The Reasoning–Creativity Trade-off: Toward Creativity-Driven Problem Solving

Max Ruiz Luyten **Mihaela van der Schaar**
University of Cambridge

Abstract

State-of-the-art post-training pipelines for reasoning LLMs rely on bootstrapped reasoning loops: they sample many traces, score them, and reinforce the highest-scoring ones, typically by correctness. This can improve accuracy while still collapsing the distribution *inside* the correct set onto a narrow family of redundant strategies, reducing creative problem-solving. To diagnose this failure mode, we introduce *Distributional Creative Reasoning* (DCR), a variational framework that casts training as gradient flow on the simplex of reasoning traces. The framework yields three core results. First, a diversity-decay analysis shows that STaR-style rejection fine-tuning and exact mean-field GRPO amplify whichever correct trace is already larger, while DPO regresses pairwise correct-trace ratios toward the reference ratios. Second, it explains why entropy and KL can slow or tether collapse but do not reward semantically distinct correct strategies for being distinct, and how a creativity kernel supplies the missing relational term. Third, under mild conditions, the resulting dynamics converge to a unique, stable, and diverse equilibrium, yielding practical guidance for kernel and hyperparameter design. DCR thus offers a principled route to training reasoning LLMs that remain both correct and creative.

1 INTRODUCTION

Reasoning pipelines can improve accuracy while collapsing diversity. Modern reasoning pipelines for large language models (LLMs) typically follow a common template: sample many solution traces, score them, and reinforce the successful ones. This recipe has been highly effective at improving benchmark accuracy, but it comes with a recurring side effect: the model's distribution over traces contracts, concentrating on a small family of templates and near-duplicates. We refer to this as **creative collapse**. It has been observed across RLHF-style post-training, GRPO-based reasoning, and self-training loops that repeatedly amplify successful outputs (Kirk et al., 2023; Shao et al., 2024; Mohammadi, 2024; Havrilla et al., 2024; Murthy et al., 2025; Yue et al., 2025).

Accuracy is not the whole story. A policy over traces has more structure than its top-1 answer or its total probability of correctness. Two models can have the same accuracy and still differ significantly: one may spread that mass across several genuinely different valid strategies, while another may collapse onto a single successful template and its paraphrases. Standard accuracy-like metrics fail to represent this.

This paper studies that hidden second axis: not only *how much* probability mass is correct, but also *how the correct mass is distributed*. That distinction matters particularly for (1) robustness to out-of-distribution reasoning and (2) multi-sample settings relevant to open-ended scientific discovery, both of which benefit from keeping multiple distinct valid approaches alive (Lehman and Stanley, 2011; Pugh et al., 2016; Wang et al., 2022; Kirk et al., 2023).

An operational notion of creativity. We use *creativity* in an operational rather than aesthetic sense: a model is creative when it maintains a *diverse portfolio of semantically distinct correct reasoning strategies*. Under this definition, a policy can become more accurate while less creative if it concentrates its correct mass on a narrow, redundant subset of valid solutions. In the theory sections, we refer to these coarse semantic groups as *strategies* or *semantic lumps*; in the experiments, they are represented either by exact

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

strategy labels or by clustering proxies, depending on the benchmark.

Our question. The central question of the paper is:

> *Can we design a training framework that explains why diversity collapse occurs under common scalar objectives, predicts how different algorithms collapse, and provides a principled remedy that preserves a diverse portfolio of correct reasoning strategies?*

Our answer. We answer this question through **Distributional Creative Reasoning** (DCR), a framework that treats training as gradient flow on the probability simplex over complete reasoning traces. The central object is the conditional distribution $p_\theta(\pi \mid x)$, not an individual trace. This change in perspective is essential because collapse is a property of the *distribution*.

DCR balances three ingredients: utility, structured diversity, and KL regularization:

$$J(p) = \mathcal{U}[p] + \lambda \mathcal{D}[p] - \beta_{\mathrm{KL}}\,\mathrm{KL}(p \| p_{\mathrm{base}}).$$

The diversity energy is

$$\mathcal{D}[p] = \alpha H[p] - \beta Q[p].$$

Here $H[p]$ is the Shannon entropy, and $Q[p] = p^\top K p$ is a kernel coverage term built from a semantic similarity kernel K, which penalizes concentration on similar traces. Entropy spreads mass; the kernel distinguishes "many strings that say the same thing" from "many strategies that solve the task in different ways."

What the theory shows. The framework yields three core messages.

First, it provides a unified probability-space lens on post-training dynamics and a simple diagnostic quantity: the pairwise log-ratio of trace probabilities. Correct–incorrect log-ratios measure correctness margins. Correct–correct log-ratios measure balance *inside* the correct set, indicating whether one valid strategy is taking over another.

Second, under the simplified models analyzed here, DCR predicts distinct collapse modes. STaR-style rejection fine-tuning and exact mean-field GRPO both amplify whichever correct trace is already larger; DPO regresses pairwise correct-trace ratios toward reference-relative values (Zelikman et al., 2022; Shao et al., 2024; Rafailov et al., 2023). Entropy and KL can damp or tether these ratios, but they do not create a force that rewards distinct correct strategies for being distinct.

Third, DCR supplies the missing force through a *gated creativity kernel* over verified-correct traces. Under mild conditions, the resulting regularized objective has a unique globally attracting interior equilibrium. With an appropriate kernel, this equilibrium preserves semantically distinct correct strategies rather than collapsing onto a reasoning monoculture.

Empirically, the same pattern appears across the appendix studies: in the held-out fixed-bank same-support MATH-500 study, DCR improves semantic coverage at fixed pass@8; the public-corpus symbolic exact-label study shows larger gains with oracle kernels than with embedding approximations; and Reasoning-Trap serves as a secondary robustness diagnostic with mixed results.

Contributions.

1. **Unified dynamical lens.** We model post-training as Shahshahani gradient flow on the simplex of complete reasoning traces (Shahshahani, 1979; Amari, 1998) and derive a universal pairwise log-ratio identity that exposes how training reshapes relative trace probabilities.
2. **Diversity decay under scalar objectives.** We show that, in the simplified models analyzed here, STaR-style rejection fine-tuning and exact mean-field GRPO amplify the currently larger correct trace, while DPO regresses pairwise correct-trace ratios toward reference-relative values.
3. **A remedy with guarantees.** We show that with positive entropic weight and a gated PSD creativity kernel, DCR admits a unique globally attracting interior equilibrium. With a semantically aligned kernel, the regularizer repels overrepresented correct strategies while entropy homogenizes redundant surface forms within a strategy.
4. **Practical design rules.** We provide concrete guidance for kernel construction and hyperparameter tuning, including conservative sufficient conditions for suppressing incorrect traces while preserving structured diversity.

Roadmap. Section 2 situates the paper within the literature on collapse in post-training and diversity-aware objectives. Section 3 introduces DCR as a distributional variational framework. Section 4 develops the diversity-decay analysis for scalar objectives. Section 5 shows how the DCR diversity energy changes the equilibrium geometry, Section 6 turns those results into practical kernel-design guidance, and Section 7 surfaces the main empirical validations. Section 8 concludes with implications and limitations. Appendix G contains the full experimental appendix, including the

toy suite, additional synthetic diagnostics, held-out same-support MATH-500, the symbolic exact-label study, and ReasoningTrap, while Appendix F maps the same mechanisms to recent empirical findings on RLVR and alignment.

2 RELATED WORK

From reward optimization to reasoning monoculture. A growing empirical literature suggests that post-training can improve small-k accuracy while narrowing the space of solutions a model actually explores. In RLHF and aligned generation more broadly, output and conceptual diversity often drop after alignment (Kirk et al., 2023; Mohammadi, 2024; Murthy et al., 2025). In reasoning-specific settings, recent work finds that RL or RLVR often reweights or amplifies reasoning paths already present in the base or SFT model rather than eliciting fundamentally new ones, and that pass@1 gains can coexist with deterioration in pass@k or other diversity-sensitive metrics (Yue et al., 2025; Havrilla et al., 2024; Dang et al., 2025; Zhao et al., 2025; Song et al., 2025). GRPO-style pipelines such as DeepSeekMath are central examples of this regime (Shao et al., 2024).

First attempts to restore diversity. Most current remedies either preserve undirected stochasticity or try to keep policies close to a reference model. Entropy bonuses, KL tethers, and newer entropy-control methods can delay or regulate entropy collapse, but by themselves they do not distinguish trivial variation from semantically distinct reasoning strategies (Xiao et al., 2025; Cui et al., 2025; Wang et al., 2025). More direct diversity-aware objectives have recently appeared for supervised fine-tuning, preference optimization, and RL—e.g. diversity-preserving SFT, DivPO, diversity-aware policy optimization for reasoning, and DARLING—but these methods remain largely algorithm-specific and do not offer a unified distributional account of why collapse occurs or what geometric ingredient is missing (Li et al., 2024; Lanchantin et al., 2025; Yao et al., 2025; Li et al., 2025). Outside LLM post-training, novelty search and quality-diversity methods argue more broadly for maintaining diverse high-performing behaviors rather than a single optimum (Lehman and Stanley, 2011; Pugh et al., 2016; Mouret and Clune, 2015).

Theoretical lenses. Replicator dynamics, Shahshahani geometry, and information-geometric natural gradients already suggest why pure utility maximization concentrates mass on high-fitness types and why entropy acts as an undirected spreading force (Hofbauer and Sigmund, 1998; Shahshahani, 1979; Amari, 1998; Harper, 2010). These perspectives are highly relevant, but they do not yet characterize the different within-correct dynamics induced by modern reasoning algorithms such as STaR, GRPO, and DPO, nor do they isolate the missing relational ingredient needed to maintain semantically distinct correct strategies.

Where DCR differs. DCR provides (i) a common probability-space dynamical lens for post-training on the simplex of complete reasoning traces; (ii) a precise diagnosis of what scalar objectives do to the *distribution inside the correct set*; and (iii) a gated kernel construction that adds the missing relational force among correct strategies. In this sense, DCR reframes diversity preservation from an ad hoc heuristic into a principled design problem.

3 DISTRIBUTIONAL CREATIVE REASONING

DCR frames LLM training as a dynamical system on the space of probability distributions over complete reasoning traces. This section introduces the framework and the key quantities that govern it.

3.1 The Landscape of Reasoning

For a fixed prompt $x \in \mathcal{X}$, let $\pi = (t_1, \ldots, t_{|\pi|})$ denote a complete trace, such as a chain of thought, a program, or an action sequence, over a finite vocabulary $\mathcal{V}$ and a maximum horizon T. The set of all such traces is finite:

$$\mathcal{S}_T, \qquad S := |\mathcal{S}_T|.$$

A policy over traces is a probability vector

$$p \in \Delta^{S-1} := \Big\{ p \in [0,1]^S : \sum_{i=1}^{S} p_i = 1 \Big\}.$$

For tasks with binary verification, we write

$$\mathcal{S}_T = \mathcal{C} \sqcup \mathcal{I},$$

where $\mathcal{C}$ and $\mathcal{I}$ are the sets of correct and incorrect traces. Two quantities then matter:

$$\rho(p) := \sum_{c \in \mathcal{C}} p_c, \qquad s_c(p) := \frac{p_c}{\rho(p)} \quad (c \in \mathcal{C},\ \rho(p) > 0).$$

The first, $\rho(p)$, is an accuracy-like quantity: total mass on correct traces. The second, $s(\cdot)$, is the distribution *inside the correct set*. Two policies can have the same $\rho(p)$ and yet differ radically in s: one may spread its correct mass across distinct strategies, while another may collapse onto one template and its paraphrases. DCR is designed to study and control both.

3.2 The DCR Objective

DCR optimizes an objective over the simplex:

$$J(p) = \mathcal{U}[p] + \lambda \mathcal{D}[p] - \beta_{\mathrm{KL}}\,\mathrm{KL}(p\|p_{\mathrm{base}}).$$

The components are:

1. **Utility:**
$$\mathcal{U}[p] = \sum_{\pi \in \mathcal{S}_T} U(\pi)\, p(\pi),$$
the expected task utility of traces.
2. **Diversity energy:** $\mathcal{D}[p]$, defined below, encourages breadth and structured distinctiveness.
3. **KL:** discourages uncontrolled drift from a reference policy p_{base}, typically an SFT checkpoint.

