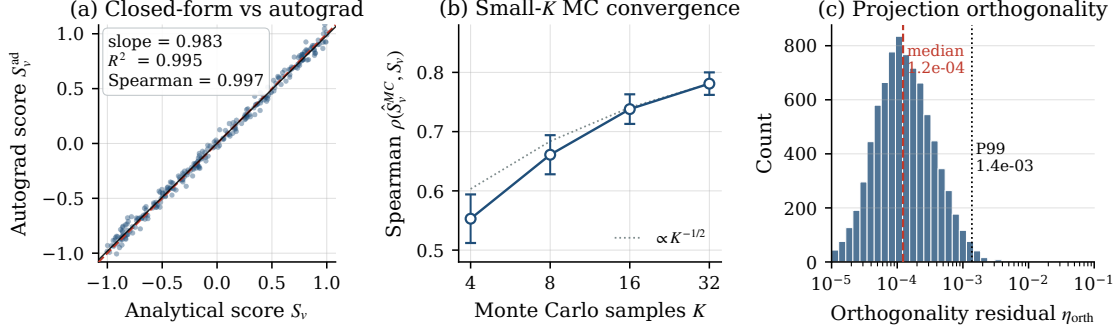


Figure 2: Implementation and trajectory-gradient consistency at 0.5B. (a) Closed-form analytical score vs. autograd evaluation (270 points). (b) Small- K Monte-Carlo rank agreement increases with K toward the closed form. (c) Projection residual is numerically orthogonal to the entropy direction in the dual Fisher metric.

Score	Entropy label			Pass@1-drop	
	AUC	AUPRC	Sp. $-\bar{H}_T$	AUC	AUPRC
Random	.50	.22	.00	—	—
Token ent. mean	.59	.30	.25	.54	.24
Gradient norm	.55	.26	.16	—	—
KL-Cov	.63	.36	.33	.58	.28
FRACT-head (χ_T)	.68	.41	.45	.61	.31
FRACT-head (ρ^{proj})	.71	.44	.49	.63	.33
FRACT-total	.74	.48	.53	.65	.36

Table 2: Predictive validity under the entropy-collapse label (30) and the non-entropy Pass@1-drop label; 1.5B, 5 seeds, all $\pm \leq .06$ omitted. The projection step (FRACT-head $\chi_T \rightarrow \rho^{\text{proj}}$) contributes +0.03 AUC under the entropy label; FRACT-total margin over chance compresses from +0.24 to +0.15 under Pass@1-drop.

Individual seeds show local inversions among adjacent baselines; the robust pattern is that adding projection weighting and Fisher coherence each improves the aggregate signal. Score-ranking visualization across both labels: Figure 4 (Appendix S).

4.2.3 Score-Guided Intervention

We test whether selectively penalizing the top- k FRACT rollouts at step 50 reduces future anchor entropy decay. The intervention is the §3.8 loss restricted to a single step. Table 3 reports $\Delta\bar{H}_T$ across budget levels. A complementary single-token deletion diagnostic at 1.5B is in Appendix P.

4.3 Component Ablations and Diagnostics

4.3.1 Projection Weighting

We isolate ρ_a via $\rho=1$ (no trajectory risk), $\rho=\chi_T$ (committor value), signed ρ^{proj} (theory-faithful), and clipped $\rho^{\text{proj}+} := \max(\rho^{\text{proj}}, 0)$. The signed projection step ($\rho=\chi_T \rightarrow \rho^{\text{proj}}$) yields the largest Acc@16 gain (+0.023, Table 4), exceeding value-

Score	$k=1\%$	3%	5%	10%
Random	+.003	+.009	+.014	+.006
Token entropy mean	+.012	+.024	+.030	+.025
KL-Cov	+.010	+.030	+.038	+.032
FRACT-head (ρ^{proj})	+.021	+.041	+.050	+.044
FRACT-full	+.024	+.045	+.061	+.053

Table 3: Score-guided intervention effect $\Delta\bar{H}_T$ at step 50, 1.5B, 5 seeds. Cell std ≤ 0.036 (omitted). FRACT’s advantage peaks at $k=5\%$; scores plateau or decrease at $k=10\%$. A budget-vs- $\Delta\bar{H}_T$ plot is in Figure 3.

Metric	$\rho=1$	$\rho=\chi_T$	ρ^{proj} signed	$\rho^{\text{proj}+}$
Acc@16	.398 $\pm$.008	.411 $\pm$.009	.434$\pm$.010	.428 $\pm$.010
RG _{1$\rightarrow$64}	.43 $\pm$.04	.47 $\pm$.04	.51$\pm$.04	.49 $\pm$.04
ρ_{ratio}	—	0	.08-.20	0

Table 4: Projection weighting ablation, 1.5B, 5 seeds.

weighting ($\rho=1 \rightarrow \rho=\chi_T$, +0.013), consistent with projection being the core mechanism.

4.3.2 Fisher Coherence Weight β

We sweep $\beta \in \{0, 0.25, 0.5, 1.0, 2.0\}$. $\beta=0.5$ is jointly optimal across all metrics (Table 5); $\beta \geq 1.0$ over-weights coherence, degrading head sharpness.

4.3.3 Anchor Strategy

We compare High-risk (default), High-entropy, First-completion, Prompt-level, and Random strategies. High-risk leads both Acc@16 (+0.011 over high-entropy) and AUC (+0.03, Table 6), confirming trajectory-level selection beyond local entropy. Random tracks the $\rho=1$ row of Table 4.

4.3.4 Diagnostic Dynamics

We monitor the four diagnostics of §3.9 throughout training. *Projection residual* η_{proj} stays below 0.3 in median during normal training and rises pre-

β	Acc@16	AUC	RG _{1→64}
0	.418±.010	.71±.04	.49±.04
0.25	.430±.011	.73±.04	.51±.04
0.5	.438±.011	.74±.04	.53±.04
1.0	.435±.012	.72±.05	.49±.05
2.0	.420±.015	.68±.05	.45±.06

Table 5: Fisher coherence ablation, 1.5B, 3 seeds.

Anchor	Acc@16	AUC	Overhead
High-risk	.439±.011	.74±.04	1.058×
High-entropy	.428±.012	.71±.04	1.052×
First-completion	.418±.013	.67±.04	1.050×
Prompt-level	.414±.012	.65±.04	1.044×
Random	.402±.015	.61±.05	1.041×

Table 6: Anchor selection strategy, 1.5B, 3 seeds.

collapse (Table 9), signaling that the entropy direction loses dominance precisely when collapse is imminent. *Head/body approximation*: median r_i^{body} remains below 0.35 across steps 20/60/100, while Spearman agreement between S^{head} and S^{flow} rises to .83–.90 and top-5% Jaccard to .62–.74 by step 100 (Table 7), supporting head-only selection. *Fisher-coherence mechanism*: G_{FC} roll-outs have substantially higher pairwise drift cosine than random (.12 vs. .02 median), and combining head and FC gives the largest entropy-preservation intervention ($\Delta \bar{H}_T = +.059$; Table 8). The FC increment is directional rather than independently decisive at the small seed budget, so we retain it as an aggregate-positive complementary correction. A quadrant view of $(\eta_{\text{proj}}, \rho_{\text{ratio}}^-)$ co-evolution is in Appendix T.

4.4 End-to-End Benchmark-wise Training

Table 10 reports each method’s Acc@16 across five reasoning benchmarks: GSM8K (Cobbe et al., 2021), MATH500 (Lightman et al., 2024), AMC and AIME24 (Li et al., 2024), and AIME25 (Zhang and Math-AI Team, 2025).

FRACT-full leads Avg by +0.037, +0.034, +0.045 over KL-Cov at 1.5B/3B/7B. FRACT-specific controls confirm directionality: Random-matched-mass tracks Entropy Bonus (+0.009 over GRPO at 1.5B; regularization mass without selection criterion), while Anti-FRACT underperforms GRPO by −0.007. All 15 FRACT-full – KL-Cov cells (scales × benchmarks) are positive (Appendix W.1, Figure 5), as are all 15 matched seed averages; paired seed gaps span +.031–+.044 / +.031–+.037 / +.035–+.058 at 1.5B/3B/7B (Ap-

Step	Med. r_i^{body}	Sp. $S^{\text{head}}, S^{\text{flow}}$	Top-5% Jacc.
20	.25–.34	.76–.84	.52–.64
60	.20–.30	.81–.88	.59–.70
100	.17–.27	.83–.90	.62–.74

Table 7: Head/body approximation diagnostics, 1.5B, 5 seeds. P90 body residual is .38–.66.

Quantity	Random	G_{FC}	G_{head}	G_{total}
Drift cos. (med.)	.02	.12	.06	.10
$\Delta \bar{H}_T$	+.012	+.024	+.046	+.059

Table 8: Fisher-coherence mechanism, 1.5B, 5 seeds. FC selection concentrates coherent batch drift; the combined intervention gives the largest entropy preservation.

pendix W), with the largest benchmark gains at AMC/AIME24 on 7B.

Fisher coherence contribution. FRACT-full – FRACT-head multi isolates the FC term. Paired per-seed FC contribution is positive in 4/5 seeds at each scale; Avg $\Delta \in \{+0.006, +0.005, +0.006\}$ (per-benchmark breakdown in Appendix W.5). With 5 seeds per scale, results are directional, not decisive; FC complements rather than drives FRACT.

4.5 Pass@k Scaling

Table 11 reports Pass@k at 7B (5 seeds) for all three methods under the same pipeline. Pass@k at 3B (5 seeds) and 1.5B (3 seeds) in Appendix X.

Test-time scaling. GRPO improves Pass@1 over Base but loses residual scaling on harder benchmarks (Base’s Pass@64 surpasses GRPO’s on AIME24/25). FRACT-full improves Pass@1 and RG_{1→64} over GRPO on all four benchmarks, exceeding Base and GRPO at every $k \in \{1, 16, 64\}$. This simultaneous Pass@1 and Pass@k improvement is the signature of selective regularization on collapse-critical tokens. Across scale, Base overtakes GRPO at Pass@64 on AIME24/AIME25, whereas FRACT-full exceeds Base at every measured $k \in \{1, 16, 64\}$ in all 12 scale–benchmark cells and improves RG_{1→64} over GRPO in 11/12 (Appendix X.6, X.7); curves are in Appendix X.1.

4.6 Computational Cost

FRACT-full adds 6.2% wall-clock at 7B over GRPO vs. 1.8% for KL-Cov, decomposing into $\sim 1.5\%$ structural cost (anchor forward, F_e multiplications, head/FC sums, ρ_a evaluation, mask

Stage	Median η_{proj}	P90
Early	.18–.26	.34–.45
Mid	.22–.31	.39–.52
Pre-collapse	.31–.43	.52–.66
Severe collapse	.34–.50	.58–.74

Table 9: Projection residual η_{proj} over training, 1.5B, 5 seeds. The pre-collapse rise from .18–.26 to .31–.43 in median is the signature monitored by Proposition 1.

Method	GSM8K	MATH	AMC	AIME24	AIME25	Avg
<i>1.5B, 5 seeds</i>						
GRPO	.702	.520	.120	.004	.000	.269±.013
+Entropy Bonus	.716	.532	.128	.012	.000	.278±.012
+R1-zero-Div	.724	.536	.132	.018	.002	.282±.014
+Pos-Adv-Reweight	.722	.534	.130	.013	.004	.281±.013
+KL-Cov	.738	.560	.138	.054	.033	.305±.014
+FRACT-head single	.768	.581	.169	.067	.046	.326±.016
+FRACT-head multi	.772	.598	.184	.083	.043	.336±.017
+FRACT-full	.781	.603	.187	.085	.052	.342±.016
Rnd.-matched-mass	.720	.532	.124	.013	.002	.278±.018
Anti-FRACT	.694	.507	.108	.000	.000	.262±.014
<i>3B, 5 seeds</i>						
GRPO	.780	.595	.245	.018	.004	.328±.013
+KL-Cov	.815	.626	.274	.086	.054	.371±.015
+FRACT-head multi	.854	.668	.310	.099	.069	.400±.015
+FRACT-full	.852	.672	.319	.106	.074	.405±.016
Anti-FRACT	.768	.580	.224	.011	.000	.317±.015
<i>7B, 5 seeds</i>						
GRPO	.825	.715	.550	.125	.083	.460±.014
+KL-Cov	.841	.750	.584	.141	.096	.482±.016
+FRACT-head multi	.879	.778	.638	.187	.123	.521±.017
+FRACT-full	.876	.785	.642	.201	.130	.527±.018
Anti-FRACT	.808	.692	.527	.105	.063	.439±.015

Table 10: Benchmark-wise Acc@16. Reported $\pm$ is pooled std across cells and seeds. Full 3B/7B comparison (Entropy Bonus, R1-zero-Div, Pos-Adv-Reweight, Random-matched-mass) in Appendix W.1.

gather), $\sim 3.0\%$ from longer completions—the policy retaining the option to extend reasoning rather than prematurely collapse—and $\sim 1.7\%$ residual system overhead (Table 37, Appendix V).

Generalization beyond RLVR. Under INTUITOR-style RLIF (Zhao et al., 2026), FRACT-full also improves Avg Acc@16 over INTUITOR + KL-Cov by $+.022/+ .018/+ .022$ at 1.5B/3B/7B, respectively, while the 7B normalized residual gain increases from .56 to .63. The gains are smaller than under RLVR but remain consistently positive across scales (Appendix U).

4.7 Summary

FRACT improves Acc@16 over KL-Cov by $+.037/+ .034/+ .045$ at 1.5B/3B/7B with positive paired gains in all 15 matched seeds. It exceeds Base at every measured $k \in \{1, 16, 64\}$ across all 12 scale–benchmark cells and improves residual scaling over GRPO in 11/12 cells (Appendix X.6). Controls confirm directionality (Anti-

Method	k	MATH500	AMC	AIME24	AIME25
Base	1	.60	.23	.07	.05
	16	.86	.65	.40	.32
	64	.92	.80	.52	.43
	RG _{1→64}	.80	.74	.48	.40
GRPO	1	.71	.55	.13	.07
	16	.89	.82	.41	.28
	64	.93	.90	.50	.36
	RG _{1→64}	.76	.78	.43	.31
FRACT-full	1	.78	.64	.21	.14
	16	.92	.90	.48	.36
	64	.95	.96	.57	.45
	RG _{1→64}	.77	.89	.46	.36

Table 11: Pass@ k and RG_{1→64} at 7B, 5 seeds.

FRACT -0.007 vs. GRPO) and that selection, not regularization mass, drives the effect (Random-matched-mass $+0.009$); the FC term adds a consistent $+0.005$ – $+0.006$ in 4/5 seeds at every scale as a complementary correction. Structural overhead is $\sim 1.5\%$ of GRPO wall-clock (6.2% with longer completions), and gains extend to RLIF with intrinsic rewards (Appendix U).

5 Conclusion

We recast entropy collapse in RLVR as a trajectory-level attribution problem. Projecting the gradient of a learned T -step committor onto the entropy direction in the dual Fisher metric of the simplex (Prop. 1) turns an intractable trajectory gradient into a tractable risk-weighted entropy surrogate plus a monitorable residual (Thm. 1), and the head term admits a matrix-free closed form requiring no extra backward through the policy (Thm. 2). The resulting per-rollout score, used as a sign-aware sparse selective-KL atop GRPO, improves Avg Acc@16 over KL-Cov by $+.037/+ .034/+ .045$ on Qwen2.5-Base 1.5B/3B/7B and raises Pass@1 and Pass@64 on all four 7B benchmarks—mitigating, not merely shifting, the accuracy–diversity trade-off—while Anti-FRACT and Random-matched-mass jointly isolate the score itself as the causal lever at $\sim 1.5\%$ structural overhead.

The geometric takeaway: once committor and entropy gradients are placed in their natural dual-Fisher pairing, the right unit of intervention is the rollout weighted by its projected collapse contribution, not the token taken locally. We expect this projection-then-attribution pattern to apply to other RL post-training pathologies that accumulate coherently across batches rather than locally.