The coefficients $\lambda, \beta_{\mathrm{KL}} \geq 0$ control the trade-off.

3.3 The Diversity Energy Functional $\mathcal{D}[p]$

The diversity term is

$$\mathcal{D}[p] = \alpha H[p] - \beta Q[p], \qquad \alpha, \beta \geq 0,$$

where

$$H[p] := -\sum_{\pi} p(\pi) \log p(\pi)$$

is Shannon entropy and

$$Q[p] = p^\top K p = \sum_{\pi,\pi'} k(\pi,\pi')\, p(\pi) p(\pi')$$

is a quadratic form induced by a symmetric similarity kernel K.

The two terms play different roles:

1. **Entropy $H[p]$:** rewards probabilistic breadth. It prevents premature concentration, but it is blind to whether traces are redundant or genuinely different.
2. **Kernel coverage $Q[p]$:** measures how much probability mass is placed on mutually similar traces. The penalty $-\beta Q[p]$ therefore discourages concentration on semantically redundant solutions.

This decomposition is important. Entropy encourages *more* traces; the kernel encourages *distinct* traces.

Proposition 3.1 (Concavity of the diversity energy)**.** *If K is positive semidefinite, then $\mathcal{D}[p]$ is concave on the simplex. It is strictly concave on the affine simplex if $\alpha > 0$, or more generally whenever the quadratic term is strictly negative on tangent directions.*

In practice, we additionally include a small entropy barrier $+\varepsilon H[p]$ with $\varepsilon > 0$. This keeps all coordinates strictly positive along trajectories and ensures a unique interior maximizer under the mild conditions formalized in Corollary A.4.1 and Lemma A.5.

3.4 Learning Dynamics and Pairwise Ratios

DCR evolves the policy by Shahshahani gradient flow, the natural gradient flow on the simplex (Shahshahani, 1979; Amari, 1998). Let

$$A := \lambda\alpha + \beta_{\mathrm{KL}} + \varepsilon > 0.$$

Barrier versus shaping entropy. The small additive term $\varepsilon H[p]$ and the diversity-weighted entropy term $\lambda\alpha H[p]$ play distinct roles even though they enter the dynamics only through their sum in

$$A = \lambda\alpha + \beta_{\mathrm{KL}} + \varepsilon.$$

We use $\varepsilon > 0$ as an interior barrier, typically chosen very small, to keep trajectories away from the boundary and pairwise log-ratios well defined. By contrast, $\lambda\alpha$ is the tunable part that intentionally broadens the equilibrium distribution. In practice, one can therefore regard ε as a stability floor and $\lambda\alpha$ as the coefficient that controls distributional breadth, while β_{KL} tethers the policy to the reference rather than creating diversity by itself.

With this notation, the regularized objective $\widetilde{J}(p) := J(p) + \varepsilon H[p]$ induces the ODE

$$\dot{p}_i = p_i\big(\phi_i(p) - \bar{\phi}(p)\big) - A\, p_i\big(\log p_i - \langle \log p \rangle\big), \quad (1)$$

where

$$\phi_i(p) := U_i - 2\lambda\beta(Kp)_i + \beta_{\mathrm{KL}} \log p_{\mathrm{base},i},$$

$$\bar{\phi}(p) := \sum_j p_j \phi_j(p), \quad \text{and} \quad \langle \log p \rangle := \sum_j p_j \log p_j.$$

The most informative observable is not p_i in isolation but the *pairwise log-ratio*

$$z_{ij}(t) := \log \frac{p_i(t)}{p_j(t)}.$$

Why this quantity? If $i \in \mathcal{C}$ and $j \in \mathcal{I}$, it measures a correctness margin. If $i, j \in \mathcal{C}$, it measures balance *inside* the correct set. Indeed, for correct traces a, b,

$$\frac{s_a}{s_b} = \frac{p_a/\rho}{p_b/\rho} = \frac{p_a}{p_b},$$

so correct–correct log-ratios are exactly the right observable for within-correct diversity.

Theorem A.2 proves the universal identity

$$\frac{d}{dt} \log \frac{p_i(t)}{p_j(t)} = \phi_i(p(t)) - \phi_j(p(t)) - A \log \frac{p_i(t)}{p_j(t)}. \quad (2)$$

This identity underlies the rest of the paper: all common coupling terms disappear in log-ratio coordinates. Pairwise ratios are controlled only by score differences and the entropy/KL barrier.

Theorem 3.1 (Global convergence of DCR training; cf. Theorem A.7)**.** *Assume $A > 0$ and that the kernel K is symmetric positive semidefinite. Then the regularized objective $\widetilde{J}$ has a unique maximizer*

$$p^\star \in \operatorname{int} \Delta^{S-1}.$$

For every interior initialization $p_0 \in \operatorname{int} \Delta^{S-1}$, the Shahshahani gradient flow of $\widetilde{J}$ admits a unique global solution p_t, remains in the interior for all $t \geq 0$, and converges to $p^\star$ as $t \to \infty$. Moreover, $\widetilde{J}(p_t)$ is strictly increasing unless $p_t = p^\star$.

This theorem gives interiority, stability, and uniqueness. Whether the equilibrium is merely broad or genuinely *strategically diverse* depends on the kernel design developed in Sections 5 and 6.

3.5 Parametric Realization and Scalability

In an LLM, the trace distribution is parameterized as p_θ. Theorems D.1, D.2, and D.5 show that the parameterization itself is well behaved: a single softmax block has exact dimension-free Jacobian and Hessian bounds, and repeated autoregressive composition contributes only an explicit linear factor in the horizon T to the logit-to-trace parameterization. Repeated softmax composition alone therefore does not imply severe conditioning.

The kernel term is also compatible with minibatch training. For a batch of sampled traces, the gradient of $Q[p_\theta]$ can be estimated with a minibatch U-statistic over all in-batch pairs at $O(B^2)$ cost (Clémençon et al., 2016). Practical kernel constructions are discussed in Section 6.

4 COLLAPSE UNDER SCALAR OBJECTIVES

This section studies what happens when the training signal contains no explicit relational term among correct traces. The point is not merely that incorrect traces are suppressed. The more interesting question is what happens *inside the correct set*.

4.1 What Quantity Do We Track?

As discussed in the previous section, the probability-space dynamics takes the generic form

$$\dot{p}_i = p_i\big(\psi_i(p) - \bar{\psi}(p)\big) - \varepsilon\, p_i\big(\log p_i - \langle \log p \rangle\big), \quad (3)$$

where $\psi_i(p)$ is the task-dependent score and $\varepsilon \geq 0$ is an entropy-like damping term, and Theorem A.2 shows that for any pair of positive-mass traces,

$$\frac{d}{dt} \log \frac{p_a(t)}{p_b(t)} = \psi_a(p(t)) - \psi_b(p(t)) - \varepsilon \log \frac{p_a(t)}{p_b(t)}. \quad (4)$$

Thus pairwise diversity is governed by a simple competition: task-induced score differences versus entropic damping.

Scope of the analysis. Appendix B analyzes different models chosen to isolate the within-class selection mechanism: STaR-style online rejection fine-tuning, exact mean-field GRPO batch-gradient dynamics in a one-step tabular trace model, and DPO under exchangeable-pair surrogates. The point is not to reproduce every implementation detail, but to expose how training signals reshape pairwise ratios inside the correct set.

4.2 Deterministic Diversity Decay

STaR-style rejection fine-tuning. Under online rejection fine-tuning, accepted traces are replayed in proportion to their current conditional frequency among correct traces. For correct traces $a, b \in \mathcal{C}$, Theorem B.1 proves

$$\frac{d}{dt} \log \frac{p_a(t)}{p_b(t)} = \frac{1 - \rho(p(t))}{\rho(p(t))}\big(p_a(t) - p_b(t)\big) - \varepsilon \log \frac{p_a(t)}{p_b(t)}. \quad (5)$$

Hence, whenever $0 < \rho < 1$ and $p_a > p_b$, the larger correct trace is amplified. In the unregularized case, STaR deterministically fixates on the initially largest positive-mass correct trace.

Exact mean-field GRPO. At first glance, GRPO may seem more symmetric because it uses group-normalized rewards. However, Proposition B.2 derives the *exact Euclidean-logit mean field* of the per-batch score-function gradient, and Theorem B.2 shows that it already contains within-class amplification. For correct traces $a, b \in \mathcal{C}$,

$$\frac{d}{dt} \log \frac{p_a(t)}{p_b(t)} = \alpha_C(\rho(p(t)))\big(p_a(t) - p_b(t)\big) - \varepsilon \log \frac{p_a(t)}{p_b(t)}, \quad (6)$$

where

$$\alpha_C(\rho) = (1-\rho)h_G(\rho) \quad \text{with} \quad h_G(\rho) > 0 \text{ for } \rho \in (0,1).$$

Thus the larger correct trace is again amplified. This is an important correction to the common intuition that GRPO is neutral within the correct set: in exact mean field, it is not.

DPO. DPO behaves differently. Under the one-sided exchangeable-pair surrogate analyzed in Theorem B.3, correct traces are updated through reference-relative margins. For $a, b \in \mathcal{C}$,

$$\frac{d}{dt} z_{ab}(t) = g_t'(\xi_{ab}(t))\Big(z_{ab}(t) - z_{ab}^{\text{base}}\Big) - \varepsilon z_{ab}(t), \quad (7)$$

where

$$z_{ab}(t) := \log \frac{p_a(t)}{p_b(t)}, \qquad z_{ab}^{\text{base}} := \log \frac{p_{\text{base},a}}{p_{\text{base},b}},$$

and $g_t'(\xi_{ab}(t)) < 0$. In the unregularized case, DPO therefore regresses pairwise correct-trace ratios toward their reference-relative values. It is not creating structured diversity among correct traces; it is preserving the reference ratio.

Robustness of the diagnosis. Subsection B.4 also shows that these conclusions persist under several more realistic variants. Replay buffers in STaR add regression toward stored frequency ratios, but not semantic coupling. Centered group-weighted GRPO, local clipping, and KL tethers modify coefficients or add base-ratio tethering, but they preserve the same rank-one within-correct geometry. Two-sided exchangeable DPO blocks preserve the same reference-relative regression picture.

4.3 Stochastic Effects: Fixation and Variance Depletion

Real training is noisy. To isolate the geometry of finite-batch fluctuations, Theorem E.1 studies *moment-matched simplex diffusion surrogates* of the form

$$dp_t = F(p_t)\,dt + \sqrt{\gamma\, q(p_t)}\,\Sigma(p_t)\,dW_t, \qquad (8)$$
$$\Sigma(p)\Sigma(p)^\top = J_p := \operatorname{diag}(p) - pp^\top.$$

These are not optimizer-specific weak-limit theorems. They are stochastic surrogates that preserve the correct tangent covariance geometry.

Two regimes are particularly informative:

1. **Constant modulation $q \equiv 1$:** this recovers classical Wright–Fisher-type fixation pressure on the simplex (Ethier and Kurtz, 1986).
2. **GRPO-style variance depletion:** with

   $$q_G(\rho) = 1 - \rho^G - (1-\rho)^G,$$

   the noise strength is proportional to the probability that a sampled group is mixed. As batches become almost pure-correct or pure-incorrect, the stochastic forcing shuts off.

This stochastic analysis is deliberately separated from the exact mean-field GRPO result above. Exact mean-field GRPO already collapses deterministically. The diffusion surrogates reveal a different effect: how finite-batch noise can either sustain fixation pressure (Wright–Fisher regime) or freeze into a path-dependent limiting portfolio when the variance budget depletes (the GRPO-inspired two-block surrogate).

4.4 Synthesis: The Diversity Decay Theorem

The deterministic and stochastic analyses point to the same conclusion: scalar objectives can improve correctness while leaving the distribution inside the correct set unprotected.

Theorem 4.1 (Diversity Decay Theorem)**.** *Consider the scalar-objective models analyzed in Appendix B, i.e. training signals with no explicit relational term among correct traces.*

(i) ***STaR-style rejection fine-tuning*** *induces winner-take-all amplification among positive-mass correct traces.*

(ii) ***Exact mean-field GRPO*** *induces the same within-correct amplification: the currently larger correct trace is deterministically reinforced.*

(iii) ***DPO****, under the exchangeable-pair models studied here, regresses pairwise correct-trace ratios toward reference-relative values.*

Entropy and KL can damp or tether these ratios, but they do not create a force that rewards semantically distinct correct strategies for being distinct.

Interpretation. The theorem does not say that every method decays in the same way. Rather, it identifies a shared structural limitation: the objectives can rescale pairwise ratios, damp them, or tether them to a reference policy, but they do not contain the missing *relational geometry* among correct traces. That is exactly the gap DCR is designed to fill. Appendix G shows these effects directly in the toy suite and then shows on held-out same-support MATH-500 that, once answer correctness is fixed, the remaining movement is in semantic coverage rather than in pass probability.

5 THE DIVERSITY ENERGY EFFECT ON THE EQUILIBRIUM STRUCTURE

The failure of scalar objectives is now clear: they can move mass toward the correct set, but they do not know whether two correct traces are redundant or genuinely different. DCR changes that geometry.

5.1 From Collapse to a Structured Interior Equilibrium

To ensure that diversity pressure is only applied where it is meaningful, DCR uses a *gated effective kernel*

$$K_{\text{eff}} := RK_{\text{sem}}R, \qquad R_{ii} = \mathbf{1}_{\{i \in \mathcal{C}\}},$$

where K_{sem} is a semantic similarity kernel and R is a verifier gate. The kernel vanishes on incorrect traces.

With such gating, the diversity pressure acts only on verified-correct traces. One could also extend diversity pressure to all traces, but we do not analyze that setting here.

Under the same mild conditions as in Theorem 3.1, the regularized DCR objective has a unique globally attracting interior equilibrium $p^\star$. The question then becomes: what does that equilibrium look like?

5.2 Balancing Correctness and Redundancy at Equilibrium

Theorem C.2 answers this question most transparently in the binary-utility, no-KL regime:

$$U_i = \mathbf{1}_{\{i\in\mathcal{C}\}}, \qquad A := \varepsilon + \lambda\alpha > 0.$$

At the unique equilibrium $p^\star$, for any incorrect trace $i \in \mathcal{I}$ and any correct trace $c \in \mathcal{C}$,

$$\frac{p_i^\star}{p_c^\star} = \exp\left(-\frac{1 - 2\lambda\beta(K_{\text{eff}}p^\star)_c}{A}\right), \tag{9}$$

and for any two correct traces $a, b \in \mathcal{C}$,

$$\log\frac{p_a^\star}{p_b^\star} = \frac{2\lambda\beta}{A}\Big((K_{\text{eff}}p^\star)_b - (K_{\text{eff}}p^\star)_a\Big). \tag{10}$$

These identities make the trade-off explicit. Correctness still matters through the unit utility gap between correct and incorrect traces. But among correct traces, relative mass now depends on semantic redundancy through the kernel response $K_{\text{eff}}p^\star$.

A useful way to read (9) is through the *correctness margin*

$$m_c^\star := 1 - 2\lambda\beta(K_{\text{eff}}p^\star)_c.$$

If $m_c^\star > 0$, then every incorrect trace is exponentially suppressed relative to c. Thus the kernel can diversify correct traces without rewarding incorrect ones, provided it is gated and not so strong that it overwhelms the basic correctness advantage.

5.3 Repel Strategies, Not Paraphrases

The key positive result is not only that DCR has an interior equilibrium, but also *how* that equilibrium allocates mass across semantic strategies.

Theorem C.8 formalizes this through a partition of traces into semantic *lumps*: traces in the same lump represent the same underlying strategy up to surface-form variation; traces in different lumps represent genuinely distinct strategies. Under a block-constant kernel model with within-lump similarity k_{in} and across-lump similarity k_{out}, where $k_{\text{in}} > k_{\text{out}} \geq 0$, two effects emerge:

Across different correct lumps. If a belongs to lump m and b to lump $n \neq m$, then

$$\phi_a(p) - \phi_b(p) = -2\lambda\beta\,(k_{\text{in}} - k_{\text{out}})\big(r_m(p) - r_n(p)\big), \tag{11}$$

where $r_m(p)$ is the total correct mass in lump m. If one correct concept becomes overrepresented, its score is pushed down relative to underrepresented concepts. The kernel therefore creates *repulsion across distinct correct strategies*.

Within the same correct lump. If a, b are two traces in the same correct lump, then the kernel does not distinguish them:

$$\phi_a(p) = \phi_b(p).$$

Their log-ratio therefore follows the pure entropic law

$$\log\frac{p_a(t)}{p_b(t)} = \log\frac{p_a(0)}{p_b(0)}\,e^{-At}. \tag{12}$$

Within a strategy, entropy exponentially equalizes redundant surface forms.

Takeaway. This cleanly separates two roles that scalar objectives conflate:

1. **repel strategies:** the kernel pushes mass away from overrepresented correct concepts toward underrepresented ones;
2. **homogenize paraphrases:** entropy equalizes redundant forms within the same concept.

This is the core geometric mechanism behind DCR.

6 THE CREATIVITY KERNEL

The theory above shows that kernel design is not a minor detail. It determines whether the interior equilibrium is merely broad or genuinely diverse in the sense of maintaining semantically distinct correct strategies.

6.1 Why Entropy Alone Is Not Enough

Entropy is useful, but it is blind. It can preserve many traces while failing to distinguish meaningful strategy diversity from superficial variation. In particular, entropy cannot tell the difference between (i) many traces that all instantiate the same idea, and (ii) many traces that solve the problem in genuinely different ways.

This is why an entropy-only remedy is incomplete. Entropy provides undirected breadth. The kernel provides directional structure.

6.2 Practical Design of the Semantic Kernel

In practice, the semantic kernel $k_{\text{sem}}(\pi, \pi')$ should reflect whether two traces are strategically similar. Two broad families are especially natural:

1. **Embedding-based kernels:** compute an embedding for the full trace or for a structured trace summary, then apply a PSD kernel in embedding space.

2. **Domain-tailored kernels:** in mathematics, coding, or planning, use features that reflect proof structure, dependency graphs, execution states, or other strategy-level invariants rather than surface wording alone.

Concrete kernel recipes. A practical general-purpose baseline is feature-first: extract a fixed representation $f(\pi)$ for each complete trace or for a structured trace summary, and then define

$$k_{\text{sem}}(\pi, \pi') = \kappa\big(f(\pi), f(\pi')\big)$$

with a PSD kernel κ, e.g. shifted cosine or Gaussian RBF on normalized embeddings from a strong frozen encoder (Lee et al., 2024, 2025). For mathematical reasoning, $f(\pi)$ can summarize proof-state trajectories, intermediate lemmas, or dependency structure, so that traces are judged similar when they instantiate the same proof strategy even if their surface realizations differ. This is especially natural in formal domains such as Lean theorem proving, where intermediate proof states are explicit objects and multi-sample success is central (He et al., 2025). For code generation, $f(\pi)$ can encode AST structure, execution traces, or unit-test fingerprints. The resulting kernel then penalizes concentration on near-identical implementations while still distinguishing genuinely different algorithms.

The symbolic study in Appendix G is consistent with this design principle: when exact strategy labels are available, oracle kernels outperform embedding-based kernels most clearly in the saturated-pass regime, which is precisely where strategy-level geometry matters most.

The key desiderata are:

1. **Intra-strategy coherence:** traces that instantiate the same idea should have similar kernel rows.

2. **Inter-strategy discrimination:** genuinely different correct strategies should be distinguishable.

The gate $K_{\text{eff}} = RK_{\text{sem}}R$ can be regarded as part of the kernel design.

6.3 Practical Tuning Rules

Subsection C.3 yields a set of interpretable, conservative tuning rules.

1. **Use PSD kernels.** Safe choices include Gram kernels, shifted cosine kernels on normalized embeddings, and Gaussian RBF kernels. Positive semidefiniteness is what preserves concavity and global convergence.

2. **Gate the kernel.** To inherit the guarantees proved here,
$$K_{\text{eff}} = RK_{\text{sem}}R.$$
Appendix G shows that gating remains the most conservative safety choice in the synthetic studies, while the held-out real-math results do not support a meaningful advantage of gated over ungated at reporting precision.

3. **Impose a conservative correctness-margin.** Since
$$|(K_{\text{eff}}p)_c| \le \|K_{\text{eff}}\|_{1\to\infty} \quad \text{for all } p \in \Delta^{S-1},$$
a sufficient dimension-free condition is
$$\lambda\beta < \frac{1}{2\|K_{\text{eff}}\|_{1\to\infty}}. \tag{13}$$
Defining
$$\eta_K := 1 - 2\lambda\beta\|K_{\text{eff}}\|_{1\to\infty},$$
this implies a uniform positive margin $m_c^\star \ge \eta_K > 0$ and therefore exponential suppression of incorrect traces. If $M := |\mathcal{C}|$ and $N := |\mathcal{I}|$, Corollary C.3 gives the bound
$$\rho_I^\star \le \frac{N}{N + Me^{\eta_K/A}},$$
where $\rho_I^\star$ is the total incorrect mass at equilibrium.

4. **Tune A as breadth versus sharpening.** The effective entropy coefficient
$$A = \varepsilon + \lambda\alpha$$
governs how broadly mass is spread once the kernel is in place in the no-KL regime considered here; more generally, $A = \varepsilon + \lambda\alpha + \beta_{\text{KL}}$. Smaller A sharpens correctness margins and suppresses incorrect traces more aggressively. Larger A spreads mass more broadly and accelerates within-strategy homogenization via (12). In short:

   K_{eff} chooses the geometry of diversity,

   $\lambda\beta$ chooses its strength,

   A chooses its breadth.

Implementation. The kernel term can be integrated into standard SGD pipelines. Given a minibatch of traces, the gradient of the quadratic term can be estimated with a minibatch U-statistic over all in-batch pairs at $O(B^2)$ per update (Clémençon et al., 2016). This is a practical route to incorporating structured diversity into post-training. For shift-invariant embedding kernels, one can additionally use random-feature or other low-rank approximations so that the quadratic term is computed in a fixed-dimensional feature space rather than from all pairwise similarities (Rahimi and Recht, 2007).

7 EMPIRICAL VALIDATION

Our empirical results are designed to test the theory at increasing levels of realism rather than to claim universal end-to-end training gains on a benchmark. Outside the toy suite, these are fixed-bank same-support reweighting studies rather than end-to-end online post-training runs. The code used in the experimental pipeline is available at `https://github.com/maxruizluyten/creative_reasoning_release`. Appendix G.1 records the code, external assets, and reproducibility details.

The toy suite in Appendix G provides the theory-faithful mechanism validation: the universal log-ratio identity, global convergence, and aligned DCR's semantic-coverage gains all appear directly, without increasing incorrect mass. On held-out same-support MATH-500, all methods attain pass@8 = 0.8912 on the full 500-prompt held-out set (Appendix Table 4), while on the 433-prompt multi-solution slice DCR (Gated) raises coverage@8 from 2.4662 (Utility/Entropy) to 2.5253 and raises N_{eff} from 2.7883 to 2.8784 (Appendix Table 5). In this held-out same-support evaluation, the gain is therefore redistribution across correct strategies rather than improved raw answer accuracy.

8 CONCLUDING INSIGHTS

This paper argues that collapse in reasoning post-training is not only about losing entropy in the abstract. It is about losing the distributional structure *inside the correct set* until a model keeps only one successful template and its near-duplicates. Accuracy can keep rising while this happens.

The experiments sharpen that claim: DCR's clearest gains appear when correctness is already saturated and the remaining degree of freedom is how mass is distributed across correct strategies.

Appendix G gives the full empirical picture: the toy suite shows the predicted geometry directly, held-out same-support MATH-500 shows improved semantic coverage at fixed pass@8, and the symbolic exact-label study sharpens the kernel-design story. Appendix F connects the same mechanics to recent empirical observations in RLVR and alignment.