Limitations

Scope of application. FRACT is a selective KL regularizer applied on top of GRPO; it is not a replacement for RL algorithms such as PPO or RLHF and is not designed to subsume existing post-training pipelines. Our experiments are restricted to mathematical reasoning (MATH, GSM8K, AMC, AIME), the Qwen2.5-Base family at 1.5B/3B/7B (decoder-only transformers), and one alternative reward regime (INTUITOR-style RLIF, Appendix U). Transferability to other reasoning domains—logical reasoning, code generation, multi-step planning—and to other architecture families (encoder-decoder, mixture-of-experts) or to scales beyond 7B remains to be established.

Theoretical scope. Proposition 1 and Theorem 1 are stated under explicit assumptions on the simplex-tangent geometry, and Theorem 2 additionally relies on the body–head factorization standard in transformer architectures. A Monte-Carlo verification of the trajectory-gradient identity at large training horizons is computationally intractable and is not attempted; the assumptions are instead monitored empirically through the matched diagnostics of §3.9 (η_{proj} , $\text{CV}(\rho_a)$, ρ_{ratio}^- , r_i^{body}). The Fisher-isotropic curvature interpretation of S_i^{FC} is offered as a complementary conditional reading; the primary tangent-coherence interpretation, which all reported conclusions rest on, requires no additional assumption beyond Proposition 1.

Component design. The Fisher coherence term is by design a complementary correction layered on top of the head score, not an independently decisive component; the full head-only variant is recoverable by setting $\beta = 0$ at no architectural cost (effect-size breakdown in §4.7 and Appendix Y). The selective-KL framing inherits the standard tradeoff of regularization-based methods: gains are conditional on the regularization budget being properly tuned within the operating regime.

Hyperparameter dependence. FRACT introduces additional hyperparameters (T , k , λ , β , γ , and the committor estimator $\hat{\chi}_T$). Defaults are chosen by pilot tuning and validated through sensitivity sweeps covering each (Appendices N, M.2, I.4, I); robustness within practical ranges is documented, but other reward regimes or architectures may require re-tuning.

Computational overhead. FRACT adds $\sim 1.5\%$ structural overhead and $\sim 6.2\%$ total overhead (including longer completions) over GRPO, dominated by the asynchronous committor-estimator training. The estimator is amortized across many training steps via a one-time pilot warm-start (§4.1, Appendix O.2); per-step compute carries no additional backward pass through the policy due to the head closed form (Theorem 2).

Ethical and societal considerations. FRACT modifies the RL training procedure of large language models without introducing new training data, altering safety alignment procedures, or changing downstream task scope. Risks associated with deploying any RL-tuned language model—factuality, bias amplification, misuse—are inherited from the base model and the reward signal and are not specifically addressed by our method.

Failure modes. A catalog of falsification criteria stated in advance of the experiments and acceptable mixed outcomes observed in the results is provided in Appendix Y.

Ethics Statement

FRACT is a methodological contribution to RLVR post-training; it modifies the RL optimization trajectory but introduces no new training data and no changes to safety alignment procedures. All datasets (MATH, GSM8K, MATH500, AMC, AIME24/25) and base models (Qwen2.5-Base 1.5B/3B/7B) are publicly available and used within their intended research scope. Risks associated with deploying any RL-tuned language model—factual errors, bias amplification, capability misuse—are inherited from the base model and reward signal and are not specifically addressed by our method. We make no claim that FRACT improves safety, factuality, or alignment. Computational overhead is documented (Appendix V), and per-seed numerical results with paired comparisons are reported (Appendix W) to support reproduction.

References

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Hugging Face. 2025. [Open R1: A Fully Open Reproduction of DeepSeek-R1](#).
- Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. 2026. [Revisiting entropy in reinforcement learning for large reasoning models](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 25300–25322, San Diego, California, United States. Association for Computational Linguistics.
- Sham M Kakade. 2001. A natural policy gradient. *Advances in neural information processing systems*, 14.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *ICLR*.
- Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q. Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. [Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions](#). Hugging Face repository.
- Qingbin Li, Rongkun Xue, Jie Wang, Ming Zhou, Zhi Li, Xiaofeng Ji, Yongqi Wang, Miao Liu, Zheming Yang, Minghui Qiu, and Jing Yang. 2025. CURE: Critical-token-guided re-concatenation for entropy-collapse prevention. *arXiv preprint arXiv:2508.11016*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *ICML*.
- Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *ICLR*.
- Jaesung R. Park, Junsu Kim, Gyeongman Kim, Jinyoung Jo, Sean Choi, Jaewoong Cho, and Ernest K. Ryu. 2025. Clip-low increases entropy and clip-high decreases entropy in reinforcement learning of large language models. *arXiv:2509.26114*.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv:2402.03300*.
- Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025. [Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning](#). In *Advances in Neural Information Processing Systems 38*, pages 127985–128019.
- Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. 2025. [Diversity-aware policy optimization for large language model reasoning](#). In *Advances in Neural Information Processing Systems 38*, pages 105100–105125.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [DAPO: An open-source LLM reinforcement learning](#)

system at scale. In *Advances in Neural Information Processing Systems* 38, pages 125532–125554.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems* 38, pages 64304–64339.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. 2025. No free lunch: Rethinking internal feedback for LLM reasoning. *arXiv:2506.17219*.

Yifan Zhang and Math-AI Team. 2025. American invitational mathematics examination (AIME) 2025. Hugging Face dataset.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2026. Learning to reason without external rewards. In *ICLR*. INTUITOR; *arXiv:2505.19590*.

A Information-Geometric Background

This appendix substantiates the choice of F_ϵ as the inner product in the projection of Proposition 1. The choice is dictated by an information-geometric primal–dual pairing on the probability simplex. We first review the relevant geometry, then verify that two natural projection formulations—one in the cotangent space, one in the tangent space—coincide and reduce to the formula (11) appearing in the main body.

A.1 Tangent and Cotangent Spaces on the Simplex

Let $\Delta = \{p \in \mathbb{R}_{>0}^{|V|} : \mathbf{1}^\top p = 1\}$ denote the open simplex of strictly positive distributions. The tangent space at p is

$$T_p\Delta = \{u \in \mathbb{R}^{|V|} : \mathbf{1}^\top u = 0\}. \quad (31)$$

The Fisher geometry on Δ distinguishes two kinds of objects with different natural norms.

Tangent (primal) norm. Policy drifts $\delta p \in T_p\Delta$ are tangent vectors. Their natural norm is the Fisher information metric:

$$\begin{aligned} \|\delta p\|_{\text{tan}}^2 &:= \delta p^\top F_\epsilon^+(p) \delta p \\ &= \sum_v \frac{(\delta p_v)^2}{\tilde{p}(v)}, \end{aligned} \quad (32)$$

a chi-squared-type distance that locally weighs drifts by the inverse of the clipped probability. This is the metric obtained by pulling back the Euclidean metric on the parameter space of a softmax model through the policy map.

Cotangent (dual) norm. Gradients of scalar functions, $\nabla_p f \in T_p^*\Delta$, are cotangent vectors. Their natural norm is the dual Fisher metric:

$$\|\nabla_p f\|_{\text{cot}}^2 := \nabla_p f^\top F_\epsilon(p) \nabla_p f. \quad (33)$$

For the natural-gradient flow $\tilde{\nabla} f := F_\epsilon \nabla_p f \in T_p\Delta$ (steepest-ascent direction in the Fisher geometry),

$$\begin{aligned} \|\tilde{\nabla} f\|_{\text{tan}}^2 &= (F_\epsilon \nabla f)^\top F_\epsilon^+(F_\epsilon \nabla f) \\ &= \nabla f^\top F_\epsilon F_\epsilon^+ F_\epsilon \nabla f \\ &= \nabla f^\top F_\epsilon \nabla f \\ &= \|\nabla_p f\|_{\text{cot}}^2, \end{aligned}$$

using the identity $F_\epsilon F_\epsilon^+ F_\epsilon = F_\epsilon$ on $T_p\Delta$.

A.2 Why F_ϵ for the Projection

Both $\nabla_p \chi_T(p_a)$ and $\nabla_p H(p_a)$ are gradients of scalar functions on the simplex and hence live in $T_{p_a}^*\Delta$. The natural inner product for projecting one onto the other is therefore the dual Fisher metric $F_\epsilon(p_a)$, not the primal $F_\epsilon^+(p_a)$. We verify below that this choice is consistent with two alternative formulations of the same projection.

Equivalence with natural-gradient projection in the tangent space. Define the natural gradients $\tilde{\nabla} \chi_T := F_\epsilon \nabla_p \chi_T$ and $\tilde{\nabla} H := F_\epsilon \nabla_p H$, both tangent vectors. Their projection coefficient in the tangent (F_ϵ^+) metric is

$$\begin{aligned} \rho_a^{\text{tan}} &:= \frac{\langle \tilde{\nabla} \chi_T, \tilde{\nabla} H \rangle_{F_\epsilon^+}}{\langle \tilde{\nabla} H, \tilde{\nabla} H \rangle_{F_\epsilon^+}} \\ &= \frac{(F_\epsilon \nabla \chi_T)^\top F_\epsilon^+(F_\epsilon \nabla H)}{(F_\epsilon \nabla H)^\top F_\epsilon^+(F_\epsilon \nabla H)} \\ &= \frac{\nabla \chi_T^\top F_\epsilon F_\epsilon^+ F_\epsilon \nabla H}{\nabla H^\top F_\epsilon F_\epsilon^+ F_\epsilon \nabla H} \\ &= \frac{\nabla \chi_T^\top F_\epsilon \nabla H}{\nabla H^\top F_\epsilon \nabla H}, \end{aligned} \quad (34)$$

where the last step uses $F_\epsilon F_\epsilon^+ F_\epsilon = F_\epsilon$. Comparing (34) with (11) gives $\rho_a = -\rho_a^{\text{tan}}$: the cotangent projection and the natural-gradient tangent projection produce the same coefficient up to the conventional sign. The dual Fisher metric is the geometrically consistent inner product for the projection step.

Equivalence with the natural-gradient direction.

A second consistency check: the natural-gradient direction $F_\epsilon \nabla \chi_T$ is the steepest-ascent direction of χ_T in the Fisher geometry. Its component along the

natural-gradient direction of H involves F_ϵ in the numerator. The projection in (11) therefore measures *how aligned the steepest-ascent direction of the committor is with the steepest-ascent direction of entropy*, in the Fisher geometry of the simplex.

A.3 Numerical Advantages of F_ϵ

Beyond the geometric justification, F_ϵ has substantial practical advantages over F_ϵ^+ .

Bounded entries. $F_\epsilon = \text{diag}(\tilde{p}) - \tilde{p}\tilde{p}^\top$ has entries bounded by $\max_v \tilde{p}(v) \leq 1$. F_ϵ^+ has spectral entries proportional to $1/\tilde{p}(v)$, growing unboundedly as the clipping $\epsilon \rightarrow 0$ on a near-zero-probability token. Using F_ϵ^+ as an inner product would amplify noise on rare tokens—precisely those whose log-probability has the largest fluctuation under stochastic policy updates.

Matrix-free implementation. F_ϵ acts on a vector as

$$F_\epsilon v = \tilde{p} \odot v - \tilde{p}(\tilde{p}^\top v), \quad (35)$$

which is $O(|V|)$ per multiplication, requires only the clipped policy vector $\tilde{p}$, and avoids materializing either F_ϵ or F_ϵ^+ as a dense $|V| \times |V|$ matrix. Forming F_ϵ^+ explicitly would require a pseudo-inverse computation cubic in $|V|$ for a naive approach and numerically unstable near the boundary.

Practical implication. The combination of bounded entries and matrix-free multiplication is what makes the projection step in FRACT computationally cheap at LM scale: each anchor contributes one $F_\epsilon \cdot v$ multiplication plus inner products, all $O(|V|)$. The total anchor pre-compute is $O(M|V|)$, dominated in practice by the anchor forward pass rather than by the projection computation itself.

A.4 Connection to Natural Policy Gradient

The choice of F_ϵ has a natural reading in the natural-policy-gradient literature (Kakade, 2001): in NPG one preconditions the parameter-space gradient by the inverse Fisher information of the policy. In FRACT, the projection step (11) is the simplex-side analogue: it asks what fraction of the (cotangent) committor gradient aligns with the (cotangent) entropy gradient in the dual Fisher metric. The resulting coefficient ρ_a then weights an entropy-surrogate gradient in parameter space (Theorem 1), producing what is effectively a natural-gradient-informed surrogate for trajectory risk.

A.5 Worked Example: $|V| = 3$

To make the geometry concrete, consider $|V| = 3$ with $\pi = (0.7, 0.2, 0.1)$ and no clipping ($\tilde{p} = \pi$). Then

$$F_\epsilon = \begin{pmatrix} 0.21 & -0.14 & -0.07 \\ -0.14 & 0.16 & -0.02 \\ -0.07 & -0.02 & 0.09 \end{pmatrix},$$

$$\nabla_p H = \begin{pmatrix} -\log 0.7 - 1 \\ -\log 0.2 - 1 \\ -\log 0.1 - 1 \end{pmatrix} \approx \begin{pmatrix} -0.643 \\ 0.609 \\ 1.303 \end{pmatrix}.$$

For a hypothetical committor gradient $\nabla_p \chi_T = (-0.1, 0.05, 0.2)^\top$, the projection coefficient is

$$\rho_a = -\frac{\nabla \chi_T^\top F_\epsilon \nabla H}{\nabla H^\top F_\epsilon \nabla H} \approx -\frac{0.027}{0.119} \approx -0.227,$$

indicating that the committor varies along the negative entropy direction (i.e., committor decreases as entropy increases) and contributing positively to the entropy surrogate Φ_ρ . The residual $r_a = \nabla \chi_T + \rho_a \nabla H \approx (-0.246, 0.188, 0.504)^\top$ satisfies $r_a^\top F_\epsilon \nabla H \approx 0$ to numerical precision, confirming dual-Fisher orthogonality.

B Proofs of Main Results

This section gives self-contained proofs of the three results stated in the main text: Proposition 1 (projection decomposition), Theorem 1 (trajectory alignment), and Theorem 2 (head closed form). Notation follows §3.

B.1 Proof of Proposition 1

Statement (restated). For each anchor a with $\nabla_p H(p_a) \neq 0$, define

$$\rho_a := -\frac{\nabla_p \chi_T(p_a)^\top F_\epsilon(p_a) \cdot \nabla_p H(p_a)}{\nabla_p H(p_a)^\top F_\epsilon(p_a) \cdot \nabla_p H(p_a)}. \quad (11 \text{ restated})$$

Then $\nabla_p \chi_T(p_a) = -\rho_a \nabla_p H(p_a) + r_a$ with $r_a^\top F_\epsilon(p_a) \nabla_p H(p_a) = 0$.

Proof. Define the residual

$$r_a := \nabla_p \chi_T(p_a) + \rho_a \nabla_p H(p_a),$$

so the decomposition $\nabla_p \chi_T = -\rho_a \nabla_p H + r_a$ holds by construction. The orthogonality claim

is a direct computation. Writing $\langle u, v \rangle_a := u^\top F_\epsilon(p_a) v$,

$$\begin{aligned}\langle r_a, \nabla_p H \rangle_a &= \langle \nabla_p \chi_T + \rho_a \nabla_p H, \nabla_p H \rangle_a \\ &= \langle \nabla_p \chi_T, \nabla_p H \rangle_a \\ &\quad + \rho_a \langle \nabla_p H, \nabla_p H \rangle_a.\end{aligned}$$

Substituting the definition of ρ_a ,

$$\begin{aligned}\langle r_a, \nabla_p H \rangle_a &= \langle \nabla_p \chi_T, \nabla_p H \rangle_a \\ &\quad - \frac{\langle \nabla_p \chi_T, \nabla_p H \rangle_a}{\langle \nabla_p H, \nabla_p H \rangle_a} \langle \nabla_p H, \nabla_p H \rangle_a \\ &= 0.\end{aligned}$$

This holds *exactly*, with no ϵ -residual: the only assumption is that the dual-Fisher norm of $\nabla_p H$ at p_a is non-zero, which fails only at the uniform distribution where $\nabla_p H \propto \mathbf{1}$ and $F_\epsilon \mathbf{1} = \mathbf{0}$ (measure-zero, excluded throughout). $\square$

Why the dual metric. Appendix A provides the full information-geometric justification: $\nabla_p \chi_T$ and $\nabla_p H$ are cotangent vectors, and the natural inner product on cotangent vectors is the dual Fisher metric F_ϵ . The same projection coefficient is obtained equivalently as a tangent-space projection of the natural gradients $F_\epsilon \nabla_p \chi_T$ and $F_\epsilon \nabla_p H$, confirming geometric consistency.