Method	Within-correct law	Distributional effect
STaR	$\dot{z}_{ab} \propto p_a - p_b$	winner-take-all
GRPO	$\dot{z}_{ab} \propto p_a - p_b$	winner-gets-amplified
DPO	$\dot{z}_{ab} \propto -(z_{ab} - z_{ab}^{\text{base}})$	reference-relative regression
DCR	$\dot{z}_{ab} = \phi_a(p) - \phi_b(p) - A z_{ab}$ with kernel term in ϕ	stable interior portfolio under gated PSD K

Table 1: **Key dynamics inside the correct set.** The exact coefficients are given in Section 4 and Appendices B–C. Scalar objectives amplify or tether pairwise correct-trace ratios; DCR adds explicit semantic coupling among verified-correct traces.

The key conceptual distinction is the following. Total correct mass measures whether a model is right. The distribution within the correct set measures whether it still retains a portfolio of valid alternatives. Pairwise log-ratios make that distinction mathematically explicit. Under the scalar objectives studied here, those ratios are amplified, damped, or tethered, but not organized by semantic structure.

DCR supplies the missing relational geometry. The gated creativity kernel does not merely add more randomness; it changes which correct traces compete with which others. With a semantically aligned kernel, it repels overrepresented strategies while entropy homogenizes redundant surface forms within a strategy. Under mild conditions, this yields a unique, stable, interior equilibrium that can remain both accurate and strategically diverse.

Scope and limitations. Our analysis is intentionally structural. It is carried out on finite trace spaces and simplified trace-level models designed to expose within-class selection mechanisms. The stochastic section uses moment-matched simplex diffusions rather than optimizer-specific SDE limit theorems. In practice, success depends on the quality of the verifier and the semantic kernel, which remain important practical limitations.

Acknowledgments

The authors thank their funders and collaborators for helpful feedback throughout the development of this work. Max Ruiz Luyten is funded by AstraZeneca. We also thank the anonymous reviewers and members of the van der Schaar Lab for valuable comments and suggestions. We used ChatGPT and Gemini for editing, polishing, and coding assistance.

References

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island, April 2007. ISBN 978-0-8218-4302-4 978-1-4704-4605-5. doi: 10.1090/mmono/191.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with Exploration: An Entropy Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30377–30385, March 2026. ISSN 2374-3468. doi: 10.1609/aaai.v40i36.40290.

Stephan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. *Journal of Machine Learning Research*, 17(76):1–36, 2016. ISSN 1533-7928.

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. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL https://arxiv.org/abs/2505.22617.

Xingyu Dang, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Assessing Diversity Collapse in Reasoning. In *Scaling Self-Improving Foundation Models without Human Supervision*, February 2025.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics. Wiley, New York, 1986. ISBN 9780470316658. doi: 10.1002/9780470316658. URL https://doi.org/10.1002/9780470316658.

Marc Harper. The replicator equation as an inference dynamic, 2010. URL https://arxiv.org/abs/0911.1763.

Alexander Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching Large Language Models to Reason with Reinforcement Learning. In *AI for Math Workshop @ ICML 2024*, June 2024.

Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25548–25560, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1298. URL https://aclanthology.org/2025.emnlp-main.1298/.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998. ISBN 978-0-521-62570-8. doi: 10.1017/CBO9781139173179.

Doohyuk Jang, Yoonjeon Kim, Chanjae Park, Hyun Ryu, and Eunho Yang. Reasoning model is stubborn: Diagnosing instruction overriding in reasoning models. arXiv preprint arXiv:2505.17225, 2025. URL https://arxiv.org/abs/2505.17225.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *The Twelfth International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PXD3FAVHJT.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization, 2025. URL https://arxiv.org/abs/2501.18101.

J. LaSalle. Some Extensions of Liapunov's Second Method. *IRE Transactions on Circuit Theory*, 7 (4):520–527, December 1960. ISSN 2331-3854. doi: 10.1109/TCT.1960.1086720.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile text embeddings distilled from large language models, 2024. URL https://arxiv.org/abs/2403.20327.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han,

Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. Gemini embedding: Generalizable embeddings from gemini, 2025. URL https://arxiv.org/abs/2503.07891.

Joel Lehman and Kenneth O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, June 2011. ISSN 1063-6560. doi: 10.1162/EVCO_a_00025.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations, 2025. URL https://arxiv.org/abs/2509.02534.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving Diversity in Supervised Fine-Tuning of Large Language Models. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. OpenAI technical report, 2023.

Tongseok Lim and Robert J. McCann. Geometrical Bounds for Variance and Recentered Moments. *Mathematics of Operations Research*, 47(1):286–296, February 2022. ISSN 0364-765X. doi: 10.1287/moor.2021.1125.

P. L. Lions and A. S. Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on Pure and Applied Mathematics*, 37(4):511–537, 1984. doi: https://doi.org/10.1002/cpa.3160370408. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160370408.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, Amit Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shiva Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: Symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017. doi: 10.7717/peerj-cs.103. URL https://peerj.com/articles/cs-103/.

Behnam Mohammadi. Creativity has left the chat: The price of debiasing language models, 2024. URL https://arxiv.org/abs/2406.05587.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015. URL https://arxiv.org/abs/1504.04909.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11241–11258, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.561. URL https://aclanthology.org/2025.naacl-long.561/.

Mitio Nagumo. Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen. *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, 24:551–559, 1942. doi: 10.11429/ppmsj1919.24.0_551. URL https://www.jstage.jst.go.jp/article/ppmsj1919/24/0/24_0_551/_pdf.

Tianwei Ni, Allen Nie, Sapana Chaudhary, Yao Liu, Huzefa Rangwala, and Rasool Fakoor. Offline learning and forgetting for reasoning with large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=RF6raEUATc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, October 2022.

Gianluigi Pillonetto, Tianshi Chen, Alessandro Chiuso, Giuseppe De Nicolao, and Lennart Ljung. Regularization in Reproducing Kernel Hilbert Spaces. In Gianluigi Pillonetto, Tianshi Chen, Alessandro Chiuso, Giuseppe De Nicolao, and Lennart Ljung, editors, *Regularized System Identification: Learning Dynamic Models from Data*, pages 181–246. Springer International Publishing, Cham, 2022. ISBN 978-3-030-95860-2. doi: 10.1007/978-3-030-95860-2_6.

Justin K. Pugh, Lisa B. Soros, and Kenneth O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, Volume 3 - 2016, 2016. ISSN 2296-9144. doi: 10.3389/frobt.2016.00040.

Qwen Team. Qwen2.5-math technical report. Hugging Face model card, 2024. URL https://huggingface.co/Qwen/Qwen2.5-Math-1.5B.

Qwen Team. Qwen3-embedding-0.6b. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen3-Embedding-0.6B.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 53728–53741, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

Yasumasa Saisho. Stochastic differential equations for multi-dimensional domain with reflecting boundary. *Probability Theory and Related Fields*, 74(3): 455–477, September 1987. ISSN 1432-2064. doi: 10.1007/BF00699100.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

S. Shahshahani. A new mathematical framework for the study of linkage and selection. *Memoirs of the American Mathematical Society*, 17(211):ix+34, 1979. doi: 10.1090/memo/0211. URL https://www.ams.org/books/memo/0211/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Yuda Song, Julia Kempe, and Remi Munos. Outcome-based exploration for llm reasoning, 2025. URL https://arxiv.org/abs/2509.06941.

Chen Wang, Zhaochun Li, Jionghao Bai, Yuzhi Zhang, Shisheng Cui, Zhou Zhao, and Yue Wang. Arbitrary entropy policy optimization breaks the exploration bottleneck of reinforcement learning, 2025. URL https://arxiv.org/abs/2510.08141.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*, September 2022.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization. *Journal of the American Statistical Association*, 120(552):2154–2164, October 2025. ISSN 0162-1459. doi: 10.1080/01621459.2025.2555067.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-aware policy optimization for large language model reasoning, 2025. URL https://arxiv.org/abs/2505.23433.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. The price of format: Diversity collapse in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15454–15468, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.836. URL https://aclanthology.org/2025.findings-emnlp.836/.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*, October 2022.

Rosie Zhao, Alexandru Meterez, Sham M. Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo Chamber: RL Post-training Amplifies Behaviors Learned in Pretraining. In *Second Conference on Language Modeling*, August 2025.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No] The linked public repository provides code, configurations, tests, and audit commands, but exact paper reproduction additionally depends on frozen release bundles and large external artifacts outside the code-only snapshot.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creators if your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

Guide to the Supplementary Material. The supplement is organized in three parts: core theory for the DCR framework and its collapse diagnosis, auxiliary technical analysis for parameterization and stochastic effects, and empirical plus contextual material that validates and situates the theory. The list below is intended as a quick navigation guide rather than a full table of contents.

How to read this supplement. For the main theoretical arc, start with Appendices A–C. For optimization and scaling technicalities, focus on Appendices D and E. For empirical validation and external grounding, go directly to Appendices F and G.

A THE DCR GRADIENT FLOW FRAMEWORK

This appendix develops the continuous-time dynamical-systems backbone of Distributional Creative Reasoning (DCR). For a fixed prompt, DCR optimizes a regularized objective over the simplex of complete reasoning traces. The continuous-time view is useful for three recurring reasons:

1. it gives an exact *log-ratio identity* for pairwise trace probabilities;
2. it isolates an *entropy-barrier criterion* that keeps trajectories away from the boundary; and
3. it yields a clean *global convergence theorem* for the regularized DCR objective.

Throughout, the trace space is finite, so all objects are finite-dimensional and all derivatives are classical. The key scalar quantity is the effective entropy coefficient

$$A := \lambda\alpha + \beta_{\mathrm{KL}} + \varepsilon,$$

which will always be assumed strictly positive. The condition $A > 0$ has two roles: it creates an interior barrier against boundary collapse, and—when combined with a positive-semidefinite creativity kernel—it makes the objective strictly concave on the simplex.

A.1 The trace simplex and the master DCR ODE

Fix a prompt x, and let $\mathcal{S}_T$ be the finite set of all traces of length at most T. Write

$$S := |\mathcal{S}_T| \geq 2.$$

A policy over traces is a probability vector

$$p \in \Delta^{S-1} := \{p \in [0,1]^S : \sum_{i=1}^{S} p_i = 1\},$$

with relative interior

$$\operatorname{int} \Delta^{S-1} := \{p \in \Delta^{S-1} : p_i > 0 \text{ for all } i\}.$$

The tangent space of the simplex is

$$T_\Delta := \mathbf{1}^\perp := \{v \in \mathbb{R}^S : \langle \mathbf{1} | v \rangle = 0\}.$$

Standing assumptions. Throughout this appendix we assume:

(A1) $K \in \mathbb{R}^{S \times S}$ is symmetric;

(A2) $\lambda, \alpha, \beta, \beta_{\mathrm{KL}}, \varepsilon \geq 0$;

(A3) $p_{\mathrm{base}} \in \operatorname{int} \Delta^{S-1}$, so $\log p_{\mathrm{base},i}$ is finite for every i;

(A4) $A = \lambda\alpha + \beta_{\mathrm{KL}} + \varepsilon > 0$.

Whenever strict concavity or global convergence is invoked, we additionally assume $K \succeq 0$.

The regularized DCR objective. We study

$$\widetilde{J}[p] = \sum_{i=1}^{S} U_i p_i + \lambda\alpha\, H[p] - \lambda\beta\, p^\top K p - \beta_{\mathrm{KL}} D_{\mathrm{KL}}(p \| p_{\mathrm{base}}) + \varepsilon H[p],$$

where

$$H[p] := -\sum_{i=1}^{S} p_i \log p_i, \qquad D_{\mathrm{KL}}(p \| p_{\mathrm{base}}) := \sum_{i=1}^{S} p_i \log \frac{p_i}{p_{\mathrm{base},i}}.$$

We use the convention $0 \log 0 := 0$, so both H and $D_{\mathrm{KL}}(\,\cdot\,\|p_{\mathrm{base}})$ extend continuously to the closed simplex.

It is convenient to collect all logarithmic terms into the single coefficient A, and to define the selective score field

$$\phi_i(p) := U_i - 2\lambda\beta(Kp)_i + \beta_{\mathrm{KL}} \log p_{\mathrm{base},i}.$$

Then, for $p \in \operatorname{int} \Delta^{S-1}$,

$$\frac{\delta \widetilde{J}}{\delta p_i}(p) = \phi_i(p) - A(1 + \log p_i).$$

The Shahshahani geometry. The Shahshahani metric is the natural Riemannian metric on the simplex (Shahshahani, 1979; Amari, 1998)

$$g_p(u, v) := \sum_{i=1}^{S} \frac{u_i v_i}{p_i}, \qquad u, v \in T_\Delta.$$

For a smooth function f on $\operatorname{int} \Delta^{S-1}$, its Shahshahani gradient is the unique vector $\nabla_{\mathrm{Sh}} f(p) \in T_\Delta$ satisfying

$$g_p(\nabla_{\mathrm{Sh}} f(p), v) = df(p) \cdot v \qquad \text{for all } v \in T_\Delta.$$

Equivalently,

$$\big(\nabla_{\mathrm{Sh}} f(p)\big)_i = p_i \left(\frac{\partial f}{\partial p_i}(p) - \sum_{j=1}^{S} p_j \frac{\partial f}{\partial p_j}(p) \right).$$

Applying this to $\widetilde{J}$ yields the DCR flow.

Definition A.1 (Master DCR ODE). The continuous-time DCR dynamics are

$$\dot{p}_i = p_i\big(\phi_i(p) - \bar{\phi}(p)\big) - A\, p_i\big(\log p_i - \langle \log p \rangle\big), \qquad i = 1, \dots, S,$$

where

$$\bar{\phi}(p) := \sum_{j=1}^{S} p_j \phi_j(p), \qquad \langle \log p \rangle := \sum_{j=1}^{S} p_j \log p_j = -H[p].$$

Equivalently, $\dot{p} = \nabla_{\mathrm{Sh}} \widetilde{J}(p)$ on $\operatorname{int} \Delta^{S-1}$.

The first term is a replicator-type selection term (Hofbauer and Sigmund, 1998), while the second is an entropic barrier. The barrier becomes large and inward-pointing whenever a coordinate becomes very small.

A.2 The universal log-ratio identity

The most useful feature of Definition A.1 is that the global coupling terms disappear in pairwise log-ratio coordinates.

Theorem A.2 (Universal log-ratio identity). Let $p(\cdot)$ be a solution of the master DCR ODE that remains in $\operatorname{int} \Delta^{S-1}$. For any pair of traces i, j, define

$$z_{ij}(t) := \log \frac{p_i(t)}{p_j(t)}.$$

Then

$$\frac{d}{dt} z_{ij}(t) = \big(\phi_i(p(t)) - \phi_j(p(t))\big) - A\, z_{ij}(t).$$

Proof. By the chain rule,

$$\frac{d}{dt} z_{ij}(t) = \frac{\dot{p}_i(t)}{p_i(t)} - \frac{\dot{p}_j(t)}{p_j(t)}.$$

From Definition A.1,

$$\frac{\dot{p}_i}{p_i} = \phi_i(p) - \bar{\phi}(p) - A\big(\log p_i - \langle \log p \rangle\big),$$

and the same identity holds with i replaced by j. Subtracting cancels both $\bar{\phi}(p)$ and $\langle \log p \rangle$, leaving

$$\frac{d}{dt} z_{ij}(t) = \big(\phi_i(p(t)) - \phi_j(p(t))\big) - A\big(\log p_i(t) - \log p_j(t)\big).$$

Since $\log p_i - \log p_j = z_{ij}$, the claim follows. □

The identity can be integrated explicitly:

$$z_{ij}(t) = z_{ij}(0) e^{-At} + \int_0^t e^{-A(t-s)} \big(\phi_i(p(s)) - \phi_j(p(s))\big)\, ds.$$

This formula will be reused repeatedly in the later appendices.

A.3 Barrier dominance and forward invariance

The log-ratio identity only makes sense while coordinates stay positive. To turn strict positivity into a quantitative statement, we work on a *trimmed simplex*:

$$\delta_\star \in (0, 1/S), \qquad \Delta^{S-1}_{\delta_\star} := \{p \in \Delta^{S-1} : \min_i p_i \geq \delta_\star\}.$$

Think of $\delta_\star$ as a computational probability floor.

Step 1: an entropy gap on active faces. Whenever a coordinate touches the floor $\delta_\star$, the entropic barrier contributes an inward drift that is uniformly bounded below.

Lemma A.3 (Entropy face gap). If $p \in \Delta^{S-1}$ and $p_i = \delta_\star$, then

$$\langle \log p \rangle - \log \delta_\star \geq L_S(\delta_\star) := (1-\delta_\star) \log \frac{1-\delta_\star}{(S-1)\delta_\star} > 0.$$

Proof. Fix i and minimize

$$f(p) := \sum_{j=1}^{S} p_j \log p_j - \log \delta_\star$$

subject to $p_i = \delta_\star$ and $\sum_{j \neq i} p_j = 1 - \delta_\star$. Because $x \mapsto x \log x$ is strictly convex, the minimum over the remaining coordinates is attained when they are uniform:

$$p_j = \frac{1-\delta_\star}{S-1}, \qquad j \neq i.$$

Substituting gives

$$\begin{aligned} \min f &= \delta_\star \log \delta_\star + (1-\delta_\star) \log \frac{1-\delta_\star}{S-1} - \log \delta_\star \\ &= (1-\delta_\star) \log \frac{1-\delta_\star}{(S-1)\delta_\star} = L_S(\delta_\star). \end{aligned}$$

Since $\delta_\star < 1/S$, the logarithm is strictly positive, so $L_S(\delta_\star) > 0$. □

Step 2: a generic forward-invariance criterion. The theorem below is stated for any locally Lipschitz score field on a neighborhood of the trimmed simplex. This is the level of regularity actually needed in the proof.

Theorem A.4 (Barrier dominance and forward invariance). Let

$$H := \{p \in \mathbb{R}^S : \langle \mathbf{1} | p \rangle = 1\}$$

be the affine simplex hyperplane, and fix $\delta_\star \in (0, 1/S)$. Let $\mathcal{U} \subset H$ be a relative-open neighborhood of $\Delta^{S-1}_{\delta_\star}$ such that

$$\mathcal{U} \subset H \cap (0, \infty)^S,$$

and assume that $\phi : \mathcal{U} \to \mathbb{R}^S$ is locally Lipschitz.

Consider on $\mathcal{U}$ the vector field

$$F_i(p) := p_i\big(\phi_i(p) - \bar{\phi}(p)\big) - A\, p_i\big(\log p_i - \langle \log p \rangle\big),$$

where

$$\bar{\phi}(p) := \sum_{j=1}^{S} p_j \phi_j(p), \qquad \langle \log p \rangle := \sum_{j=1}^{S} p_j \log p_j.$$

Define the maximal outward score pressure on the trimmed simplex by

$$M_{\mathrm{out}}(\delta_\star) := \sup_{p \in \Delta^{S-1}_{\delta_\star}} \max_{1 \leq i \leq S} \big(\bar{\phi}(p) - \phi_i(p)\big).$$

Because ϕ is continuous and $\Delta^{S-1}_{\delta_\star}$ is compact, $M_{\mathrm{out}}(\delta_\star) < \infty$.

If

$$A\, L_S(\delta_\star) \geq M_{\mathrm{out}}(\delta_\star),$$

then $\Delta^{S-1}_{\delta_\star}$ is forward-invariant under the ODE $\dot{p} = F(p)$. If the inequality is strict, then every active face is strictly repelling: whenever $p_i = \delta_\star$,

$$\dot{p}_i > 0.$$

Proof. Let $\mathcal{K} := \Delta^{S-1}_{\delta_\star}$. Since ϕ is locally Lipschitz on $\mathcal{U}$ and the logarithmic term is C^1 on $\mathcal{U} \subset H \cap (0,\infty)^S$, the vector field F is locally Lipschitz on $\mathcal{U}$.

First, F is tangent to the affine hyperplane H: indeed,

$$\sum_{i=1}^{S} F_i(p) = \sum_{i=1}^{S} p_i(\phi_i - \bar{\phi}) - A \sum_{i=1}^{S} p_i(\log p_i - \langle \log p \rangle) = 0.$$

Hence $F(p) \in T_\Delta = \mathbf{1}^\perp$ for every $p \in \mathcal{U}$.

Now fix $p \in \partial\mathcal{K}$, and let

$$\mathcal{I}_{\mathrm{act}}(p) := \{i : p_i = \delta_\star\}$$

be the active set. The Bouligand tangent cone of $\mathcal{K}$ at p, viewed inside the affine space H, is

$$T_{\mathcal{K}}(p) = \{v \in \mathbf{1}^\perp : v_i \geq 0 \text{ for all } i \in \mathcal{I}_{\mathrm{act}}(p)\}.$$

So it is enough to show $F_i(p) \geq 0$ for every active coordinate i.

If $i \in \mathcal{I}_{\mathrm{act}}(p)$, then $p_i = \delta_\star$, and

$$F_i(p) = \delta_\star \Big[\phi_i(p) - \bar{\phi}(p) + A\big(\langle \log p \rangle - \log \delta_\star\big) \Big].$$

By Lemma A.3,

$$\langle \log p \rangle - \log \delta_\star \geq L_S(\delta_\star),$$

and by definition of $M_{\mathrm{out}}(\delta_\star)$,

$$\phi_i(p) - \bar{\phi}(p) \geq -M_{\mathrm{out}}(\delta_\star).$$

Therefore

$$F_i(p) \geq \delta_\star \Big[-M_{\mathrm{out}}(\delta_\star) + A\, L_S(\delta_\star) \Big] \geq 0.$$

Thus $F(p) \in T_{\mathcal{K}}(p)$ for every $p \in \partial\mathcal{K}$. Since $\mathcal{K}$ is a compact convex subset of the finite-dimensional affine space H, the standard Nagumo viability theorem (Nagumo, 1942) implies that $\mathcal{K}$ is forward-invariant under $\dot{p} = F(p)$.

If $A\, L_S(\delta_\star) > M_{\mathrm{out}}(\delta_\star)$, the same computation yields $F_i(p) > 0$ on every active face. □

For DCR itself, the hypotheses of Theorem A.4 are automatic: the score field ϕ is affine in p, hence globally Lipschitz on $\mathbb{R}^S$, and therefore locally Lipschitz on every relative-open neighborhood of a trimmed simplex.

Corollary A.4.1 (Uniform interior confinement for DCR). Under the standing assumptions of this appendix, let $p(0) \in \operatorname{int} \Delta^{S-1}$. Then there exists

$$\delta_\star = \delta_\star(p(0)) \in \big(0, \min_i p_i(0)\big)$$

such that the unique solution of the master DCR ODE exists for all $t \geq 0$ and satisfies

$$p(t) \in \Delta^{S-1}_{\delta_\star} \qquad \text{for all } t \geq 0.$$

In particular, no coordinate can hit 0 in finite time.

Proof. Because $p_{\mathrm{base}} \in \operatorname{int} \Delta^{S-1}$, the map $p \mapsto \phi(p)$ is continuous on the closed simplex: U is fixed, $p \mapsto Kp$ is linear, and every $\log p_{\mathrm{base},i}$ is finite. Since Δ^{S-1} is compact,

$$M_\phi := \sup_{p \in \Delta^{S-1}} \max_{1 \leq i \leq S} |\phi_i(p)| < \infty.$$

Hence, for every $p \in \Delta^{S-1}$,

$$\bar{\phi}(p) - \phi_i(p) \leq \big|\bar{\phi}(p)\big| + |\phi_i(p)| \leq 2M_\phi,$$

so

$$M_{\mathrm{out}}(\delta) \leq 2M_\phi \qquad \text{for every } \delta \in (0, 1/S).$$

By Lemma A.3, $L_S(\delta) \to \infty$ as $\delta \downarrow 0$. Therefore we may choose

$$0 < \delta_\star < \min_i p_i(0)$$

such that

$$A\, L_S(\delta_\star) \geq 2M_\phi.$$

Then $p(0) \in \Delta^{S-1}_{\delta_\star}$, and Theorem A.4 shows that $\Delta^{S-1}_{\delta_\star}$ is forward-invariant for the DCR flow.