Implementation note. The proof above uses the exact $\nabla_p \chi_T$. In practice FRACT uses the learned $\nabla_p \hat{\chi}_T$ and adds a numerical regularizer ε_{num} to the denominator of (11):

$$\rho_a^{\text{impl}} := - \frac{\nabla_p \hat{\chi}_T(p_a)^\top F_\epsilon(p_a) \nabla_p H(p_a)}{\nabla_p H(p_a)^\top F_\epsilon(p_a) \nabla_p H(p_a) + \varepsilon_{\text{num}}}. \quad (36)$$

This yields ρ_a^{impl} that deviates from the exact ρ_a by $O(\varepsilon_{\text{num}})$ when the true denominator is bounded away from zero, and gracefully damps the projection when $\nabla_p H$ approaches the flat direction. We monitor η_{proj} and $\text{CV}(\rho_a)$ to track the resulting calibration (§3.9 and §4.3.4).

Sign of ρ_a . Proposition 1 does not constrain the sign of ρ_a . When the committor varies along $-\nabla_p H$ in the dual metric, $\rho_a > 0$ (the typical ESS regime). When the committor varies along $+\nabla_p H$ at some anchor, $\rho_a < 0$ at that anchor. The signed convention is essential for Theorem 1 to hold as an exact identity; clipping the sign would break the chain-rule decomposition. The fraction of anchors with $\rho_a < 0$ is reported as ρ_{ratio}^- (§3.9).

B.2 Proof of Theorem 1

Statement (restated). With ρ_a, r_a from Proposition 1 and $\Phi_\rho(\theta) := \frac{1}{M} \sum_a \rho_a H_a(\theta)$ (each ρ_a detached),

$$\nabla_\theta \Psi_T(\theta) = -\nabla_\theta \Phi_\rho(\theta) + \mathcal{E}_{\text{proj}}(\theta), \quad (15 \text{ restated})$$

where $\mathcal{E}_{\text{proj}}(\theta) := \frac{1}{M} \sum_a J_\pi(a; \theta)^\top r_a$.

Proof. By the chain rule applied to $\chi_T(p_a(\theta))$ at a single anchor,

$$\nabla_\theta \chi_T(p_a(\theta)) = J_\pi(a; \theta)^\top \nabla_p \chi_T(p_a), \quad (37)$$

where $J_\pi(a; \theta) = \partial \pi_\theta(\cdot | a) / \partial \theta \in \mathbb{R}^{|V| \times d}$ is the policy Jacobian at the anchor. The gradient $\nabla_p \chi_T(p_a)$ lives in $T_{p_a}^* \Delta$ (well-defined because $J_\pi(a) \mathbf{1} = \mathbf{0}$). Substituting the projection decomposition (12),

$$\begin{aligned}\nabla_\theta \chi_T(p_a(\theta)) &= J_\pi(a)^\top [-\rho_a \nabla_p H(p_a) + r_a] \\ &= -\rho_a J_\pi(a)^\top \nabla_p H(p_a) \\ &\quad + J_\pi(a)^\top r_a \\ &= -\rho_a \nabla_\theta H_a(\theta) + J_\pi(a)^\top r_a,\end{aligned}$$

where the final equality applies the chain rule to $H_a(\theta) = H(p_a(\theta))$ in the opposite direction. Averaging over $a \in \mathcal{A}$,

$$\begin{aligned}\nabla_\theta \Psi_T(\theta) &= \frac{1}{M} \sum_a \nabla_\theta \chi_T(p_a(\theta)) \\ &= -\frac{1}{M} \sum_a \rho_a \nabla_\theta H_a(\theta) \\ &\quad + \frac{1}{M} \sum_a J_\pi(a)^\top r_a.\end{aligned}$$

Because each ρ_a is treated as a scalar detached from the autograd graph, $\nabla_\theta(\rho_a H_a) = \rho_a \nabla_\theta H_a$, so the first sum equals $\nabla_\theta \Phi_\rho(\theta)$. The second sum is $\mathcal{E}_{\text{proj}}(\theta)$, completing the proof. $\square$

Operator-norm bound on the residual. By sub-multiplicativity and the triangle inequality,

$$\|\mathcal{E}_{\text{proj}}\|_2 \leq \frac{1}{M} \sum_a \|J_\pi(a)\|_{\text{op}} \|r_a\|_2. \quad (38)$$

Two observations: (i) when the committor gradient lies in the entropy direction at every anchor (the ESS regime), $r_a = 0$ for all a and the residual vanishes identically; (ii) via softmax-Fisher,

$$\begin{aligned}\|J_\pi(a)\|_{\text{op}} &= \|F(\pi_a) J_z(a)\|_{\text{op}} \\ &\leq \|F(\pi_a)\|_{\text{op}} \|J_z(a)\|_{\text{op}} \\ &\leq \frac{1}{2} \|J_z(a)\|_{\text{op}}\end{aligned}$$

by Lemma 2. A uniform bound on the logit Jacobian (Assumption SGD) therefore yields a uniform bound on $\|\mathcal{E}_{\text{proj}}\|_2$.

Detachment of ρ_a . The detachment of ρ_a is essential: if ρ_a depended on θ through autograd, the identity (15) would acquire additional terms involving $\partial \rho_a / \partial \theta$. The detached convention makes Φ_ρ a clean entropy surrogate—only the θ -dependence inside H_a contributes to $\nabla_\theta \Phi_\rho$.

Per-rollout flow score and first-order expansion. The flow score $S_i^{\text{flow}} := -\alpha_i \langle \nabla_\theta \Phi_\rho, g_i \rangle$ inherits the residual via Theorem 1. A first-order Taylor expansion of $\Psi_T(\theta + \Delta\theta)$ along $\Delta\theta = \frac{\eta}{N} \sum_i \alpha_i g_i$ yields

$$\begin{aligned} \Psi_T(\theta + \Delta\theta) - \Psi_T(\theta) &= \frac{\eta}{N} \sum_i S_i^{\text{flow}} + \langle \mathcal{E}_{\text{proj}}, \Delta\theta \rangle \\ &\quad + O(\|\Delta\theta\|^2), \end{aligned} \quad (39)$$

where the $O(\|\Delta\theta\|^2)$ remainder is controlled via Lemma 4 and Theorem 4. The middle term vanishes in the ESS regime and is otherwise bounded by $\|\mathcal{E}_{\text{proj}}\|_2 \cdot \|\Delta\theta\|_2$.

Relation to the policy gradient theorem. Theorem 1 can be read as a special-case projection of the policy gradient theorem (Sutton et al., 1999): the long-horizon return is replaced by the finite-horizon committor χ_T , and the projection step replaces a direct Monte-Carlo estimate of the gradient. The advantage is that the projection formulation produces a tractable closed-form surrogate Φ_ρ whose parameter-space gradient is computable from forward statistics (Theorem 2), at the cost of the explicit residual $\mathcal{E}_{\text{proj}}$.

B.3 Proof of Theorem 2

Statement (restated). Let $S_i^{\text{head}} := -\alpha_i \langle \nabla_{\theta_{\text{out}}} \Phi_\rho, g_i^{\text{out}} \rangle$ with $g_i^{\text{out}} := \nabla_{\theta_{\text{out}}} \log \pi_\theta(y_i | c_i)$, and $\kappa(a, c_i) := h_a^\top h_{c_i} + \mathbf{1}_{\text{bias}}$. Then

$$\begin{aligned} S_i^{\text{head}} &= \frac{\alpha_i}{M} \sum_{a \in \mathcal{A}} \rho_a \cdot \kappa(a, c_i) \\ &\quad \cdot (\phi_a(y_i) - \bar{\phi}_a^{(i)}). \end{aligned} \quad (19 \text{ restated})$$

Proof. The proof has four steps: head gradients of anchor entropy, head gradients of token log-probability, the per-anchor inner product, and the risk-weighted sum.

Step 1 (anchor head gradients). Differentiating H_a in w_u using $\nabla_{w_u} \pi_a(v) = \pi_a(v)(\delta_{u,v} - \pi_a(u))h_a$,

$$\begin{aligned} \nabla_{w_u} H_a &= - \sum_v [\nabla_{w_u} \pi_a(v)] (\log \pi_a(v) + 1) \\ &= - \sum_v \pi_a(v) (\delta_{u,v} - \pi_a(u)) \\ &\quad \cdot (\log \pi_a(v) + 1) h_a. \end{aligned}$$

The constant term $\sum_v \pi_a(v)(\delta_{u,v} - \pi_a(u))$ vanishes by simplex normalization, and using $s_v^{(a)} = \log \pi_a(v) + H_a$,

$$\nabla_{w_u} H_a = -\pi_a(u) s_u^{(a)} h_a = -\phi_a(u) h_a, \quad (40)$$

where $\phi_a(u) := \pi_a(u) s_u^{(a)}$. An analogous calculation for the bias gives

$$\nabla_{b_u} H_a = -\phi_a(u) \mathbf{1}_{\text{bias}}. \quad (41)$$

Step 2 (token log-prob head gradients). By a standard softmax identity,

$$\begin{aligned} \nabla_{w_u} \log \pi_{c_i}(y_i) &= (\delta_{u,y_i} - \pi_{c_i}(u)) h_{c_i}, \\ \nabla_{b_u} \log \pi_{c_i}(y_i) &= (\delta_{u,y_i} - \pi_{c_i}(u)) \mathbf{1}_{\text{bias}}. \end{aligned} \quad (42)$$

Step 3 (single-anchor inner product). Summing over $u \in \mathcal{V}$ using (40), (41), (42):

$$\begin{aligned} &\langle \nabla_{\theta_{\text{out}}} H_a, g_i^{\text{out}} \rangle \\ &= \sum_u \left[-\phi_a(u) h_a^\top (\delta_{u,y_i} - \pi_{c_i}(u)) h_{c_i} \right. \\ &\quad \left. - \phi_a(u) (\delta_{u,y_i} - \pi_{c_i}(u)) \mathbf{1}_{\text{bias}} \right] \\ &= -(h_a^\top h_{c_i} + \mathbf{1}_{\text{bias}}) \\ &\quad \cdot \sum_u \phi_a(u) (\delta_{u,y_i} - \pi_{c_i}(u)) \\ &= -\kappa(a, c_i) (\phi_a(y_i) - \bar{\phi}_a^{(i)}), \end{aligned}$$

where $\bar{\phi}_a^{(i)} := \mathbb{E}_{v \sim \pi_{c_i}} [\phi_a(v)] = \sum_u \pi_{c_i}(u) \phi_a(u)$.

Step 4 (risk-weighted sum). Since $\Phi_\rho = \frac{1}{M} \sum_a \rho_a H_a$ with detached ρ_a ,

$$\nabla_{\theta_{\text{out}}} \Phi_\rho = \frac{1}{M} \sum_a \rho_a \nabla_{\theta_{\text{out}}} H_a.$$

Substituting Step 3 and applying $S_i^{\text{head}} := -\alpha_i \langle \nabla_{\theta_{\text{out}}} \Phi_\rho, g_i^{\text{out}} \rangle$ produces (19). $\square$

Forward-only structure. Every quantity in (19) is a forward statistic at $\theta^{(t)}$: hidden states h_a, h_{c_i} , policy values π_a, π_{c_i} , anchor scores ϕ_a , projection weight ρ_a . No backward through the policy is required to evaluate S_i^{head} ; the only backward pass needed is the standard GRPO update.

Interpretation of the closed form. $\kappa(a, c_i)$ measures hidden-state similarity between anchor and rollout’s prompt; large values mean the rollout’s prompt is similar to the anchor and its head update will strongly influence the anchor’s distribution. $\phi_a(y_i) - \bar{\phi}_a^{(i)}$ measures the centered surprisal of the rollout’s selected token under the anchor’s distribution: large positive values indicate the rollout selected a high-probability anchor token whose surprisal exceeds the rollout’s own expectation. The product is then risk-weighted by ρ_a across anchors and advantage-weighted by α_i .

C Body Residual and Supporting Lemmas

This section closes out the body-head factorization by providing the closed form of S_i^{body} , three supporting lemmas (Bregman identity, vocabulary-independent Hessian bound, logit-linearization remainder), and a ranking-margin theorem justifying head-only sufficiency.

C.1 Body Residual Closed Form

Theorem 3 (Body residual). *With $J_a := \partial h_a / \partial \theta_{\text{body}}$, $J_c := \partial h_c / \partial \theta_{\text{body}}$, and the body-side Gram matrix*

$$G_{\text{body}}(a, c) := W^\top J_a J_c^\top W \in \mathbb{R}^{|V| \times |V|}, \quad (43)$$

the body component of the flow score equals

$$S_i^{\text{body}} = \frac{\alpha_i}{M} \sum_a \rho_a \phi_a^\top G_{\text{body}}(a, c_i) (e_{y_i} - \pi_{c_i}). \quad (44)$$

Consequently $S_i^{\text{flow}} = S_i^{\text{head}} + S_i^{\text{body}}$.

Proof. From $z_{a,v} = w_v^\top h_a + b_v$, $\nabla_{\theta_{\text{body}}} z_{a,v} = J_a^\top w_v$, and chain-rule through the softmax,

$$\nabla_{\theta_{\text{body}}} H_a = -J_a^\top W \phi_a, \quad (45)$$

where the negative sign and ϕ_a come from the identities used in Step 1 of the head proof. Symmetrically for token log-probability at c_i ,

$$\nabla_{\theta_{\text{body}}} \log \pi_{c_i}(y_i) = J_{c_i}^\top W (e_{y_i} - \pi_{c_i}). \quad (46)$$

Their inner product is $-\phi_a^\top G_{\text{body}}(a, c_i) (e_{y_i} - \pi_{c_i})$. Risk-weighted summation and $S_i^{\text{body}} := -\alpha_i \langle \nabla_{\theta_{\text{body}}} \Phi_\rho, \nabla_{\theta_{\text{body}}} \log \pi_i \rangle$ give (44). The flow-score decomposition follows from the orthogonal split $\theta = (\theta_{\text{body}}, \theta_{\text{out}})$. $\square$

Why we do not compute S_i^{body} in the main loop.

Exact computation requires the $|V| \times |V|$ Gram matrix $G_{\text{body}}(a, c_i)$ for every (anchor, rollout) pair, costing $O(MN|V|^2)$ per step. At LM scale ($|V| \sim 10^5$) this is prohibitive. FRACT instead uses S_i^{head} and validates head sufficiency empirically through $r_i^{\text{body}} = |S_i^{\text{body}}| / (|S_i^{\text{head}}| + \varepsilon_{\text{num}})$, sampled on a subset of rollouts (§4.3.4).

Spectral reduction (sketch). Under two simplifying conditions, often approximately satisfied in trained LMs, S_i^{body} admits a head-like reduction:

1. *Separability:* $G_{\text{body}}(a, c) \approx \gamma(a, c) \cdot I$ for an effective scalar coupling.
2. *Token-diagonal action:* the dominant component of $W^\top J_a J_c^\top W$ acts as a rank-1 operator in the ϕ_a -direction.

Under both, $S_i^{\text{body}} \propto S_i^{\text{head}}$ up to an effective scaling, and r_i^{body} stays bounded. The validity of this reduction is not assumed by Theorem 1 or Theorem 2; it is empirically monitored throughout training, with median $r_i^{\text{body}} \leq 0.35$ and top-5% Jaccard agreement above 0.5 in our runs (Table 7).

Pairwise bound. For ranking, what matters is the pairwise gap $|S_i^{\text{body}} - S_j^{\text{body}}|$. By the triangle inequality and (44),

$$\begin{aligned} |S_i^{\text{body}} - S_j^{\text{body}}| &\leq \frac{1}{M} \sum_a |\rho_a| \|\phi_a\|_2 \\ &\quad \cdot \|G_{\text{body}}(a, c)\|_{\text{op}} \\ &\quad \cdot (|\alpha_i| + |\alpha_j|) \sqrt{2} \sup_c \|e_y - \pi_c\|_2, \end{aligned} \quad (47)$$

bounded by $2\sqrt{2}A_{\text{max}}\sigma_{\text{max}}G_{\text{body},\text{max}}\bar{\rho}_{|\cdot|}$ for the constants defined in the ranking margin theorem below.

C.2 Bregman Endpoint Identity

The following identity converts finite-step entropy changes into a linear part plus a KL remainder, underlying the second-order analysis of Appendix D.4.

Lemma 1 (Bregman endpoint identity). *For policies π_0, π_1 on the same support,*

$$\begin{aligned} H(\pi_0) - H(\pi_1) &= \langle \pi_1 - \pi_0, \log \pi_0 \rangle + D_{\text{KL}}(\pi_1 \parallel \pi_0). \end{aligned} \quad (48)$$

Proof. $H(\pi_0) - H(\pi_1) = \langle \pi_1, \log \pi_1 \rangle - \langle \pi_0, \log \pi_0 \rangle$. Adding and subtracting $\langle \pi_1, \log \pi_0 \rangle$:

$$\begin{aligned} H(\pi_0) - H(\pi_1) &= \langle \pi_1 - \pi_0, \log \pi_0 \rangle \\ &\quad + \langle \pi_1, \log \pi_1 - \log \pi_0 \rangle \\ &= \langle \pi_1 - \pi_0, \log \pi_0 \rangle \\ &\quad + D_{\text{KL}}(\pi_1 \parallel \pi_0). \quad \square \end{aligned}$$

Corollary 1 (Anchor entropy step decomposition). *With $\Delta_a^{\text{end}} := \pi_a(\theta + \Delta\theta) - \pi_a(\theta)$,*

$$H_a(\theta) - H_a(\theta + \Delta\theta) = \langle \Delta_a^{\text{end}}, \log \pi_a(\theta) \rangle + D_{\text{KL}}.$$

The risk-weighted sum $\sum_a \rho_a [H_a(\theta) - H_a(\theta + \Delta\theta)]$ inherits this decomposition exactly: the linear part produces the flow score, the KL part produces the Bregman correction analyzed in Appendix D.4.

C.3 Hessian Bounds on Entropy

We bound the second-order remainder when expanding H_a or $\chi_T(p_a)$ around θ . The two key constants are the operator norm of the entropy Hessian in logit space (Lemma 2) and a vocabulary-independent bound combining it with a curvature correction (Theorem 4).

Lemma 2 (Logit-space entropy Hessian). *For any $\pi = \text{softmax}(z)$ over $|V|$ tokens,*

$$\begin{aligned} \nabla_z^2 H(\pi) &= -F(\pi) + \mathcal{R}(\pi), \\ \mathcal{R}(\pi) &:= \phi\pi^\top + \pi\phi^\top - \text{diag}(\phi), \end{aligned} \quad (49)$$

with $F(\pi) = \text{diag}(\pi) - \pi\pi^\top$. Moreover $\|F(\pi)\|_{\text{op}} \leq 1/2$ for every π .

Proof. Differentiate $H(\pi(z)) = -\sum_v \pi_v \log \pi_v$ twice in z using $\partial \pi_v / \partial z_u = \pi_v(\delta_{uv} - \pi_u)$. The first derivative is $\partial H / \partial z_u = -\phi_u$. The second derivative, after collecting terms, gives $\nabla_z^2 H = -F(\pi) + \mathcal{R}(\pi)$. The bound on $\|F\|_{\text{op}}$ follows from

$$x^\top F x = \sum_v \pi_v x_v^2 - \left(\sum_v \pi_v x_v \right)^2 = \text{Var}_\pi[X],$$

$$\text{and } \text{Var}_\pi[X] \leq \frac{1}{4}(\max x - \min x)^2 \leq \frac{1}{2}\|x\|_2^2. \quad \square$$

Lemma 3 (Score-moment bounds). *With $\sigma_c^2 := \sum_v \pi_c(v)(s_v^{(c)})^2$,*

$$\|\phi^{(c)}\|_2 \leq \sigma_c, \quad \|\phi^{(c)}\|_\infty \leq \sigma_c,$$

$$\|\mathcal{R}(\pi_c)\|_{\text{op}} \leq 3\sigma_c.$$

Proof. By Cauchy–Schwarz, $\|\phi\|_2^2 = \sum_v \pi_v^2 s_v^2 \leq \max_v \pi_v \cdot \sum_v \pi_v s_v^2 \leq \sigma^2$. The ℓ_∞ bound is similar: $\|\phi\|_\infty^2 = \max_v \pi_v^2 s_v^2 \leq \pi_v \cdot \pi_v s_v^2 \leq \sigma^2$. For $\mathcal{R}$:

$$\begin{aligned} \|\phi\pi^\top\|_{\text{op}} &\leq \|\phi\|_2 \|\pi\|_2 \leq \sigma, \\ \|\pi\phi^\top\|_{\text{op}} &\leq \sigma, \\ \|\text{diag}(\phi)\|_{\text{op}} &= \|\phi\|_\infty \leq \sigma. \end{aligned}$$

By the triangle inequality, $\|\mathcal{R}\|_{\text{op}} \leq 3\sigma$. $\square$

Theorem 4 (Vocabulary-independent Hessian bound). *For any π with score variance σ ,*

$$\|\nabla_z^2 H(\pi)\|_{\text{op}} \leq 3\sigma + \frac{1}{2}. \quad (50)$$

The bound is independent of $|V|$.

Proof. Combine Lemmas 2 and 3 by the triangle inequality on $\nabla_z^2 H = -F + \mathcal{R}$. $\square$

Why vocabulary-independence matters. Naive bounds on Hessians of softmax-based functions often scale with $|V|$ via $\|s\|_2$. Theorem 4 depends only on σ , the *distributional* score variance under π , which is finite (typically $O(1)$) even for very large vocabularies. This is essential for FRACT’s residual estimates to remain meaningful at LM scale.

C.4 Ranking Margin for Head-Only Sufficiency

The diagnostic r_i^{body} checks empirically whether S_i^{head} ranks rollouts consistently with S_i^{flow} . The following result gives a deterministic sufficient condition under which the head-score *ordering* of two rollouts agrees with the ordering of their (true) leave-one-out trajectory risk differences.

Lemma 4 (Logit-linearization remainder). *Under Assumption SGD (bounded $\|\Delta\theta\|_2 \leq \eta_0$ and $\sup \|\nabla_\theta^2 z\|_{\text{op}} \leq L_J^{(2)}$), for any anchor a and any $s \in [0, 1]$,*

$$\|z_a(s) - z_a^{\text{lin}}(s)\|_2 \leq \frac{1}{2}L_J^{(2)}\|\Delta\theta\|^2 \cdot s(1-s), \quad (51)$$

where $z_a(s) := z_a(\theta + s\Delta\theta)$ and $z_a^{\text{lin}}(s) := z_a(\theta) + sJ_z(a)\Delta\theta$.

Proof (sketch). Define $\psi(s) := z_a(s) - z_a^{\text{lin}}(s)$, which satisfies $\psi(0) = \mathbf{0}$ and $\psi(1)$ is the first-order Taylor remainder. Standard second-order Taylor in a normed space gives $\|\psi(1)\|_2 \leq \frac{1}{2}L_J^{(2)}\|\Delta\theta\|^2$. For $s \in (0, 1)$ the corresponding remainder scales as $s(1-s)$ by integral-remainder applied to $\psi(s)$. $\square$

Theorem 5 (Head-only ranking margin). *Fix two rollouts i, j and assume bounded constants $A_{\max}$ (advantage magnitude), $\sigma_{\max}$ (anchor score-variance), $G_{\text{body},\max}$ (op. norm of G_{body}), $\bar{\rho}_{|\cdot|} := \frac{1}{M} \sum_a |\rho_a|$, L_z^{SGD} , $G_{\max} := \sup_a \|J_\pi(a)\|_{\text{op}} \cdot \|g_i\|_2$. If*

$$\begin{aligned} |S_i^{\text{head}} - S_j^{\text{head}}| &> 2\sqrt{2} A_{\max} \sigma_{\max} G_{\text{body},\max} \bar{\rho}_{|\cdot|} \\ &+ L_z^{\text{SGD}} \|\Delta z\|_{\mathcal{A},2} \\ &\cdot (\|\Delta z_i\| + \|\Delta z_j\|) \frac{N}{\eta} \\ &+ 2|\alpha_i \vee \alpha_j| \|\mathcal{E}_{\text{proj}}\| G_{\max}, \end{aligned} \quad (52)$$

then $\text{sign}(\Psi_T^{(-i)} - \Psi_T^{(-j)}) = -\text{sign}(S_i^{\text{head}} - S_j^{\text{head}})$.

Proof (sketch). Decompose $\Psi_T^{(-i)} - \Psi_T^{(-j)}$ into three error sources, each bounded separately: (1) body residual $S_i^{\text{flow}} - S_i^{\text{head}} = S_i^{\text{body}}$, pairwise bounded by (47), contributing the first right-hand term; (2) logit-linearization remainder (Lemma 4 with Theorem 4), contributing the second; (3) projection residual from (39) carrying $\langle \mathcal{E}_{\text{proj}}, \Delta\theta \rangle$, contributing the third. When the head-score gap exceeds the sum, the sign of the LOO difference is determined by the head-score difference. $\square$

Interpretation. Theorem 5 is the formal counterpart of the empirical observation in §4.3.4 that top-5% Jaccard agreement between S^{head} and S^{flow} rankings stays above 0.5. When all three error sources are simultaneously small—as monitored by r_i^{body} , the Hessian-bound diagnostic, and η_{proj} respectively—head-only ranking is provably faithful at the level of pairwise LOO trajectory-risk comparisons.

D Fisher-Isotropic Curvature

This section substantiates the curvature interpretation of S_i^{FC} alluded to in §3.7. The interpretation requires an isotropic-curvature assumption on the learned committor; we make it explicit, derive the resulting second-order trajectory decomposition, and identify the term S_i^{FC} approximates. Crucially, S_i^{FC} retains a forward-only mechanistic meaning—risk-weighted tangent-Fisher coherence of push-forward drifts—*regardless* of whether the curvature assumption holds; the second-order reading is an additional interpretation, monitored via η_{FIC} .

D.1 Softmax–Fisher Identity, Dual Reading

Lemma 5 (Softmax–Fisher identity, dual reading). *With $\delta := J_\pi(a)\Delta\theta \in T_{\pi_a}\Delta$ and $\Delta z := J_z(a)\Delta\theta \in \mathbb{R}^{|V|}$,*

$$\delta^\top F_\epsilon^+(\pi_a) \delta = \Delta z^\top F_\epsilon(\pi_a) \Delta z. \quad (53)$$

Proof. From $J_\pi = FJ_z$ (writing $F := F_\epsilon(\pi_a)$), $\delta = F\Delta z$. Then

$$\begin{aligned} \delta^\top F_\epsilon^+ \delta &= (F\Delta z)^\top F_\epsilon^+ (F\Delta z) \\ &= \Delta z^\top F F_\epsilon^+ F \Delta z \\ &= \Delta z^\top F \Delta z, \end{aligned}$$

using $F F_\epsilon^+ F = F$ on $T_{\pi_a}\Delta$. $\square$

Geometric reading. Lemma 5 is the primal–dual pairing: the tangent-space (F_ϵ^+) norm of the policy drift δ equals the cotangent-space (F_ϵ) norm of the logit direction Δz . This duality makes $q^\top F_\epsilon q$ in S_i^{FC} equal to the tangent-side inner product $\langle Fq_i, Fq_j \rangle_{F_\epsilon^+}$ of push-forward policy drifts, allowing the score to be computed entirely from cotangent-side logit directions.

D.2 Exact Second-Order Expansion of the Committor

Lemma 6 (Second-order expansion of χ_T). *For sufficiently smooth χ_T and any $\Delta \in T_{p_a}\Delta$,*

$$\begin{aligned} \chi_T(p_a + \Delta) - \chi_T(p_a) &= \langle \nabla_p \chi_T(p_a), \Delta \rangle \\ &+ \frac{1}{2} \Delta^\top \nabla_p^2 \chi_T(p_a) \Delta \\ &+ o(\|\Delta\|^2). \end{aligned} \quad (54)$$

If, in addition, χ_T factors as $\chi_T(p) = F_T(H(p))$ for some univariate F_T , then

$$\begin{aligned} \nabla_p^2 \chi_T(p_a) &= F_T''(H(p_a)) \nabla_p H \nabla_p H^\top \\ &+ F_T'(H(p_a)) \nabla_p^2 H. \end{aligned} \quad (55)$$

Proof. The first identity is the standard Taylor expansion of a C^2 function. The factored form follows from $\nabla_p \chi_T = F_T'(H) \nabla_p H$ and differentiating once more via the product rule. $\square$

D.3 Fisher-Isotropic Curvature Approximation

The exact $\nabla_p^2 \chi_T(p_a)$ is rarely available analytically and is expensive to estimate from data. We use a single-parameter isotropic approximation.

Definition 1 (Fisher-Isotropic Curvature). *At anchor a , the committor Hessian is Fisher-isotropic with curvature $\kappa_a \in \mathbb{R}$ if*

$$\nabla_p^2 \chi_T(p_a) \approx \kappa_a F_\epsilon^+(p_a) \quad (56)$$

as operators on $T_{p_a} \Delta$.

Equation (56) is a *tangent-metric* statement, consistent with $\nabla_p^2 \chi_T$ acting on tangent vectors. Two observations:

1. When $\chi_T = F_T(H)$, the exact Hessian (Lemma 6) is the sum of a rank-1 entropy-direction term and the (rescaled) entropy Hessian. The FIC approximation (56) replaces this with a single scalar multiple of F_ϵ^+ —a substantial simplification, justified when the entropy direction and the natural-gradient flow are sufficiently aligned.
2. The approximation introduces a quality residual

$$\begin{aligned} \eta_{\text{FIC}}(p_a) &:= \frac{\|\nabla_p^2 \hat{\chi}_T(p_a) - \kappa_a F_\epsilon^+(p_a)\|_{\text{Frob}}}{\|\nabla_p^2 \hat{\chi}_T(p_a)\|_{\text{Frob}} + \varepsilon_{\text{num}}}, \end{aligned} \quad (57)$$

reported as an empirical diagnostic (Appendix J).

Fitting κ_a . Given the learned $\hat{\chi}_T$, the FIC curvature is fit by least-squares: $\kappa_a := \arg \min_{\kappa} \|\nabla_p^2 \hat{\chi}_T(p_a) - \kappa F_\epsilon^+(p_a)\|_{\text{Frob}}^2$, with closed-form solution

$$\kappa_a = \frac{\text{tr}(F_\epsilon^+ \nabla_p^2 \hat{\chi}_T)}{\text{tr}(F_\epsilon^+ F_\epsilon^+)}.$$

The fitting is performed per anchor at a subsampled cadence during training; in practice κ_a is approximately constant over training windows of ~ 10 steps, allowing amortization.

D.4 Curvature Interpretation of S_i^{FC}

We now show that under FIC, S_i^{FC} approximates the per-rollout contribution to a cumulative second-order trajectory-curvature term. Fix an anchor a . Write $\delta_i^{(a,t)} := F_\epsilon(\pi_a^{(t)})q_i$ for the push-forward policy drift induced by token i at step t , $\delta^{(a,t)} := \sum_i \delta_i^{(a,t)}$ for the per-step drift, and $\Delta_a^{(T)} := \sum_{t=0}^{T-1} \delta^{(a,t)}$ for the cumulative anchor drift.

Theorem 6 (Cumulative drift decomposition under FIC). *Under Definition 1 at anchor a , the T -horizon committor change satisfies*

$$\begin{aligned} \chi_T(p_a^{(t+T)}) - \chi_T(p_a^{(t)}) &\stackrel{\text{FIC}}{\approx} \rho_a \langle \nabla_p H(p_a), \Delta_a^{(T)} \rangle \\ &\quad + \frac{\kappa_a}{2} \|\Delta_a^{(T)}\|_{F_\epsilon^+}^2 + r^{(\text{FIC})} \\ &\quad + o(\|\Delta_a^{(T)}\|^2), \end{aligned} \quad (58)$$

where $r^{(\text{FIC})}$ collects the approximation residual.