It remains to justify global existence and uniqueness. On the relative-open set

$$\mathcal{U}_{\delta_\star/2} := \{p \in H : p_i > \delta_\star/2 \text{ for all } i\},$$

the DCR vector field is C^1, hence locally Lipschitz. Standard ODE theory on the affine space H therefore gives a unique maximal solution through $p(0)$. Because the trajectory never leaves the compact set $\Delta^{S-1}_{\delta_\star}$, no finite-time blow-up is possible, so the maximal solution extends to all $t \geq 0$. The already-proved forward invariance then yields

$$p(t) \in \Delta^{S-1}_{\delta_\star} \qquad \forall t \geq 0.$$

□

A.4 Global convergence of DCR training

We now turn to the asymptotic behavior of DCR when the creativity kernel is positive semidefinite. The main result is that the regularized objective has a unique interior maximizer and that every interior trajectory converges to it.

Step 1: strict concavity and the unique interior maximizer.

Lemma A.5 (Strict concavity and unique interior maximizer). Assume $A > 0$, $K = K^\top$, and $K \succeq 0$. Then the restriction of $\widetilde{J}$ to $\operatorname{int} \Delta^{S-1}$ is A-strongly concave relative to the affine hull

$$H = \{p \in \mathbb{R}^S : \langle \mathbf{1} | p \rangle = 1\}.$$

Consequently, $\widetilde{J}$ admits a unique global maximizer

$$p^\star \in \operatorname{int} \Delta^{S-1}.$$

Proof. Let $p \in \operatorname{int} \Delta^{S-1}$ and $v \in T_\Delta = \mathbf{1}^\perp$. Differentiating twice along the affine hyperplane H gives

$$\left\langle v \middle| \nabla^2 \widetilde{J}(p)\, v \right\rangle = -2\lambda\beta\, v^\top K v - A \sum_{i=1}^S \frac{v_i^2}{p_i}.$$

Since $K \succeq 0$ and $\lambda\beta \geq 0$,

$$-2\lambda\beta\, v^\top K v \leq 0.$$

Also $p_i \leq 1$ for every coordinate, so $1/p_i \geq 1$, hence

$$-A \sum_{i=1}^S \frac{v_i^2}{p_i} \leq -A \sum_{i=1}^S v_i^2 = -A\|v\|_2^2.$$

Therefore

$$\left\langle v \middle| \nabla^2 \widetilde{J}(p)\, v \right\rangle \leq -A\|v\|_2^2 \qquad \forall\, v \in T_\Delta.$$

Since $\operatorname{int} \Delta^{S-1}$ is convex, integrating this Hessian bound along the line segment from p to q yields, for all $p, q \in \operatorname{int} \Delta^{S-1}$,

$$\widetilde{J}(q) \leq \widetilde{J}(p) + \left\langle \Pi_{\mathbf{1}^\perp} \nabla \widetilde{J}(p) \middle| q - p \right\rangle - \frac{A}{2}\|q - p\|_2^2.$$

This is the standard first-order strong-concavity inequality relative to the affine hull H.

Because $\widetilde{J}$ extends continuously to the compact simplex Δ^{S-1}, at least one maximizer exists. It remains to show that no boundary point can maximize $\widetilde{J}$.

Let $q \in \partial\Delta^{S-1}$, and define

$$Z := \{k : q_k = 0\}, \qquad m := |Z| \geq 1.$$

Choose any j with $q_j > 0$, and set

$$v := \sum_{k \in Z} e_k - m e_j, \qquad q(t) := q + tv, \qquad 0 < t < q_j/m.$$

Then $q(t) \in \operatorname{int}\Delta^{S-1}$ for all such t: every zero coordinate becomes $t > 0$, the j-th coordinate remains positive, and all other coordinates are unchanged.

Write

$$\widetilde{J}(p) = \Psi(p) - A\sum_{i=1}^{S} p_i \log p_i,$$

where

$$\Psi(p) := \sum_{i=1}^{S} U_i p_i - \lambda\beta\, p^\top K p + \beta_{\mathrm{KL}} \sum_{i=1}^{S} p_i \log p_{\mathrm{base},i}.$$

The function Ψ is C^1 on a neighborhood of the simplex, so

$$\Psi(q(t)) = \Psi(q) + O(t) \qquad \text{as } t \downarrow 0.$$

For the entropy term, a direct expansion gives

$$-\sum_{i=1}^{S} q_i(t) \log q_i(t) = -\sum_{i=1}^{S} q_i \log q_i - m\, t \log t + O(t) \qquad \text{as } t \downarrow 0.$$

Hence

$$\widetilde{J}(q(t)) - \widetilde{J}(q) = A\, m(-t \log t) + O(t).$$

Since $-t\log t \gg t$ as $t \downarrow 0$, the right-hand side is strictly positive for all sufficiently small $t > 0$. Thus every boundary point admits an inward perturbation that increases $\widetilde{J}$, so no boundary point can maximize it.

Therefore every maximizer lies in $\operatorname{int}\Delta^{S-1}$. Since every maximizer is interior, the strong concavity inequality on $\operatorname{int}\Delta^{S-1}$ implies uniqueness. The unique maximizer is denoted $p^\star$. □

Step 2: a strict Lyapunov identity.

Theorem A.6 (Strict Lyapunov identity). Along any solution $t \mapsto p(t) \in \operatorname{int}\Delta^{S-1}$ of the master DCR ODE,

$$\frac{d}{dt}\widetilde{J}\big(p(t)\big) = \sum_{i=1}^{S} p_i(t) \left(\frac{\delta \widetilde{J}}{\delta p_i}(p(t)) - \sum_{j=1}^{S} p_j(t) \frac{\delta \widetilde{J}}{\delta p_j}(p(t)) \right)^2 \geq 0.$$

Moreover, equality holds if and only if $p(t)$ is a stationary point of the DCR flow.

Proof. Let $p_t := p(t)$. Since $\dot{p}_t = \nabla_{\mathrm{Sh}}\widetilde{J}(p_t)$,

$$\frac{d}{dt}\widetilde{J}(p_t) = d\widetilde{J}(p_t) \cdot \dot{p}_t = g_{p_t}\big(\nabla_{\mathrm{Sh}}\widetilde{J}(p_t), \nabla_{\mathrm{Sh}}\widetilde{J}(p_t)\big).$$

By definition of the Shahshahani metric,

$$g_{p_t}\big(\nabla_{\mathrm{Sh}}\widetilde{J}, \nabla_{\mathrm{Sh}}\widetilde{J}\big) = \sum_{i=1}^{S} \frac{\dot{p}_i(t)^2}{p_i(t)}.$$

Substituting the coordinate formula for $\dot{p}_i$ yields the stated identity.

Since every $p_i(t)$ is strictly positive, the sum vanishes if and only if $\dot{p}_i(t) = 0$ for every i, i.e. if and only if $p(t)$ is stationary. □

Step 3: global convergence and an exponential rate on trimmed simplices.

Theorem A.7 (Global convergence; proof of Theorem 3.1). Assume $A > 0$, $K = K^\top$, and $K \succeq 0$. For every initial condition

$$p(0) \in \operatorname{int} \Delta^{S-1},$$

the master DCR ODE admits a unique global solution $p(t) \in \operatorname{int} \Delta^{S-1}$ for all $t \geq 0$, and

$$p(t) \to p^\star \qquad \text{as } t \to \infty,$$

where $p^\star$ is the unique maximizer from Lemma A.5.

Moreover, if the trajectory is contained in a forward-invariant trimmed simplex $\Delta_{\delta_\star}^{S-1}$, then the convergence is exponentially fast:

$$\widetilde{J}(p^\star) - \widetilde{J}(p_t) \leq e^{-2A\delta_\star t} \left(\widetilde{J}(p^\star) - \widetilde{J}(p_0)\right),$$

and

$$\|p_t - p^\star\|_1 \leq \sqrt{\frac{2S}{A}} \left(\widetilde{J}(p^\star) - \widetilde{J}(p_0)\right)^{1/2} e^{-A\delta_\star t}.$$

Proof. We split the argument into four steps.

Step 1: well-posedness and confinement. By Corollary A.4.1, there exists $\delta_\star > 0$ (depending on $p(0)$) such that the unique DCR trajectory through $p(0)$ stays in $\Delta_{\delta_\star}^{S-1}$ for all $t \geq 0$. In particular, the solution is global and remains in the interior.

Step 2: LaSalle reduction to the derivative-zero set. By Theorem A.6, the map $t \mapsto \widetilde{J}(p_t)$ is nondecreasing. Since the trajectory is contained in the compact invariant set $\Delta_{\delta_\star}^{S-1}$ and $\widetilde{J}$ is continuous, $\widetilde{J}(p_t)$ is bounded above. Standard LaSalle invariance (LaSalle, 1960) therefore gives that the ω-limit set of the trajectory is contained in

$$\mathcal{Z} := \{p \in \Delta_{\delta_\star}^{S-1} : \frac{d}{dt}\widetilde{J}(p) = 0\}.$$

Step 3: identification of the invariant set. By Theorem A.6,

$$\frac{d}{dt}\widetilde{J}(p) = 0 \qquad \Longleftrightarrow \qquad \dot{p} = 0.$$

Because $\Delta_{\delta_\star}^{S-1} \subset \operatorname{int} \Delta^{S-1}$, the stationarity condition

$$0 = p_i \left(\frac{\delta \widetilde{J}}{\delta p_i}(p) - \sum_{j=1}^{S} p_j \frac{\delta \widetilde{J}}{\delta p_j}(p) \right)$$

is equivalent to

$$\frac{\delta \widetilde{J}}{\delta p_i}(p) = \mu \qquad \text{for all } i$$

for some scalar μ. This is exactly the first-order optimality condition for an interior extremum on the simplex. By Lemma A.5, there is only one such point, namely $p^\star$. Hence the largest invariant subset of $\mathcal{Z}$ is $\{p^\star\}$, and LaSalle yields

$$p_t \to p^\star \qquad \text{as } t \to \infty.$$

Step 4: exponential rate on a trimmed simplex. Let

$$\Pi_{\mathbf{1}^\perp} := I - \frac{1}{S}\mathbf{1}\mathbf{1}^\top$$

denote the Euclidean projection onto $T_\Delta = \mathbf{1}^\perp$.