Proof (sketch). Apply Lemma 6 with $\Delta = \Delta_a^{(T)}$. The linear term decomposes via Proposition 1 as $-\rho_a \langle \nabla_p H, \Delta_a^{(T)} \rangle + \langle r_a, \Delta_a^{(T)} \rangle$ (the first survives the body, the second contributes to $r^{(\text{FIC})}$). The quadratic term $\frac{1}{2} \Delta_a^{(T)\top} \nabla_p^2 \chi_T \Delta_a^{(T)}$ is approximated by $\frac{\kappa_a}{2} \|\Delta_a^{(T)}\|_{F_\epsilon^+}^2$ via FIC. The sign convention of ρ_a accounts for the leading minus sign. $\square$

Decomposition of the quadratic term. Expanding the cumulative tangent norm,

$$\begin{aligned} \|\Delta_a^{(T)}\|_{F_\epsilon^+}^2 &= \sum_t (\mathcal{S}_a^{(t)} + 2\mathcal{C}_a^{(t)}) \\ &\quad + 2 \sum_{t < t'} \langle \delta^{(a,t)}, \delta^{(a,t')} \rangle_{F_\epsilon^+}, \end{aligned} \quad (59)$$

where

$$\begin{aligned} \mathcal{S}_a^{(t)} &:= \sum_i \|\delta_i^{(a,t)}\|_{F_\epsilon^+}^2 \quad (\text{self-energy}), \\ \mathcal{C}_a^{(t)} &:= \sum_{i < j} \langle \delta_i^{(a,t)}, \delta_j^{(a,t)} \rangle_{F_\epsilon^+} \\ &\quad (\text{within-step coherence}). \end{aligned}$$

By Lemma 5,

$$\langle \delta_i^{(a,t)}, \delta_j^{(a,t)} \rangle_{F_\epsilon^+} = q_i^\top F_\epsilon(\pi_a) q_j,$$

which is the inner term in S_i^{FC} (23).

Per-rollout attribution. Substituting into Theorem 6 and risk-weighting ρ_a on the quadratic gives a per-step within-coherence contribution of the form

$$\rho_a \kappa_a \mathcal{C}_a^{(t)} = \rho_a \kappa_a \sum_{i < j} q_i^\top F_\epsilon(\pi_a) q_j.$$

Symmetrizing over (i, j) pairs and absorbing the advantage weights α_i , the context–anchor coupling

$\kappa(a, c_i)$, and the scaling constants into β , one recovers (23):

$$S_i^{\text{FC}} = \frac{\alpha_i}{M} \sum_a \rho_a \kappa(a, c_i) \cdot q_i^\top F_\epsilon(\pi_a) (u_a - \alpha_i \kappa(a, c_i) q_i),$$

where $u_a := \sum_j \alpha_j \kappa(a, c_j) q_j$ is the batch drift and the subtraction $-\alpha_i \kappa(a, c_i) q_i$ removes the self-interaction $i = j$. Under FIC, S_i^{FC} is therefore a forward-only approximation of token i 's contribution to second-order trajectory curvature in χ_T .

Without FIC. Even when FIC is violated (large η_{FIC}), S_i^{FC} retains the mechanistic meaning of risk-weighted tangent-Fisher coherence of push-forward policy drifts: a large positive S_i^{FC} flags rollouts whose policy update aligns with the batch's aggregate update at the anchor, weighted by the projection coefficient ρ_a . This interpretation requires no curvature assumption.

E Multi-Feature Committor

Proposition 1 and Theorem 1 are stated for an arbitrary committor χ_T , but the practical default uses $\hat{\chi}_T$ trained as a logistic on three entropy-aligned core features ($H, m(p), g_{12}(p)$; see §3.2 and Appendix K.3). When the diagnostic ρ_{ratio}^- is large—indicating that the committor varies orthogonally to or against entropy at many anchors—this core feature set becomes unfaithful. The multi-feature variant addresses this by extending the feature set with reference-relative and score-variance terms, without changing the projection geometry.

E.1 Construction

Let $x(p) \in \mathbb{R}^{d_x}$ be a feature vector summarizing relevant simplex properties of p . We use features of three types:

- **Entropy-like:** $H(p), \sum_v p(v)^\alpha$ for $\alpha \in \{2, 1/2\}$ (Renyi entropies).
- **Mode-concentration-like:** $m(p) := \max_v p(v), \sum_v p(v) \mathbf{1}[p(v) \geq \tau]$ for $\tau \in \{0.1, 0.3\}$.
- **Reference-relative:** $D_{\text{KL}}(p \| p_{\text{base}})$ for a fixed reference policy p_{base} (e.g., initial pre-RL policy).

The multi-feature committor estimator is

$$\hat{\chi}_T(p) = \sigma(w^\top x(p)), \quad (60)$$

with $w \in \mathbb{R}^{d_x}$ fit by logistic regression on the trajectory pairs $(p^{(t)}, Y^{(t,T)})$ accumulated by the replay-update block of Algorithm 1. The sigmoid output keeps $\hat{\chi}_T$ in $[0, 1]$ as a committor must be.

E.2 Projection

Proposition 1 is unchanged: replace $\nabla_p \chi_T$ by $\nabla_p \hat{\chi}_T$, which now has contributions from every feature:

$$\nabla_p \hat{\chi}_T(p_a) = \sigma'(w^\top x(p_a)) \sum_{k=1}^{d_x} w_k \nabla_p x_k(p_a). \quad (61)$$

The projection (11) uses this multi-feature gradient on its numerator. Each gradient $\nabla_p x_k$ is a closed-form simplex vector, so no additional autograd cost is introduced:

$$\begin{aligned} \nabla_p H &= -\log p - \mathbf{1}, \\ \nabla_p \sum_v p(v)^\alpha &= \alpha p^{\alpha-1}, \end{aligned}$$

$$\nabla_p m(p) = e_{v^*}, \quad v^* = \arg \max_v p(v),$$

$$\nabla_p D_{\text{KL}}(p \| p_{\text{base}}) = \log p - \log p_{\text{base}} + \mathbf{1}.$$

(For non-differentiable points such as $\nabla_p m$, we use a subgradient; in practice ties at the mode are essentially measure-zero events for trained LMs.)

E.3 When to Use

We recommend switching to the multi-feature variant when ρ_{ratio}^- exceeds an operational threshold (we use 0.3 as a default, with sensitivity reported in §4), or when η_{proj} exceeds 0.5 on a substantial fraction of anchors. In our main experiments these regimes occur rarely; the univariate-in-entropy default is sufficient for the reported results in §4.

E.4 Compatibility with Theorems

Theorem 1 and Theorem 2 are proved under the assumption that ρ_a, r_a come from *some* projection decomposition; they are otherwise agnostic to the source of $\nabla_p \chi_T$. Replacing the univariate-in-entropy estimator with (60) therefore preserves both identities verbatim, with ρ_a now reflecting projection of a richer multi-feature gradient onto the entropy direction. The body residual closed form (Theorem 3) and the ranking margin theorem (Theorem 5) also hold unchanged.

E.5 Why Project onto Entropy, Not onto a Multi-Feature Direction

A natural alternative is to project $\nabla_p \hat{\chi}_T$ onto a multi-feature direction in the simplex, rather than

onto $\nabla_p H$. We chose the entropy direction for three reasons:

1. *Theorem 2 requires the entropy direction.* The forward-only closed form of S_i^{head} relies on $\Phi_\rho = \frac{1}{M} \sum_a \rho_a H_a$ being a sum of *entropies*, so that the centered surprisal $\phi_a(y_i) - \bar{\phi}_a^{(i)}$ emerges from the head gradient. Projecting onto a multi-feature direction would require a more complex surrogate without a corresponding closed form.
2. *Entropy is the canonical collapse measure.* Entropy is the universally accepted scalar measure of policy concentration, with a direct interpretation in terms of generation diversity. Other features can refine the committor but the projection target is best kept canonical.
3. *The multi-feature committor still uses entropy through projection.* Even when $\hat{\chi}_T$ depends on multiple features, the projection coefficient ρ_a extracts the *component* of the committor gradient that aligns with entropy. Information from other features enters indirectly through how they shape $\hat{\chi}_T$'s predictions but does not appear as a separate projection axis. This is the design choice that keeps FRACT both theoretically clean and computationally efficient.

F Experimental Appendix Overview

Table 12 maps each appendix section to its content and the corresponding location in the main text.

Reporting convention. The §4.3 component ablations use the 5-benchmark unweighted mean of Acc@16 at the FRACT-head multi base (absolute Avg 0.41–0.44 at 1.5B); the §4.4 E2E tables use the same mean for full GRPO + variant pipelines (Avg 0.26–0.34 at 1.5B). The main-text aggregate std in §4.4 combines benchmark-level cell dispersion and seed-level variation; the seed-only std used for paired comparisons is reported in Appendix W. The two need not match. All Pass@ k values in §4.5 and Appendix X are directly measured under our protocol for all three methods. The normalized residual gain $\text{RG}_{1 \rightarrow 64} = (\text{Pass}@64 - \text{Pass}@1) / (1 - \text{Pass}@1)$ measures how much of the residual failure mass at $k = 1$ is recovered by additional samples and is invariant to absolute Pass@1 level.

Seed allocation. Table 13 lists seed budgets for each experiment.

Algorithm 1 FRACT score per GRPO step

Require: Current $\theta^{(t)}$; batch $\{(c_i, y_i, \alpha_i)\}_{i=1}^N$; anchor set $\mathcal{A}$; learned committor $\hat{\chi}_T$.

- 1: **for** $a \in \mathcal{A}$ **do** ▷ anchor pre-compute
- 2: Compute $h_a, \pi_a, \log \pi_a, H_a, \phi_a$.
- 3: Build $F_e(\pi_a)$ via (9).
- 4: $\nabla_p \hat{\chi}_T(p_a) \leftarrow$ autograd of $\hat{\chi}_T$ at π_a .
- 5: $\nabla_p H(p_a) \leftarrow -\log \pi_a - \mathbf{1}$.
- 6: $\rho_a \leftarrow$ Eq. (11); $r_a \leftarrow \nabla_p \hat{\chi}_T + \rho_a \nabla_p H$.
- 7: $\eta_{\text{proj}}(p_a) \leftarrow$ Eq. (27). ▷ diagnostic
- 8: **end for**
- 9: **for** $i = 1, \dots, N$ **do** ▷ per-rollout score
- 10: $\kappa_{i,a} \leftarrow h_a^\top h_{c_i} + \mathbf{1}_{\text{bias}}$.
- 11: $S_i^{\text{head}} \leftarrow$ Eq. (19).
- 12: $S_i^{\text{FC}} \leftarrow$ Eq. (23).
- 13: $S_i^* \leftarrow S_i^{\text{head}} + \beta S_i^{\text{FC}}$.
- 14: **end for**
- 15: $\mathcal{P} \leftarrow \{i : S_i^* > 0\}$.
- 16: $m_i \leftarrow \mathbb{1}[i \in \mathcal{P} \cap \text{TopK}_{k\%}(\{S_j^*\}_{j \in \mathcal{P}})]$.
- 17: Update $\theta^{(t+1)}$ with $\mathcal{L} = \mathcal{L}_{\text{GRPO}} + \lambda \cdot \text{masked KL}$.
- 18: **if** $t \geq T$ **then** ▷ replay-update of $\hat{\chi}_T$
- 19: Add $\{(p_a^{(t-T)}, Y_a^{(t-T, T)})\}_{a \in \mathcal{A}}$ to the $\hat{\chi}_T$ replay buffer.
- 20: Take one AdamW step on a $\hat{\chi}_T$ mini-batch (§K.3).
- 21: **end if**

G Experimental Setup Details

Table 14 gives the full configuration referenced in §4.1.

H FRACT Algorithm

Algorithm 1 gives the full pseudocode summarized in §3.8.

I Hyperparameter Sweep

All “default” rows match the §4.3.2 configuration ($k = 5\%$, $\lambda = 5 \times 10^{-2}$, $\beta = 0.5$): Avg Acc@16 = 0.438 ± 0.011 .

I.1 Sweep over selective KL budget k

Table 15 sweeps $k \in \{1, 3, 5, 10, 20\}\%$ at 1.5B with $\beta = 0.5$ and $\lambda = 5 \times 10^{-2}$; $k = 5\%$ achieves the best balance across Acc@16, $\text{RG}_{1 \rightarrow 64}$, and ΔH .

I.2 Sweep over KL coefficient λ

Table 16 sweeps $\lambda \in \{1 \times 10^{-2}, 2.5 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}\}$; the default 5×10^{-2} is jointly best on Acc@16 and RG.

I.3 Sweep over Fisher-coherence weight β

This sweep duplicates §4.3.2 Table 5 for cross-reference. $\beta = 0.5$ is jointly optimal across Acc@16 (0.438), predictive AUC (0.74), and $\text{RG}_{1 \rightarrow 64}$ (0.53); $\beta \geq 1.0$ already shows declining

App.	Content	Main link
I	Hyperparameter sweep: $k, \lambda, \beta, H_{\min}$	§4.1 , §4.3.2
J	Empirical FIC diagnostics	§4.3.4
K	Weight estimator variants	§4.3.1
L	ESS diagnostic and multi-feature ablation	§4.3.1
M	Anchor strategy full results and M -sensitivity	§4.3.3
N	Horizon T -sensitivity	§4.1 , §4.2
O	Committer generalization and warmup reuse	§4.1
P	Single-token deletion sub-experiment	§4.2.3 , §4.3.4
Q	Implementation consistency: full results	§4.2.1
R	Small- K Monte Carlo verification	§4.2.1
S	Predictive validity visualization	§4.2.2
T	Diagnostic quadrants	§4.3.4
U	RLIF generalization	§4.4
V	Compute decomposition and memory profile	§4.6
W	Per-seed E2E results and paired comparisons	§4.4
X	Full Pass@ k results for 1.5B / 3B / 7B	§4.5
Y	Failure modes and negative-result analysis	§4.7

Table 12: Appendix map.

Experiment	Scale	Seeds	Component	Setting
§4.2.1 Impl. consistency	0.5B	3	Main models	Qwen2.5-Base 1.5B / 3B / 7B
§4.2.2 Entropy-label validity	1.5B	5	Toy verification	Qwen2.5-0.5B
§4.2.2 Pass@1-drop validity	1.5B	5	RL algorithm	GRPO
§4.2.3 Score-guided intervention	1.5B	5	Training data	MATH train, 7,500 problems
§4.3.1 Projection-weight ablation	1.5B	5	Training steps	117
§4.3.2 FC weight ablation	1.5B	3	Rollouts / step	512 (64 prompts $\times$ 8 gen.)
§4.3.3 Anchor strategy ablation	1.5B	3	Max completion length	3072
§4.3.4 Diagnostics	1.5B	5	Optimizer / LR	AdamW / 3×10^{-6}
§4.4 E2E benchmark-wise	1.5B/3B/7B	5	Precision / Parallel.	bf16 / FSDP + grad ckpt
§4.5 Pass@ k scaling (main)	7B	5	Anchors	$M = 8$ MATH-format probes
§4.5 Pass@ k cross-scale	3B	5	Committer warmup	100-step pilot GRPO (held out)
§4.5 Pass@ k cross-scale	1.5B	3	Committer estimator	$\hat{\chi}_T$ logistic (Appendix K.3)
			Collapse threshold	$H_{\min} = \gamma H_a^{(0)}$, $\gamma = 0.5$
			Horizon	$T = 20$ GRPO steps
			KL budget	$k = 5\%$, $\lambda = 5 \times 10^{-2}$
			Defaults	signed ρ^{proj} , $\beta = 0.5$

Table 13: Seed allocation across experiments. Primary E2E comparisons use matched seeds across methods.

Table 14: Experimental setup. The collapse threshold $H_{\min}$ is set per anchor as a fraction γ of its pre-training entropy $H_a^{(0)}$; sensitivity to γ is in [Appendix I.4](#).

returns. We retain $\beta = 0.5$ as the default without invoking joint-criterion arbitration; it is the Acc-best, AUC-best, and RG-best configuration.

I.4 Sweep over collapse threshold $H_{\min}$ (committer label)

The collapse threshold $H_{\min}$ determines the ground-truth committer label $Y^{(t,T)} = \mathbb{1}[\exists \tau \in (t, t+T] : H(\pi_\tau) \leq H_{\min}]$ and therefore directly affects the supervision target for $\hat{\chi}_T$. We parameterize $H_{\min} = \gamma \cdot H_a^{(0)}$ per anchor, where $H_a^{(0)}$ is the anchor’s entropy at the pre-training checkpoint. The default $\gamma = 0.5$ corresponds to “policy has dropped to half of its initial entropy at this anchor”; smaller γ requires deeper collapse, larger γ accepts shallower entropy drops as positive labels.

The sweep shows a mild inverted-U pattern: $\gamma = 0.5$ gives the highest mean Acc@16 and AUC, but the Acc@16 differences (≤ 0.010) are at the edge of one SD. We therefore interpret this as a

robustness check rather than evidence that $\gamma = 0.5$ is uniquely optimal: strict thresholds ($\gamma = 0.3$) yield too few positive examples ($\approx 10\%$ of anchored windows), starving the committer of supervision; loose thresholds ($\gamma = 0.7$) inflate the positive rate to 40%, where the committer conflates “near collapse” with “already collapsed.” The default $\gamma = 0.5$ places the threshold at the midpoint of the policy’s available entropy range. The bounded Acc@16 sensitivity (± 0.010) supports the [§4.4](#) effectiveness claims being robust to moderate $H_{\min}$ variation; we do not claim the inverted-U is statistically resolved at the 3-seed budget.

k	Avg Acc@16	RG _{1→64} ↑	ΔH from base
1%	0.422±.010	0.46±.04	−0.15±.03
3%	0.435±.011	0.49±.04	−0.13±.03
5%	0.438±.011	0.50±.04	−0.12±.04
10%	0.434±.014	0.47±.05	−0.11±.04
20%	0.421±.018	0.42±.06	−0.10±.05

Table 15: k -sweep at 1.5B, FRACT-full, fixed $\beta = 0.5$ and $\lambda = 5 \times 10^{-2}$, 5 seeds. $k = 5\%$ best balance.

λ	Avg Acc@16	RG _{1→64} ↑	Mean KL
1×10^{-2}	0.426±.010	0.44±.05	low
2.5×10^{-2}	0.434±.011	0.48±.04	moderate
5×10^{-2}	0.438±.011	0.50±.04	moderate
1×10^{-1}	0.430±.014	0.51±.04	high

Table 16: λ -sweep at 1.5B, 5 seeds. Default 5×10^{-2} is jointly best on Acc@16 and RG.

J Empirical FIC Diagnostics

The pairwise Fisher interaction, averaged over the M anchors:

$$G_{\text{add}} \approx \frac{1}{M} \sum_{a=1}^M \sum_{i < j} \delta_{i,a}^\top F_\epsilon^+(\pi_a) \delta_{j,a}, \quad (62)$$

where $F_\epsilon^+(\pi_a)$ is the ϵ -regularized pseudo-inverse of the simplex Fisher metric at π_a (i.e., on the categorical simplex $\Delta^{|V|-1}$ rather than the parameter Fisher in $\mathbb{R}^P$), and $\delta_{i,a}$ is the push-forward distributional drift induced by token instance i at anchor a . The outer average reflects multi-anchor aggregation; the inner sum captures pairwise interactions at the same anchor.

J.1 FIC quality

Table 18 reports FIC-quality diagnostics in the 0.5B toy setting across 3 seeds, including Spearman correlation of C^{FC} with measured drift and top-5% overlap with oracle coherence.

J.2 Second-order interpretation

$S_i^* = S_i^{\text{head}} + \beta S_i^{\text{FC}}$. The expected intervention pattern: selecting tokens by S^* should outperform selecting by S^{head} on average, while S^{FC} alone (the pure pairwise interaction term) should be weaker than S^{head} . The relation refers to *downstream intervention effect*, not raw score magnitude.

K Weight Estimator Variants

This appendix catalogs the candidate forms for the committor weight estimator $\hat{\chi}_T$ (§K.1), reports

their calibration quality (§K.2), and specifies the training schedule used for the default logistic estimator (§K.3).

K.1 Variants

Table 19 summarizes five candidate forms for $\hat{\chi}_T$ with their trade-offs; we use the logistic-on-core-features default.

K.2 Calibration

Table 20 compares Brier, ECE, AUC, and AUPRC across the five variants; the default logistic is causal, smooth, and minimally parameterized, while the multi-feature variant achieves +0.02–0.03 AUC and serves only as the ESS-violation fallback (Appendix L).

K.3 Training schedule for the default estimator

The default committor estimator $\hat{\chi}_T$ used throughout the paper is a logistic regression on a three-dimensional feature vector with anchor-relative offsets, fit as follows.

Architecture. $\hat{\chi}_T(p) = \sigma(a_0 + a_1(\tau_H - H(p)) + a_2(m(p) - \tau_m) + a_3(g_{12}(p) - \tau_g))$, where $m(p) := \max_v p(v)$ is the mode probability and $g_{12}(p) := p_{(1)}(v) - p_{(2)}(v)$ is the top-two gap. The thresholds (τ_H, τ_m, τ_g) are anchor-specific and set to the warmup-median value of the corresponding feature; $(a_0, \dots, a_3)$ are fit by maximum likelihood on labeled trajectory pairs.

Warmup phase. Before main FRACT training begins, $\hat{\chi}_T$ is fit on a held-out pool of trajectory pairs $(p_a^{(s)}, Y_a^{(s,T)})$. The default pool is the “GRPO + mixed” source of Appendix O.1: a 100-step pilot GRPO run on the same training distribution, combined with a small held-out batch of FRACT previous-checkpoint trajectories where available, yielding approximately 640 labeled pairs (the pool is constructed only from steps with $s + T$ within the pilot horizon so that every pair has its T -step label observed). Warmup fitting uses scikit-learn’s logistic regression with L2 regularization (inverse strength $C = 1.0$) and converges in < 5 seconds at all three model scales. The thresholds (τ_H, τ_m, τ_g) are set to the warmup-median value of each feature, anchor by anchor. The pilot run is held out from main training to prevent temporal leakage; see Appendix O.2.

γ	$H_{\min}/H_a^{(0)}$	Acc@16	AUC	Label + rate
0.3	strict	0.428 $\pm$.013	0.70 $\pm$.04	$\approx$ 10%
0.5	default	0.438$\pm$.011	0.74$\pm$.04	$\approx$ 22%
0.7	loose	0.431 $\pm$.012	0.71 $\pm$.05	$\approx$ 40%

Table 17: $H_{\min}$ sensitivity via γ -sweep, 1.5B, 3 seeds. “Label + rate” is the fraction of anchored windows labeled positive ($Y_a^{(t,T)} = 1$) under each threshold. Label-positive rates match the AUPRC baseline (Table 2, Random 0.22 at $\gamma = 0.5$) within noise.

Metric	Range
Spearman: C^{FC} vs. measured drift	0.45–0.62
Top-5% overlap with oracle coherence	0.28–0.42
Random top-5% overlap	0.05–0.08
Mean pairwise cosine, top C^{FC}	0.07–0.16
Mean pairwise cosine, random	0.00–0.04

Table 18: FIC quality, 0.5B toy setting, 3 seeds.

Online update phase. During FRACT training, $\hat{\chi}_T$ receives the newly available trajectory pair $(p_a^{(t-T)}, Y_a^{(t-T,T)})$ for each anchor at every GRPO step $t \geq T$, producing $M = 8$ new pairs per step. Online updates are performed using AdamW (learning rate 10^{-3} , weight decay 10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$) on mini-batches of 64 pairs sampled with replacement from a rolling buffer of the most recent 512 available pairs (FIFO eviction). One AdamW step is taken per GRPO step. We use AdamW rather than LBFGS or Newton-step despite the small parameter count: the streaming label regime makes batch second-order methods awkward, and momentum smoothing absorbs the noise of single-step labels from $M = 8$ anchors. The thresholds (τ_H, τ_m, τ_g) are held fixed at their warmup values throughout main training; only the four logit coefficients $(a_0, \dots, a_3)$ are updated.

Compute cost. Online updates contribute $< 0.1\%$ to wall-clock at all three scales, dominated entirely by the GRPO forward/backward passes; see Appendix V for the full decomposition.

L ESS Diagnostic and Multi-feature Ablation

ESS denotes the **entropy-sufficiency assumption**: collapse risk at a given token can be predicted primarily from entropy-aligned features (token entropy $H(p)$, margin $m(p)$, top-2 gap $g_{12}(p)$). When this holds, the default logistic committor suffices; when it fails, additional features (KL divergence from base, log-prob variance) improve predictive AUC.

L.1 ESS diagnostic

Let $\text{ESSGap} := \text{AUC}(\hat{\chi}_{\text{multi}}) - \text{AUC}(\hat{\chi}_{\text{entropy}})$. The corresponding entropy-only and multi-feature results are reported in Table 22.

L.2 Multi-feature fallback

Feature vector:

$$x(p) = [H(p), m(p), g_{12}(p), D_{\text{KL}}(p \parallel p_{\text{base}}), \text{Var}_{v \sim p}[\log p(v)]].$$

M Anchor Strategy Full Results

This appendix expands §4.3.3 with the full-pipeline benchmark-wise breakdown (§M.1) and the anchor-count sensitivity (§M.2).

M.1 Full benchmark-wise comparison

Table 23 reports anchor-strategy results at the full-pipeline scale, with the High-risk row matching the §4.4 FRACT-full Avg of 0.342.

This table reports full-pipeline Acc@16 on the §4.4 scale, whereas §4.3.3 Table 6 reports ablation-base quantities (component ablation against the FRACT-head multi base, before full E2E composition). The absolute values therefore differ (Table 23 High-risk 0.342 vs. Table 6 High-risk 0.439); within-table comparisons remain consistent but cross-table absolute numbers should not be compared directly.

M.2 M-sensitivity

Table 24 sweeps the anchor count $M \in \{2, 4, 8, 16\}$ at the ablation-base scale; $M = 8$ is retained as the default since $M = 16$ yields only $+0.002$ Acc@16 (within one SD) at $\sim 3\%$ additional overhead.

N Horizon Sensitivity

Table 25 sweeps the committor horizon $T \in \{5, 10, 20, 40\}$; $T = 20$ is jointly best on AUC, AUPRC, and Acc@16, and is retained as the default.

Variant	Form	Trade-off
Logistic-on-features	$\sigma(a_0 + a^\top x)$	simple, stable; limited expressivity
Isotonic	monotone bins	well calibrated; less smooth
Brownian hitting	first-passage approx.	interpretable; stronger assumptions
MLP	shallow network	expressive; overfit risk
Multi-feature logistic	$H, m, g_{12}, \text{KL}, \text{Var}$	strong fallback; more features

Table 19: Weight estimator variants. Default: $\hat{\chi}_T(p) = \sigma(a_0 + a_1(\tau_H - H(p)) + a_2(m(p) - \tau_m) + a_3(g_{12}(p) - \tau_g))$.

Estimator	Brier↓	ECE↓	AUC↑	AUPRC↑
Entropy fallback	.18±.03	.090±.02	.61±.04	.31±.04
Logistic	.13±.02	.045±.015	.70±.04	.44±.05
Isotonic	.12±.03	.038±.018	.69±.05	.43±.05
Brownian	.15±.03	.060±.02	.66±.05	.39±.05
MLP	.12±.03	.050±.02	.72±.05	.45±.06
Multi-feature log.	.11±.02	.040±.015	.73±.04	.46±.05

Table 20: Calibration comparison, 1.5B, 5 seeds. The default logistic-on-core-features estimator is causal, smooth, and minimally parameterized; the multi-feature variant achieves +0.02–0.03 AUC at the cost of audit complexity and is used only as the ESS-violation fallback (Appendix L).

ESSGap	Interpretation
< 0.02	multi-feature unnecessary
0.02–0.05	multi-feature recommended
> 0.05	multi-feature required

Table 21: ESS diagnostic regimes.

Method	AUC	Acc@16	RG _{1→64} ↑
FRACT default	.74±.04	.438±.011	.53±.04
FRACT multi-feat.	.76±.04	.442±.012	.54±.04

Table 22: Multi-feature fallback effect, 1.5B, 5 seeds.

O Commitor Generalization and Warmup Reuse

This appendix studies how the committor estimator $\hat{\chi}_T$ generalizes across warmup conditions (§O.1) and verifies absence of temporal leakage (§O.2).

O.1 Warmup source

Table 26 compares three warmup sources for fitting $\hat{\chi}_T$; mixing GRPO trajectories with FRACT previous-checkpoint data is jointly best.

O.2 Temporal leakage check

Table 27 contrasts the causal $\hat{\chi}_T$ estimator (used throughout the paper) against a non-causal oracle reported as an invalid upper bound for reference.

The oracle gap (0.80 vs. 0.74, +0.06) quantifies the amount of future-information leakage that would be obtained by fitting the committor non-causally on the same trajectories used for scoring. All FRACT scores reported in this paper use the

causal $\hat{\chi}_T$ estimator (trained only on pairs with $s \leq t - T$, §3.2); the oracle row is reported strictly as an invalid upper bound for reference.

P Single-token Deletion Sub-experiment

Table 28 reports the single-token deletion diagnostic at 1.5B across 3 seeds, comparing entropy preservation when removing tokens selected by FRACT, token-entropy, or random ranking.

This experiment is a local diagnostic rather than a causal estimate of training dynamics; deleting a single token does not reproduce the multi-token selective-KL intervention used in training and should not be interpreted as a counterfactual estimate of FRACT’s training effect.

Companion visualization (Figure 3). Figure 3 plots the score-guided intervention effect $\Delta \bar{H}_T$ vs. intervention budget k , complementing Table 3 in §4.2.3.

Q Implementation Consistency: Full Results

This appendix expands the implementation-consistency summary in §4.2.1, which verifies that the closed-form FRACT score (as coded in our repository) matches the analytical expression evaluated via PyTorch autograd on the same model state. This is a numerical-faithfulness check at the level of the implementation; it is *not* a full Monte-Carlo verification of the trajectory-gradient identity (Appendix R).

Anchor	GSM8K	MATH500	AMC	AIME24	AIME25	Avg	AUC
High-risk	.781	.603	.187	.085	.052	.342	.71
High-entropy	.785	.605	.184	.080	.053	.341	.69
First-completion	.768	.585	.169	.067	.045	.327	.66
Prompt-level	.762	.578	.161	.061	.040	.320	.65
Random	.748	.566	.150	.052	.034	.310	.62

Table 23: Anchor strategy benchmark-wise Acc@16, 1.5B, 3 seeds (full-pipeline scale, matches §4.4 at 1.5B). The High-risk row matches §4.4 FRACT-full Avg = 0.342.

M	Avg Acc@16	AUC	Overhead
2	0.425±.013	0.67±.05	1.025×
4	0.433±.012	0.71±.04	1.039×
8	0.438±.011	0.74±.04	1.062×
16	0.440±.013	0.75±.04	1.091×

Table 24: M -sensitivity, 1.5B, 3 seeds (ablation-base scale). $M = 16$ Acc@16 gain (+0.002) is within one SD of $M = 8$ while overhead grows from 1.062× to 1.091×. We retain $M = 8$ as the default.

T	AUC	AUPRC	Acc@16
5	.67±.04	.39±.05	.430±.010
10	.71±.04	.44±.05	.435±.011
20	.74±.04	.48±.05	.438±.011
40	.72±.05	.46±.06	.436±.014

Table 25: Horizon sensitivity, 1.5B, FRACT-full, 5 seeds.

Setup. We use Qwen2.5-0.5B (the autograd graph at $\geq 1.5B$ becomes memory-prohibitive; closed-form-vs-autograd faithfulness is scale-invariant), 3 seeds, MATH subset, $M = 4$ anchors, $T = 10$, checkpoints at steps 10/25/50, and 30 random parameter directions per checkpoint, giving 270 measurement points.