Because $\widetilde{J}$ is A-strongly concave on $\operatorname{int}\Delta^{S-1}$ relative to its affine hull H, for every $p \in \operatorname{int}\Delta^{S-1}$,

$$\widetilde{J}(p^\star) \le \widetilde{J}(p) + \Big\langle \Pi_{\mathbf{1}^\perp} \nabla \widetilde{J}(p) \Big| p^\star - p \Big\rangle - \frac{A}{2}\|p^\star - p\|_2^2.$$

Maximizing the right-hand side over $h := p^\star - p \in \mathbf{1}^\perp$ gives the Euclidean Polyak–Łojasiewicz inequality

$$\Big\|\Pi_{\mathbf{1}^\perp} \nabla \widetilde{J}(p)\Big\|_2^2 \ge 2A\big(\widetilde{J}(p^\star) - \widetilde{J}(p)\big).$$

Next, the squared Shahshahani gradient norm is

$$\Big\|\nabla_{\mathrm{Sh}} \widetilde{J}(p)\Big\|_{\mathrm{Sh}}^2 = \sum_{i=1}^S p_i \left(\frac{\delta \widetilde{J}}{\delta p_i}(p) - \sum_{j=1}^S p_j \frac{\delta \widetilde{J}}{\delta p_j}(p)\right)^2.$$

Since $p_i \ge \delta_\star$ on $\Delta_{\delta_\star}^{S-1}$,

$$\Big\|\nabla_{\mathrm{Sh}} \widetilde{J}(p)\Big\|_{\mathrm{Sh}}^2 \ge \delta_\star \sum_{i=1}^S \left(\frac{\delta \widetilde{J}}{\delta p_i}(p) - \sum_{j=1}^S p_j \frac{\delta \widetilde{J}}{\delta p_j}(p)\right)^2.$$

For any vector $g \in \mathbb{R}^S$, the function $c \mapsto \sum_i (g_i - c)^2$ is minimized at the arithmetic mean $c = \frac{1}{S}\sum_i g_i$. Therefore

$$\begin{aligned}\sum_{i=1}^S \left(\frac{\delta \widetilde{J}}{\delta p_i} - \sum_{j=1}^S p_j \frac{\delta \widetilde{J}}{\delta p_j}\right)^2 &\ge \sum_{i=1}^S \left(\frac{\delta \widetilde{J}}{\delta p_i} - \frac{1}{S}\sum_{j=1}^S \frac{\delta \widetilde{J}}{\delta p_j}\right)^2 \\ &= \Big\|\Pi_{\mathbf{1}^\perp} \nabla \widetilde{J}(p)\Big\|_2^2.\end{aligned}$$

Combining the last two displays with the Euclidean Polyak–Łojasiewicz inequality yields the Shahshahani Polyak–Łojasiewicz bound

$$\Big\|\nabla_{\mathrm{Sh}} \widetilde{J}(p)\Big\|_{\mathrm{Sh}}^2 \ge 2A\delta_\star\big(\widetilde{J}(p^\star) - \widetilde{J}(p)\big).$$

By Theorem A.6,

$$\frac{d}{dt}\widetilde{J}(p_t) = \Big\|\nabla_{\mathrm{Sh}} \widetilde{J}(p_t)\Big\|_{\mathrm{Sh}}^2.$$

Define the objective gap

$$E(t) := \widetilde{J}(p^\star) - \widetilde{J}(p_t).$$

Then

$$\dot{E}(t) = -\frac{d}{dt}\widetilde{J}(p_t) \le -2A\delta_\star E(t),$$

and Grönwall's lemma gives

$$E(t) \le E(0) e^{-2A\delta_\star t}.$$

Finally, strong concavity at the interior maximizer implies

$$\widetilde{J}(p) \le \widetilde{J}(p^\star) + \Big\langle \nabla \widetilde{J}(p^\star) \Big| p - p^\star \Big\rangle - \frac{A}{2}\|p - p^\star\|_2^2.$$

Because $p^\star$ is an interior simplex maximizer, there exists $\mu \in \mathbb{R}$ such that

$$\nabla \widetilde{J}(p^\star) = \mu\,\mathbf{1}.$$

Since $p - p^\star \in \mathbf{1}^\perp$, the linear term vanishes, hence

$$\widetilde{J}(p^\star) - \widetilde{J}(p) \ge \frac{A}{2}\|p - p^\star\|_2^2.$$

Applying this at $p = p_t$ and using $\|x\|_1 \le \sqrt{S}\|x\|_2$ gives

$$\|p_t - p^\star\|_1 \le \sqrt{\frac{2S}{A}}\, E(t)^{1/2} \le \sqrt{\frac{2S}{A}}\, E(0)^{1/2} e^{-A\delta_\star t}.$$

This proves the stated rate. □

Remark. The exponential rate in Theorem A.7 is global in time along each interior trajectory, but it is not uniform over all interior initializations: the floor $\delta_\star$ comes from the confinement argument and therefore depends on the starting point $p(0)$.

B PROOF OF THE DIVERSITY-DECAY THEOREM

This appendix analyzes three scalar training mechanisms on the finite trace space

$$\mathcal{S}_T = \mathcal{C} \sqcup \mathcal{I}, \qquad \rho(p) := \sum_{c \in \mathcal{C}} p_c,$$

where $\mathcal{C}$ is the set of correct traces and $\mathcal{I}$ the set of incorrect traces. Our goal is to understand whether a scalar training signal preserves or destroys *diversity inside the correct class*. Throughout this appendix, the DCR creativity kernel is turned off, the KL tether is absent, and an optional entropy barrier with coefficient $\varepsilon \geq 0$ may be added to the probability-space dynamics as an algorithm-agnostic regularizer.

The key message is simple. When the objective depends only on scalar correctness, scalar normalized reward, or scalar preferred–dispreferred margins, it has no access to semantic relations among correct traces. As a result, it cannot actively spread mass across genuinely distinct correct strategies. Depending on the algorithm, one gets either:

- **winner-take-all amplification** inside the correct class (STaR-style rejection fine-tuning and exact mean-field GRPO batch-gradient dynamics), or
- **reference-ratio regression** (DPO under one-sided or two-sided exchangeable-pair surrogates).

Entropy can damp relative log-ratios, but it does not create a semantic balancing force.

Generic logit-to-probability reduction. Let $\theta \in \mathbb{R}^{|\mathcal{S}_T|}$ be trace-level logits and $p_\theta = \mathrm{softmax}(\theta)$ the induced policy over traces. Any deterministic logit drift

$$\dot{\theta}_i = \psi_i(p)$$

induces the probability-space dynamics

$$\dot{p}_i = p_i\Big(\psi_i(p) - \bar{\psi}(p)\Big), \qquad \bar{\psi}(p) := \sum_j p_j \psi_j(p).$$

If we additionally include the entropy barrier $\varepsilon H[p]$, the resulting probability flow becomes

$$\dot{p}_i = p_i\Big(\psi_i(p) - \bar{\psi}(p)\Big) - \varepsilon p_i\Big(\log p_i - \langle \log p \rangle\Big), \qquad \langle \log p \rangle := \sum_j p_j \log p_j. \tag{14}$$

Therefore, whenever $p_a(t), p_b(t) > 0$,

$$\frac{d}{dt} \log \frac{p_a(t)}{p_b(t)} = \psi_a(p(t)) - \psi_b(p(t)) - \varepsilon \log \frac{p_a(t)}{p_b(t)}. \tag{15}$$

Equation (15) is the only structural fact from Appendix A that we will use below.

Remark B.1 (Modeling scope)**.** *Each subsection states the* exact *simplified model it analyzes. Section B.1 studies the online rejection-finetuning core of STaR-style self-training (Zelikman et al., 2022). Section B.2 studies the unclipped, no-KL, outcome-level GRPO* per-batch surrogate *in a one-step tabular trace model and derives its exact Euclidean-logit mean field (Shao et al., 2024). Section B.3 studies a* one-sided *exchangeable-pair DPO surrogate chosen to isolate preferred-side reference-margin geometry (Rafailov et al., 2023). The purpose of this appendix is not to reproduce every engineering detail of each implementation, but to expose the within-class selection mechanism governing diversity decay.*

B.1 STaR-style online rejection fine-tuning: winner-take-all fixation

Context. We isolate the online rejection-finetuning core of STaR-style self-training. The policy samples traces, a verifier filters for correctness, and the model is fine-tuned by supervised maximum likelihood on the accepted traces. Because accepted traces are replayed *in proportion to their conditional sampling frequency*, the update amplifies whichever correct trace is already more common.

Proposition B.1 (STaR-style rejection-finetuning score field)**.** *Under self-sampling with rejection, the accepted target distribution is*

$$q_c = \frac{p_c}{\rho(p)} \quad (c \in \mathcal{C}), \qquad q_i = 0 \quad (i \in \mathcal{I}),$$

whenever $\rho(p) > 0$. If the model takes an infinitesimal Euclidean logit-ascent step on the log-likelihood against q, then the induced trace-level logit drift is

$$\psi_i^{\mathrm{RFT}}(p) = q_i - p_i = \begin{cases} p_i \dfrac{1-\rho(p)}{\rho(p)}, & i \in \mathcal{C}, \\ -p_i, & i \in \mathcal{I}. \end{cases}$$

Proof. For multinomial logits, gradient ascent on the log-likelihood against target distribution q gives

$$\dot{\theta}_k = q_k - p_k.$$

Substituting the accepted-target law yields, for $c \in \mathcal{C}$,

$$\psi_c^{\mathrm{RFT}}(p) = \frac{p_c}{\rho(p)} - p_c = p_c \frac{1-\rho(p)}{\rho(p)},$$

and for $i \in \mathcal{I}$,

$$\psi_i^{\mathrm{RFT}}(p) = 0 - p_i = -p_i.$$

□

Theorem B.1 (Correct-correct log-ratio dynamics for STaR-style rejection fine-tuning)**.** *Let $a, b \in \mathcal{C}$ and assume $p_a(t), p_b(t) > 0$. Then under the flow (14) with score field ψ^{RFT},*

$$\frac{d}{dt} \log \frac{p_a(t)}{p_b(t)} = \frac{1-\rho(p(t))}{\rho(p(t))} \Big(p_a(t) - p_b(t) \Big) - \varepsilon \log \frac{p_a(t)}{p_b(t)}. \tag{16}$$

Equivalently, with $z_{ab}(t) := \log \frac{p_a(t)}{p_b(t)}$,

$$\dot{z}_{ab}(t) = p_b(t) \big(e^{z_{ab}(t)} - 1 \big) \frac{1-\rho(p(t))}{\rho(p(t))} - \varepsilon z_{ab}(t).$$

Proof. Apply the generic identity (15) with

$$\psi_a^{\mathrm{RFT}} - \psi_b^{\mathrm{RFT}} = \frac{1-\rho}{\rho}(p_a - p_b).$$

Since $p_a = p_b e^{z_{ab}}$, the equivalent form follows immediately. □

Remark B.2 (Cold-start singularity)**.** *The accepted-target distribution contains the factor $1/\rho(p)$. Hence the rejection-finetuning score field becomes singular as $\rho(p) \downarrow 0$. This formalizes the familiar cold-start fragility of rejection fine-tuning: if the policy puts vanishing mass on correct traces, the online supervised target itself becomes ill-conditioned.*

Corollary B.1 (Winner-take-all among positive-mass correct traces)**.** *Assume $\varepsilon = 0$ and $0 < \rho(0) < 1$. Let*

$$\mathcal{C}_+ := \{ c \in \mathcal{C} : p_c(0) > 0 \}$$

be the set of correct traces with positive initial mass, and assume there exists a unique

$$a \in \mathcal{C}_+ \qquad \text{such that} \qquad p_a(0) > p_c(0) \quad \forall c \in \mathcal{C}_+ \setminus \{a\}.$$

Then:

1. *$\rho(t)$ is strictly increasing on $\{0 < \rho < 1\}$ and satisfies $\rho(t) \uparrow 1$.*
2. *Every $c \in \mathcal{C} \setminus \mathcal{C}_+$ remains zero for all t, and for every $c \in \mathcal{C}_+ \setminus \{a\}$,*

$$\frac{p_c(t)}{p_a(t)} \to 0.$$

   Equivalently, the normalized correct-class composition converges to the vertex e_a on the support $\mathcal{C}_+$.

3. *Consequently,*

$$p(t) \to e_a.$$

Proof. Fix $c \in \mathcal{C}_+$. Since the probability-space dynamics have the form $\dot{p}_c = p_c(\cdots)$, any coordinate that starts at zero remains zero. Thus traces in $\mathcal{C} \setminus \mathcal{C}_+$ are absorbing and may be ignored.

Define the normalized correct and incorrect compositions

$$s_c := \frac{p_c}{\rho} \quad (c \in \mathcal{C}), \qquad r_i := \frac{p_i}{1-\rho} \quad (i \in \mathcal{I},\ \rho < 1),$$

and the quadratic concentration functionals

$$S_2 := \sum_{c \in \mathcal{C}} s_c^2, \qquad R_2 := \sum_{i \in \mathcal{I}} r_i^2.$$

A direct computation from Proposition B.1 gives

$$\dot{\rho} = \rho(1-\rho)^2\big(S_2 + R_2\big). \tag{17}$$

Since $S_2 \geq 1/|\mathcal{C}_+|$ whenever $\rho > 0$, the right-hand side is strictly positive on $0 < \rho < 1$. Hence ρ is increasing and has a limit $L \in (0, 1]$. If $L < 1$, then for all large t we have

$$\rho(t) \geq \frac{L}{2}, \qquad 1 - \rho(t) \geq \frac{1-L}{2}, \qquad S_2(t) \geq \frac{1}{|\mathcal{C}_+|},$$

so (17) implies

$$\dot{\rho}(t) \geq \frac{L}{2}\Big(\frac{1-L}{2}\Big)^2 \frac{1}{|\mathcal{C}_+|} > 0$$

for all large t, contradicting convergence of $\rho(t)$ to L. Therefore $\rho(t) \uparrow 1$.