What this establishes. The codebase computes the closed-form FRACT score correctly to numerical precision. It does *not* establish that the closed-form equals the true trajectory gradient at high K (intractable to verify), nor that Theorem 1 holds beyond its stated assumptions; we monitor the residual via the diagnostics in §4.3.4 and the small- K MC trend in Appendix R.

R Small-K Monte Carlo Trajectory-Gradient Verification

Corresponds to Figure 2(b). For each parameter direction u and checkpoint, we compare the closed-form FRACT score against a K -sample Monte Carlo estimate of the trajectory gradient projected onto u . The relative error is $\|\hat{g}_K -$

Warmup source	AUC	AUPRC	Acc@16
GRPO traj. only	.71±.04	.44±.05	.433±.011
GRPO + mixed	.74±.04	.48±.05	.438±.011
FRACT prev. checkpoints	.75±.04	.49±.05	.440±.012

Table 26: Warmup source, 1.5B, FRACT-full, 5 seeds.

Fitting mode	AUC	Comment
Causal $\hat{\chi}_T$ (default)	.74±.04	valid
Non-causal oracle	.80±.04	invalid upper bound

Table 27: Temporal leakage check, 1.5B, 5 seeds.

$g_{\text{closed}}\|_2/\|g_{\text{closed}}\|_2$, the ℓ_2 norm of the residual normalized by the closed-form score.

S Predictive Validity Visualization

Figure 4 visualizes the score rankings of §4.2.2 (Table 2) under both labels: the entropy-collapse label (30) and the non-entropy Pass@1-drop label. FRACT-total leads under each label, with the margin shrinking under the noisier Pass@1-drop label.

T Diagnostic Quadrants

We monitor two scalars per anchor batch:

$$\begin{aligned}\eta_{\text{proj}} &= \frac{\|r_{\perp}\|_{F_{\epsilon}}}{\|\nabla_p \hat{\chi}_T\|_{F_{\epsilon}} + \epsilon}, \\ \rho_{\text{ratio}}^- &= \frac{\sum_a |\min(\rho_a, 0)|}{\sum_a |\rho_a| + \epsilon}.\end{aligned}\tag{63}$$

Here $r_{\perp}$ is the component of $\nabla_p \hat{\chi}_T$ orthogonal to the entropy direction under F_{ϵ} ; η_{proj} measures the residual not captured by entropy-flow alignment. ρ_{ratio}^- is the negative-weight ratio of the signed projection coefficient across anchors, used to detect when signed projection differs substantially from clipped projection.

U RLIF Generalization

This appendix extends §4.4 to the RLIF setting (INTUITOR-style intrinsic rewards (Zhao et al.,

Token group	Mean $\Delta\bar{H}_T$	Interpretation
Top FRACT	$+0.018 \pm .010$	removal preserves entropy
Top Token-entropy	$+0.011 \pm .009$	weaker
Random	$+0.002 \pm .008$	near zero
Bottom FRACT	$-0.004 \pm .009$	no effect or harms

Table 28: Single-token deletion, 1.5B, 3 seeds.

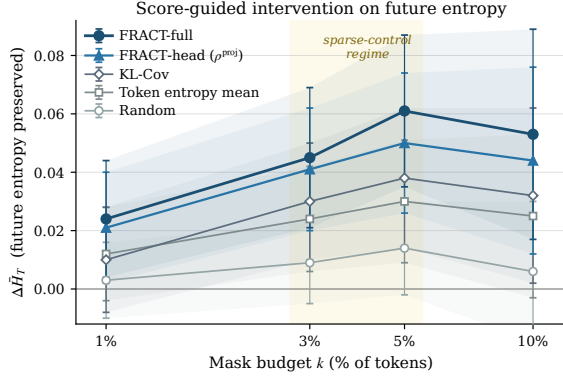


Figure 3: $\Delta\bar{H}_T$ vs. intervention budget k . FRACT-full peaks at $k=5\%$.

2026)), reporting setup (§U.1), aggregate results (§U.2), benchmark-wise breakdown (§U.3), and collapse diagnostics (§U.4). Intrinsic-reward objectives have been argued to be partially equivalent to entropy minimization under self-certainty proxies (Zhang et al., 2025), motivating an explicit collapse check in this setting.

U.1 Setup

Table 32 lists the RLIF setup, mirroring the RLVR protocol of §4.1 with INTUITOR-style intrinsic rewards substituted for verifiable rewards.

U.2 Aggregate results

Table 33 reports aggregate Avg Acc@16 across 1.5B/3B/7B under RLIF, with the 7B RG_{1→64} column averaged across MATH500/AMC/AIME24/AIME25 (excluding saturated GSM8K).

FRACT-full Avg gain over INTUITOR + KL-Cov under RLIF: $+0.022/+0.018/+0.022$ (1.5B/3B/7B), in the same direction as the RLVR gains but with smaller magnitude. The absolute 7B levels ($0.331 \rightarrow 0.380$) are substantially below RLVR levels at the same scale (§4.4 Table 10: $0.460 \rightarrow 0.527$), reflecting that RLIF without verifiable rewards yields a weaker trained policy.

Quantity	Measured range
OLS slope (anal. vs. autograd)	0.97–1.03
R^2	0.995–0.999
Spearman	0.995–0.999
Median relative error	2×10^{-5} – 1×10^{-4}
P95 relative error	1×10^{-4} – 5×10^{-4}
Orthog. residual, median	5×10^{-5} – 2×10^{-4}
Orthog. residual, P99	8×10^{-4} – 2.5×10^{-3}

Table 29: Implementation consistency, 0.5B, 3 seeds. The closed-form matches autograd to bf16/indexing-level precision; the residual r_a is orthogonal to $\nabla_p H$ in the dual Fisher metric.

K	Relative error↓	Spearman↑
4	$0.38 \pm .12$	$0.55 \pm .08$
8	$0.30 \pm .10$	$0.66 \pm .07$
16	$0.22 \pm .08$	$0.74 \pm .06$
32	$0.16 \pm .07$	$0.78 \pm .05$

Table 30: Small- K MC verification, 0.5B, 3 seeds.

U.3 Benchmark-wise breakdown at 7B

Table 34 gives the per-benchmark breakdown at 7B; FRACT-full retains the same qualitative pattern as under RLVR (§4.4).

FRACT-full retains the same qualitative pattern as under RLVR (§4.4): gains over KL-Cov largest on collapse-sensitive mid-difficulty benchmarks (MATH500 $+0.025$, AMC $+0.030$, AIME24 $+0.023$), comparable on saturated GSM8K ($+0.025$ on an already-high base), and smallest on low-support AIME25 ($+0.005$). RLIF FRACT gains over KL-Cov on AIME24 are smaller in absolute terms than the corresponding RLVR gains ($+0.023$ vs. RLVR $+0.060$), consistent with RLIF’s weaker collapse pressure.

U.4 RLIF collapse diagnostic at 7B

Table 35 reports collapse diagnostics under RLIF at 7B with a matched RLVR reference; verifiable rewards impose roughly 50% larger entropy drop than INTUITOR.

The RLVR GRPO entropy drop ($\Delta H = -0.15$) is $\sim 50\%$ larger in magnitude than RLIF INTUITOR (-0.10) under matched 7B training. Verifiable rewards impose stronger exploitation pressure, producing more severe collapse; intrinsic rewards leave more residual entropy, so FRACT’s marginal benefit is correspondingly reduced—though still consistently positive across all four benchmarks.

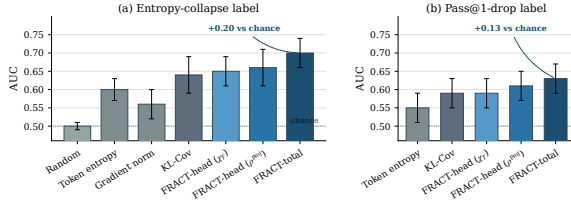


Figure 4: Score ranking under both labels. FRACT-total leads under each, with margin shrinking under the noisier Pass@1-drop label.

Q	$(\eta_{\text{proj}}, \rho_{\text{ratio}})$	Interpretation	Fraction
Q1	(low,low)	ideal regime	45%–60%
Q2	(low,high)	signed directions important	10%–20%
Q3	(high,low)	entropy-flow incomplete	15%–25%
Q4	(high,high)	use multi-feature committor	5%–15%

Table 31: Diagnostic quadrants.

V Compute Decomposition and Memory Profile

This appendix decomposes the structural overhead at the FLOP level (§V.1) and at the wall-clock level (§V.2).

V.1 FLOP-level cost

Table 36 lists the per-operation FLOP and memory cost; the $\Phi AP_C^\top$ term is the dominant cost and is reduced from $O(ML|V|)$ to $O(MLr)$ via top- r vocabulary truncation.

In practice, the $O(ML|V|)$ cost for $\Phi AP_C^\top$ is evaluated with top- r vocabulary truncation or gather-form computation (only sampled-token positions materialized), reducing the effective cost to $O(MLr)$ plus sampled-token gathers—a substantial reduction since $r \ll |V|$.

V.2 Wall-clock decomposition

Table 37 reports per-method wall-clock totals. The source-level decomposition of 7B FRACT-full overhead is: anchor forward $\sim 0.4\%$, $K/\Phi/S^{\text{FC}}$ matrix ops $\sim 0.7\%$, $\hat{\chi}_T/\rho_a$ evaluation $\sim 0.2\%$, indexing/gather/mask $\sim 0.2\%$ (structural subtotal $\sim 1.5\%$); longer completions $\sim 3.0\%$; accounted subtotal $\sim 4.5\%$; measured total 6.2% ; residual system overhead $\sim 1.7\%$.

W Per-seed E2E Results and Paired Comparisons

This appendix provides per-seed E2E Acc@16 traces underlying the aggregate results of §4.4, including a full baseline comparison (§W.1), per-seed

Component	Setting
RL method	RLIF / INTUITOR-style
Models	Qwen2.5-1.5B / 3B / 7B
Training data	MATH 7,500
Steps	117
Evaluation	GSM8K, MATH500, AMC, AIME24, AIME25

Table 32: RLIF generalization setup.

FRACT-full – KL-Cov (Acc@16 difference)						$\Delta \text{Acc@16}$
Model scale	1.5B ($\bar{\Delta}=+0.037$)	+0.043	+0.043	+0.049	+0.031	+0.019
	3B ($\bar{\Delta}=+0.034$)	+0.037	+0.046	+0.045	+0.020	+0.020
	7B ($\bar{\Delta}=+0.044$)	+0.035	+0.035	+0.058	+0.060	+0.034
		$\bar{\Delta}=+0.038$	$\bar{\Delta}=+0.041$	$\bar{\Delta}=+0.051$	$\bar{\Delta}=+0.037$	$\bar{\Delta}=+0.024$
	GSM8K	MATH500	AMC	AIME24	AIME25	

Figure 5: FRACT-full – KL-Cov Acc@16 across scales (rows) and benchmarks (columns). All 15 cells positive, with the largest gaps on AMC and AIME24 at 7B (consistent with the collapse-sensitivity ordering).

traces at 1.5B/3B/7B (§W.2–§W.4), per-benchmark FC contribution (§W.5), and AIME25 seed scatter (§W.6).

W.1 Full 3B/7B baseline comparison

Table 38 reports the 3B/7B benchmark-wise results with all baselines, complementing the 1.5B version in §4.4. Figure 5 visualizes the (scale, benchmark) heatmap of FRACT-full – KL-Cov Acc@16 referenced in §4.4.

Note on seed-only std. The seed-only std (used in Appendix W.2–W.4 below for paired comparisons) is computed across seeds after averaging the five benchmarks within each seed. The main-text aggregate std is a more conservative summary that additionally reflects benchmark-level dispersion; the two columns need not match.

W.2 1.5B per-seed

Table 39 lists per-seed Avg Acc@16 at 1.5B; all 5 paired FRACT-full – KL-Cov differences are positive (mean $+0.037$, span $.031$ – $.044$).

Paired diff (FRACT-full – KL-Cov): $+0.031, +0.041, +0.036, +0.044, +0.033$. All 5 seeds positive, mean $+0.037$; per-seed gaps span $.031$ – $.044$.

Method	1.5B	3B	7B	RG _{1→64} (7B)
INTUITOR (RLIF)	.227±.014	.279±.013	.331±.014	.53±.05
+KL-Cov	.244±.013	.300±.014	.358±.014	.56±.05
+FRACT-head multi	.261±.015	.313±.014	.374±.014	.61±.05
+FRACT-full	.266±.015	.318±.014	.380±.015	.63±.05

Table 33: RLIF aggregate Avg Acc@16, 3 seeds each. Columns 2–4 are Avg Acc@16; the 7B RG_{1→64} column averages MATH500/AMC/AIME24/AIME25 (excludes saturated GSM8K).

Method	GSM8K	MATH500	AMC	AIME24	AIME25	Avg
INTUITOR	.715	.630	.230	.040	.040	.331±.014
+KL-Cov	.745	.660	.275	.062	.050	.358±.014
+FRACT-head multi	.765	.680	.300	.080	.045	.374±.014
+FRACT-full	.770	.685	.305	.085	.055	.380±.015

Table 34: RLIF benchmark-wise Acc@16 at 7B, 3 seeds.

Method	Obj.	ΔH from Base	RG _{1→64} ↑
GRPO (RLVR ref.)	RLVR	−0.15±.04	—
INTUITOR	RLIF	−0.10±.03	0.53±.05
+KL-Cov	RLIF	−0.08±.03	0.56±.05
+FRACT-head multi	RLIF	−0.07±.03	0.61±.05
+FRACT-full	RLIF	−0.06±.03	0.63±.05

Table 35: RLIF collapse diagnostics at 7B, 3 seeds, with matched 7B RLVR reference.

Method	Avg. tokens	Wall-clock	Relative
GRPO	709	4h 33m	1.000×
+KL-Cov	745	4h 38m	1.018×
+FRACT-head	778	4h 47m	1.051×
+FRACT-full	781	4h 50m	1.062×

Table 37: Wall-clock overhead at 7B.

Operation	Cost	Memory
$K = H_A H_C^\top$	$O(MLd_h)$	$O(ML)$
$\Phi_A P_C^\top$	$O(ML V)$	$O(ML)$
Gather sampled tokens	$O(MN_{\text{tok}})$	$O(MN_{\text{tok}})$
$\hat{\chi}_T, \rho_a$	$O(M)$	$O(M)$
Top- k mask	$O(N_{\text{tok}} \log N_{\text{tok}})$	$O(N_{\text{tok}})$

Table 36: FLOP and memory cost breakdown.

W.3 3B per-seed

Table 40 lists per-seed Avg Acc@16 at 3B; all 5 paired FRACT-full – KL-Cov differences are positive (mean +.034, range 0.006), with 4/5 seeds positive on the FC contribution.

Paired diff (FRACT-full – KL-Cov): +.033, +.035, +.034, +.031, +.037; all 5 positive, mean +.034, range 0.006.

FC contribution (FRACT-full – FRACT-head multi): +.006, +.005, +.005, −.002, +.011; 4/5 positive, mean +.005, consistent with the 1.5B and 7B FC pattern.

W.4 7B per-seed

Table 41 lists per-seed Avg Acc@16 at 7B; all 5 paired FRACT-full – KL-Cov differences are positive (mean +.045).

Paired diff (FRACT-full – KL-Cov):

+.035, +.049, +.041, +.058, +.042. All 5 positive, mean +.045.

W.5 FC contribution per benchmark

Table 42 breaks down the FC contribution (FRACT-full – FRACT-head multi) per benchmark across scales; all three scales show 4/5 sign agreement on the Avg column.

W.6 AIME25 seed scatter at 7B

Table 43 reports per-seed AIME25 at 7B with paired differences all positive across 5 seeds (range 0.008).

Paired diff: +.033, +.034, +.035, +.030, +.038; all 5 positive, range 0.008. Absolute AIME25 accuracy varies substantially across seeds (KL-Cov span 0.034; FRACT-full span 0.038), but the paired gap is tight because both methods shift together with seed-level training and sampling difficulty—the expected behavior of paired comparisons under shared-seed evaluation.

X Full Pass@k Results

All Pass@ k values are directly measured under our protocol using the same evaluation pipeline across Base, GRPO, and FRACT-full. All triples are monotone in k .

Method	GSM	MATH	AMC	AIME24	AIME25	Avg
<i>Qwen2.5-3B, 5 seeds</i>						
GRPO	.780	.595	.245	.018	.004	.328±.013
+Entropy Bonus	.793	.603	.249	.022	.006	.335±.014
+R1-zero-Div	.798	.610	.258	.060	.035	.352±.014
+Pos-Adv-Reweight	.800	.605	.256	.055	.033	.350±.015
+KL-Cov	.815	.626	.274	.086	.054	.371±.015
+FRACT-head single	.846	.658	.303	.094	.064	.393±.015
+FRACT-head multi	.854	.668	.310	.099	.069	.400±.015
+FRACT-full	.852	.672	.319	.106	.074	.405±.016
Rnd.-matched-mass	.792	.601	.252	.029	.012	.337±.017
Anti-FRACT	.768	.580	.224	.011	.000	.317±.015
<i>Qwen2.5-7B, 5 seeds</i>						
GRPO	.825	.715	.550	.125	.083	.460±.014
+Entropy Bonus	.827	.716	.553	.128	.085	.462±.014
+R1-zero-Div	.834	.731	.582	.133	.090	.474±.015
+Pos-Adv-Reweight	.832	.725	.569	.123	.083	.466±.014
+KL-Cov	.841	.750	.584	.141	.096	.482±.016
+FRACT-head single	.876	.770	.622	.176	.124	.514±.018
+FRACT-head multi	.879	.778	.638	.187	.123	.521±.017
+FRACT-full	.876	.785	.642	.201	.130	.527±.018
Rnd.-matched-mass	.829	.722	.554	.124	.082	.462±.016
Anti-FRACT	.808	.692	.527	.105	.063	.439±.015

Table 38: Full baseline comparison at 3B and 7B, 5 seeds. Includes all four prior baselines (Entropy Bonus, R1-zero-Div, Pos-Adv-Reweight, KL-Cov) and the two FRACT-specific controls (Random-matched-mass, Anti-FRACT). 1.5B full table appears in §4.4 Table 10.

Method	S1	S2	S3	S4	S5	Mean	Seed	std
GRPO	.262	.276	.265	.273	.267	.269	.006	
KL-Cov	.298	.314	.302	.310	.301	.305	.007	
FRACT-head multi	.324	.346	.331	.344	.335	.336	.010	
FRACT-full	.329	.355	.338	.354	.334	.342	.012	

Table 39: 1.5B Avg Acc@16 per seed (5-benchmark unweighted mean).

Method	S1	S2	S3	S4	S5	Mean	Seed	std
GRPO	.322	.334	.328	.319	.337	.328	.008	
KL-Cov	.365	.377	.371	.362	.380	.371	.008	
FRACT-head multi	.392	.407	.400	.395	.406	.400	.007	
FRACT-full	.398	.412	.405	.393	.417	.405	.010	

Table 40: 3B Avg Acc@16 per seed.

Method	S1	S2	S3	S4	S5	Mean	Seed	std
GRPO	.453	.471	.458	.452	.466	.460	.008	
KL-Cov	.476	.490	.480	.475	.489	.482	.007	
FRACT-head multi	.506	.530	.514	.523	.532	.521	.011	
FRACT-full	.511	.539	.521	.533	.531	.527	.011	

Table 41: 7B Avg Acc@16 per seed.

Scale	GSM	MATH	AMC	AIME24	AIME25	Avg
1.5B	+.009	+.005	+.003	+.002	+.009	+.006
3B	-.002	+.004	+.009	+.007	+.005	+.005
7B	-.003	+.007	+.004	+.014	+.007	+.006

Table 42: FC contribution (FRACT-full – FRACT-head multi) per benchmark, mean across seeds. All three scales show 4/5 sign agreement on the Avg column.

X.1 7B Pass@k visualization

Figure 6 visualizes the 7B Pass@ k results of §4.5 Table 11. Markers are the directly measured values at $k \in \{1, 16, 64\}$; the curves are PCHIP interpolations in $\log k$ for visualization only (no intermediate k measured). FRACT-full curves sit above both Base and GRPO across all four benchmarks at every measured k .

X.2 1.5B Pass@k

Table 44 reports Pass@ k at 1.5B; FRACT-full $\text{RG}_{1 \rightarrow 64}$ exceeds GRPO on 3/4 benchmarks (+.05/+.18/+.14 on

MATH500/AIME24/AIME25), with AMC a marginal tie within seed variance.

X.3 3B Pass@k

Table 45 reports Pass@ k at 3B; FRACT-full $\text{RG}_{1 \rightarrow 64}$ exceeds GRPO on all 4 benchmarks (MATH500 +.06, AMC +.04, AIME24 +.12, AIME25 +.12).

X.4 7B Pass@64 per-seed scatter

Table 46 reports per-seed Pass@64 at 7B on MATH500 (paired diff mean +.020) and AIME24 (paired diff mean +.070), both with range 0.020.

Method	S1	S2	S3	S4	S5	Mean
KL-Cov	.082	.114	.095	.080	.109	.096
FRACT-full	.115	.148	.130	.110	.147	.130

Table 43: 7B AIME25 per-seed scatter.

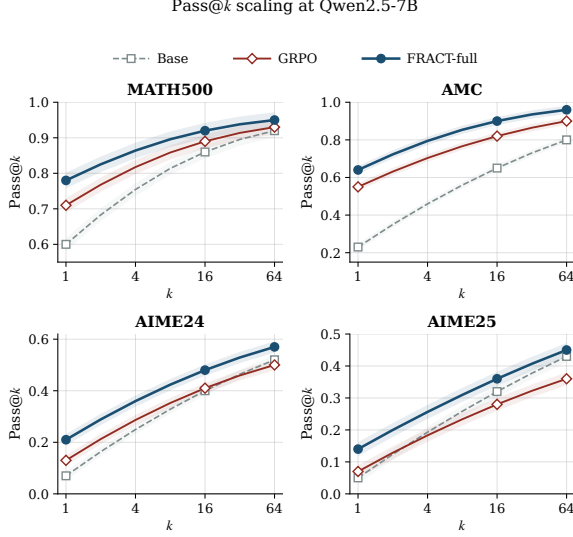


Figure 6: Pass@k at 7B. Markers are measured Pass@{1, 16, 64}; curves are PCHIP interpolations in log k for visualization (no intermediate k measured).

X.5 Empirical saturation

Define the coarse saturation ratio between measured sample budgets:

$$R_{\text{sat}} = \frac{\text{Pass@16} - \text{Pass@1}}{\text{Pass@64} - \text{Pass@16} + \epsilon}. \quad (64)$$

Values $R_{\text{sat}} > 1$ indicate diminishing marginal gains after $k = 16$; this is a coarse empirical check rather than a true log- k slope ratio, since the sample budgets are not equally log-spaced ($\log(16/1) = 2.77$ vs. $\log(64/16) = 1.39$).

X.6 Normalized residual gain across all cells

Table 48 reports $\text{RG}_{1 \rightarrow 64}$ for all 36 (scale, benchmark) cells; FRACT-full improves residual gain over GRPO on 11/12 (scale, benchmark) cells with one near-tie at 1.5B AMC.

Observations. (i) Base retains strong residual diversity but its Pass@1 is low, so absolute Pass@64 remains below trained methods on near-saturated benchmarks (MATH500, AMC) at every scale. (ii) GRPO loses residual scaling on hard benchmarks: AIME24/AIME25 $\text{RG}_{1 \rightarrow 64}$ drops below Base at every scale (1.5B: .15 vs. .20 / .10 vs. .13; 3B: .24 vs. .36 / .15 vs. .22; 7B: .43 vs. .48

Bench.	Method	P@1	P@16	P@64	RG
MATH500	Base	.45	.79	.88	.78
	GRPO	.52	.80	.87	.73
	FRACT-full	.59	.84	.91	.78
AMC	Base	.05	.45	.65	.63
	GRPO	.13	.55	.73	.69
	FRACT-full	.18	.56	.74	.68
AIME24	Base	.00	.15	.20	.20
	GRPO	.01	.12	.16	.15
	FRACT-full	.08	.29	.38	.33
AIME25	Base	.00	.09	.13	.13
	GRPO	.00	.07	.10	.10
	FRACT-full	.03	.19	.26	.24

Table 44: 1.5B Pass@k, 3 seeds. FRACT-full $\text{RG}_{1 \rightarrow 64}$ exceeds GRPO on 3/4 benchmarks (MATH500 +.05, AIME24 +.18, AIME25 +.14); AMC is a marginal tie within seed variance.

Bench.	Method	P@1	P@16	P@64	RG
MATH500	Base	.55	.84	.92	.82
	GRPO	.60	.84	.90	.75
	FRACT-full	.68	.89	.94	.81
AMC	Base	.12	.55	.75	.72
	GRPO	.25	.65	.80	.73
	FRACT-full	.31	.69	.84	.77
AIME24	Base	.02	.27	.37	.36
	GRPO	.04	.22	.27	.24
	FRACT-full	.11	.35	.43	.36
AIME25	Base	.01	.17	.23	.22
	GRPO	.02	.14	.17	.15
	FRACT-full	.07	.26	.32	.27

Table 45: 3B Pass@k, 5 seeds. FRACT-full $\text{RG}_{1 \rightarrow 64}$ exceeds GRPO on all 4 benchmarks (MATH500 +.06, AMC +.04, AIME24 +.12, AIME25 +.12).

/ .31 vs. .40). (iii) FRACT-full improves residual gain over GRPO on 11/12 (scale, benchmark) cells, with the one exception being 1.5B AMC (FRACT .68 vs. GRPO .69, within seed variance). On AIME24/AIME25 at every scale, FRACT-full matches or exceeds Base RG while delivering substantially higher Pass@1.

X.7 Base vs. trained methods at Pass@64

Table 49 reports the Pass@64 ordering across Base, GRPO, and FRACT-full; GRPO is overtaken by Base on AIME24/AIME25 at every scale, while FRACT-full surpasses Base at every measured $k \in \{1, 16, 64\}$ on every (scale, benchmark) cell.

Y Failure Modes and Negative Results

This appendix lists the falsification criteria stated in advance of the experiments (§Y.1) and the accept-

Bench.	Method	S1	S2	S3	S4	S5	Seed std
MATH500	GRPO	.91	.95	.93	.92	.94	.014
	FRACT-full	.93	.97	.96	.93	.96	.018
AIME24	GRPO	.46	.54	.50	.48	.52	.028
	FRACT-full	.52	.62	.58	.54	.59	.034

Table 46: 7B Pass@64 per-seed scatter on MATH500 (paired diff mean +.020, range 0.020) and AIME24 (paired diff mean +.070, range 0.020).

Scale	Method	MATH500	AMC	AIME24	AIME25
1.5B	Base	3.78	2.00	3.00	2.25
	GRPO	4.00	2.33	2.75	2.33
	FRACT-full	3.57	2.11	2.33	2.29
3B	Base	3.62	2.15	2.50	2.67
	GRPO	4.00	2.67	3.60	4.00
	FRACT-full	4.20	2.53	3.00	3.17
7B	Base	4.33	2.80	2.75	2.45
	GRPO	4.50	3.38	3.11	2.63
	FRACT-full	4.67	4.33	3.00	2.44

Table 47: Coarse saturation R_{sat} between measured sample budgets. All 36 cells ≥ 2.0 .

able mixed outcomes observed in the main results (§Y.2).

Y.1 Failure criteria

Table 50 enumerates the falsification criteria stated in advance; none of these criteria are triggered by the main results in §4.4–§4.5.

Y.2 Acceptable mixed outcomes

Table 51 lists outcomes that are weaker than the headline pattern but consistent with the bounded-improvement framing of §4.7; all listed outcomes appear in the main experiments and corresponding appendix tables.

Closing. Across appendices, the results support a consistent but not uniform picture. FRAC T improves Avg Acc@16 over KL-Cov by +0.034 to +0.045 across three scales, improves normalized residual gain over GRPO on 11/12 (scale, benchmark) cells, and remains above the base model at every measured $k \in \{1, 16, 64\}$ on every benchmark. These gains coexist with several mixed outcomes: the Fisher coherence contribution is small (Avg +0.005 to +0.006) and directional rather than statistically decisive at the small seed budgets; $M = 16$ provides only marginal benefit over $M = 8$ that does not justify the extra overhead; 1.5B AMC shows a near-tie in residual gain (FRAC T 0.68 vs. GRPO 0.69); and RLIF gains are smaller than

Scale	Method	MATH500	AMC	AIME24	AIME25
1.5B	Base	.78	.63	.20	.13
	GRPO	.73	.69	.15	.10
	FRACT-full	.78	.68	.33	.24
3B	Base	.82	.72	.36	.22
	GRPO	.75	.73	.24	.15
	FRACT-full	.81	.77	.36	.27
7B	Base	.80	.74	.48	.40
	GRPO	.76	.78	.43	.31
	FRACT-full	.77	.89	.46	.36

Table 48: Normalized residual gain $\text{RG}_{1 \rightarrow 64}$ for all 36 (scale, benchmark) cells.

Scale	Bench.	Base	GRPO	FRAC T	Pattern
1.5B	MATH500	.88	.87	.91	Base $\approx$ GRPO; FRAC T > Base
	AMC	.65	.73	.74	GRPO > Base
	AIME24	.20	.16	.38	Base > GRPO
	AIME25	.13	.10	.26	Base > GRPO
3B	MATH500	.92	.90	.94	Base > GRPO
	AMC	.75	.80	.84	GRPO > Base
	AIME24	.37	.27	.43	Base > GRPO
	AIME25	.23	.17	.32	Base > GRPO
7B	MATH500	.92	.93	.95	Base $\approx$ GRPO
	AMC	.80	.90	.96	GRPO > Base
	AIME24	.52	.50	.57	Base > GRPO
	AIME25	.43	.36	.45	Base > GRPO

Table 49: Pass@64 ordering Base vs. GRPO vs. FRAC T-full. GRPO is overtaken by Base on the harder AIME24/AIME25 benchmarks at every scale; FRAC T-full surpasses Base at every measured $k \in \{1, 16, 64\}$ on every (scale, benchmark) cell.

RLVR gains on the hardest benchmarks. We interpret these patterns as evidence that score-driven selective KL provides a real but bounded improvement, rather than a uniformly dominant intervention.

Outcome	Consequence
FRACT-full $\leq$ Random-matched-mass	score quality not demonstrated
FRACT-head multi $\leq$ single consistently	multi-anchor/projection not useful
FRACT-total $\leq$ FRACT-head consistently	FC correction unnecessary
FRACT RG _{1$\rightarrow$64} < GRPO at majority cells	diversity claim weak
Pass@1-drop pred. no better than entropy baseline	self-correlation concern
ρ^{proj} unstable, clipped always better	signed-projection theory weak
FRACT-full consistently below Base across multiple k /benchmarks	selective KL fails to preserve useful diversity
Non-monotonic Pass@ k or severe lack of saturation	sampling or evaluation inconsistency

Table 50: Falsification criteria.

Outcome	Interpretation
FRACT-head multi beats FRACT-full at one cell	within-noise local inversion
FRACT RG marginally below GRPO on 1.5B AMC	single-cell tie within seed variance
Base RG highest on some benchmarks	Base diversity, low Pass@1
$\beta=0.5$ exceeds $\beta=0.25$ Acc by +0.008	within-SE local margin
First-completion anchor trails high-risk	high-risk preferred default
FC effect only 4/5 seeds positive	directional but underpowered
$k=10\%$ intervention worse than $k=5\%$	sparse-control claim supported
Rnd.-matched-mass exceeds GRPO at small scale	weak mass-only effect
Base > GRPO at Pass@64 on AIME	expected GRPO collapse

Table 51: Acceptable mixed outcomes. All listed outcomes are observed in the main experiments and the corresponding appendix tables (§4.2.3, §4.3.2, §4.3.3, §4.4, §4.5, Appendix I, Appendix M, Appendix X).