Now set $u := 1 - \rho$. Since $\rho \leq 1$ and $S_2 + R_2 \leq 2$,

$$\dot{u} = -\rho u^2(S_2 + R_2) \geq -2u^2.$$

Hence

$$\frac{d}{dt}\frac{1}{u(t)} = -\frac{\dot{u}(t)}{u(t)^2} \leq 2,$$

so

$$u(t) \geq \frac{1}{u(0)^{-1} + 2t}. \tag{18}$$

Therefore the effective time

$$\tau(t) := \int_0^t (1 - \rho(s))\, ds = \int_0^t u(s)\, ds$$

diverges as $t \to \infty$.

Fix $c \in \mathcal{C}_+ \setminus \{a\}$ and define

$$q_c := \frac{p_c}{p_a}.$$

By Theorem B.1 with $\varepsilon = 0$,

$$\frac{d}{dt}\log q_c = \frac{1-\rho}{\rho}(p_c - p_a) = (1-\rho)(s_c - s_a).$$

Passing to the rescaled time $d\tau = (1-\rho)dt$ gives

$$\frac{d}{d\tau}\log q_c = s_c - s_a = s_a(q_c - 1), \qquad \frac{dq_c}{d\tau} = q_c s_a(q_c - 1).$$

Since $q_c(0) < 1$, uniqueness of ODE solutions implies $q_c(\tau) < 1$ for all τ, so q_c is strictly decreasing. Because a remains the largest correct trace on the support $\mathcal{C}_+$, we have

$$s_a(\tau) \geq \frac{1}{|\mathcal{C}_+|}.$$

Since q_c decreases, also $1 - q_c(\tau) \geq 1 - q_c(0)$. Therefore

$$\frac{d}{d\tau}\log q_c = -s_a(1-q_c) \leq -\frac{1-q_c(0)}{|\mathcal{C}_+|}.$$

Integrating in τ yields exponential decay in effective time,

$$q_c(\tau) \leq q_c(0)\exp\Big(-\kappa_c\tau\Big) \quad \text{for some } \kappa_c > 0.$$

Since $\tau(t) \to \infty$, we obtain $q_c(t) \to 0$ for every $c \neq a$ in $\mathcal{C}_+$. Thus the correct-class composition converges to e_a, and since $\rho(t) \to 1$, we conclude $p(t) \to e_a$. □

Interpretation. The amplification factor in (16) is proportional to $(1-\rho)/\rho$. Thus the selective pressure is strongest when the model still makes many errors and weakens near mastery. The slowdown is real, but Corollary B.1 shows that it is not strong enough to prevent deterministic fixation.

B.2 GRPO: exact Euclidean-logit mean field and deterministic within-class amplification

Context. Group Relative Policy Optimization (GRPO) samples a group of outputs for the same prompt, scores them jointly, normalizes rewards by the group mean and standard deviation, and then applies a score-function gradient step. The crucial structural point is that the exact Euclidean-logit mean field of the *per-batch gradient* contains the usual factor $\nabla_\theta \log p_\theta(Y_m)$. In the one-step tabular trace model, that factor contributes an extra p_i term in the mean field. This is what turns apparent class symmetry into *winner-take-all competition inside the correct class.*

We analyze the outcome-level GRPO core under binary outcome supervision,

$$U_c = 1 \quad (c \in \mathcal{C}), \qquad U_i = 0 \quad (i \in \mathcal{I}),$$

with group size $G \geq 2$, no clipping, and no KL term.

Proposition B.2 (Class-conditional normalized advantages and exact mean-field logit drift)**.** *Fix $\rho \in [0,1]$. For a mixed group containing exactly t correct traces, the within-group standardized advantages are*

$$A_C(t) = \sqrt{\frac{G-t}{t}}, \qquad A_I(t) = -\sqrt{\frac{t}{G-t}}, \qquad 1 \leq t \leq G-1,$$

with the convention $A_C(G) = A_I(0) = 0$ on pure groups.

Define the class-conditional expected standardized advantages

$$\alpha_C(\rho) := \mathbb{E}_{S\sim\mathrm{Binom}(G-1,\rho)}\left[\sqrt{\frac{G-1-S}{S+1}}\right],$$

$$\alpha_I(\rho) := -\mathbb{E}_{S\sim\mathrm{Binom}(G-1,\rho)}\left[\sqrt{\frac{S}{G-S}}\right].$$

Then:

1. $\rho\,\alpha_C(\rho) + (1-\rho)\,\alpha_I(\rho) = 0$ *for all* $\rho \in [0,1]$.
2. *There exists a continuous function* $h_G : [0,1] \to (0,\infty)$ *such that*

$$\alpha_C(\rho) = (1-\rho)h_G(\rho), \qquad \alpha_I(\rho) = -\rho h_G(\rho).$$

In particular, $h_G(1) = \sqrt{G-1}$.

3. *For a sampled group* $Y_{1:G} \sim p^{\otimes G}$*, define the unclipped, no-KL per-batch GRPO surrogate*

$$\widehat{\mathcal{L}}_{\mathrm{GRPO}}(\theta; Y_{1:G}) := \frac{1}{G}\sum_{m=1}^{G} \widehat{A}_m \log p_\theta(Y_m),$$

where the sampled normalized advantages $\widehat{A}_m$ *are treated as fixed inside the batch gradient, exactly as in the usual score-function estimator (Williams, 1992). Then the exact Euclidean-logit mean field of this sample gradient is*

$$\psi_i^{\mathrm{GRPO}}(p) := \mathbb{E}_{Y_{1:G}\sim p^{\otimes G}}\left[\partial_{\theta_i}\widehat{\mathcal{L}}_{\mathrm{GRPO}}(\theta; Y_{1:G})\right] = \begin{cases} p_i\,\alpha_C(\rho(p)), & i \in \mathcal{C}, \\ p_i\,\alpha_I(\rho(p)), & i \in \mathcal{I}. \end{cases}$$

Proof. If a group contains exactly t correct traces, then the group mean reward is t/G and the empirical standard deviation is

$$\sigma(t) = \sqrt{\frac{t(G-t)}{G^2}}.$$

Hence for a correct trace in that group,

$$A_C(t) = \frac{1 - t/G}{\sigma(t)} = \sqrt{\frac{G-t}{t}},$$

and for an incorrect trace,

$$A_I(t) = \frac{-t/G}{\sigma(t)} = -\sqrt{\frac{t}{G-t}}.$$

On pure groups ($t = 0$ or $t = G$), the variance is zero, and we follow the standard convention of setting the normalized advantage to zero.

Now condition on the class of one sampled trace. If $c \in \mathcal{C}$, then after conditioning on $Y_1 = c$, the number of additional correct traces among $Y_2, \ldots, Y_G$ is

$$S \sim \mathrm{Binom}(G-1, \rho),$$

and the total number of correct traces in the group is $T = S + 1$. Therefore

$$\mathbb{E}[\widehat{A}_1 \mid Y_1 = c] = \alpha_C(\rho).$$

Similarly, if $i \in \mathcal{I}$, then the total number of correct traces equals $T = S$ with the same $S \sim \mathrm{Binom}(G-1, \rho)$, so

$$\mathbb{E}[\widehat{A}_1 \mid Y_1 = i] = \alpha_I(\rho).$$

This proves the displayed formulas.

For the identity $\rho\alpha_C + (1-\rho)\alpha_I = 0$, note that the normalized advantages in any group sum to zero:

$$\sum_{m=1}^{G} \widehat{A}_m = 0.$$

By exchangeability, $\mathbb{E}[\widehat{A}_1] = 0$, but also

$$\mathbb{E}[\widehat{A}_1] = \rho\,\mathbb{E}[\widehat{A}_1 \mid Y_1 \in \mathcal{C}] + (1-\rho)\,\mathbb{E}[\widehat{A}_1 \mid Y_1 \in \mathcal{I}] = \rho\alpha_C + (1-\rho)\alpha_I.$$

Hence the identity follows.

Next, factor out $(1-\rho)$ from the explicit binomial sum for α_C:

$$\alpha_C(\rho) = (1-\rho)\sum_{s=0}^{G-2}\binom{G-1}{s}\rho^s(1-\rho)^{G-2-s}\sqrt{\frac{G-1-s}{s+1}}.$$

This defines a continuous positive function h_G on $[0,1]$, with

$$h_G(1) = (G-1)\sqrt{\frac{1}{G-1}} = \sqrt{G-1}.$$

The formula for α_I then follows from the zero-mean identity.

Finally, fix one sampled group $Y_{1:G}$. Differentiating the per-batch surrogate gives

$$\partial_{\theta_k}\widehat{\mathcal{L}}_{\mathrm{GRPO}}(\theta; Y_{1:G}) = \frac{1}{G}\sum_{m=1}^{G}\widehat{A}_m\,\partial_{\theta_k}\log p_\theta(Y_m).$$

Since

$$\partial_{\theta_k}\log p_\theta(Y_m) = \mathbf{1}\{Y_m = k\} - p_k,$$

we obtain

$$\partial_{\theta_k}\widehat{\mathcal{L}}_{\mathrm{GRPO}}(\theta; Y_{1:G}) = \frac{1}{G}\sum_{m=1}^{G}\widehat{A}_m\big(\mathbf{1}\{Y_m = k\} - p_k\big).$$

Because $\sum_{m=1}^{G}\widehat{A}_m = 0$ pointwise, the baseline term vanishes exactly, leaving

$$\partial_{\theta_k}\widehat{\mathcal{L}}_{\mathrm{GRPO}}(\theta; Y_{1:G}) = \frac{1}{G}\sum_{m=1}^{G}\widehat{A}_m\mathbf{1}\{Y_m = k\}.$$

Taking expectations and using exchangeability gives

$$\psi_k^{\mathrm{GRPO}}(p) = \mathbb{E}\left[\frac{1}{G}\sum_{m=1}^{G}\widehat{A}_m\mathbf{1}\{Y_m = k\}\right] = \mathbb{E}\big[\widehat{A}_1\mathbf{1}\{Y_1 = k\}\big] = p_k\,\mathbb{E}[\widehat{A}_1 \mid Y_1 = k].$$

Substituting the class-conditional expectations gives the stated drift. □

Theorem B.2 (Correct-correct log-ratio dynamics under exact mean-field GRPO)**.** *Let $a, b \in \mathcal{C}$ and assume $p_a(t), p_b(t) > 0$. Under the flow (14) with the exact GRPO mean field from Proposition B.2,*

$$\frac{d}{dt}\log\frac{p_a(t)}{p_b(t)} = \alpha_C\big(\rho(p(t))\big)\Big(p_a(t) - p_b(t)\Big) - \varepsilon\log\frac{p_a(t)}{p_b(t)}. \tag{19}$$

Equivalently,

$$\frac{d}{dt}\log\frac{p_a(t)}{p_b(t)} = (1-\rho(p(t)))h_G(\rho(p(t)))\,p_b(t)\Big(\frac{p_a(t)}{p_b(t)} - 1\Big) - \varepsilon\log\frac{p_a(t)}{p_b(t)}.$$

In particular, in the unregularized case $\varepsilon = 0$, whenever $0 < \rho(p(t)) < 1$ and $p_a(t) > p_b(t)$, the log-ratio drift is strictly positive.

Proof. Apply the generic identity (15) with

$$\psi_a^{\mathrm{GRPO}} - \psi_b^{\mathrm{GRPO}} = \alpha_C(\rho)(p_a - p_b).$$

The equivalent form uses $\alpha_C = (1-\rho)h_G$. □

Corollary B.2 (Deterministic fixation under exact mean-field GRPO)**.** *Assume $\varepsilon = 0$ and $0 < \rho(0) < 1$. Let*

$$\mathcal{C}_+ := \{c \in \mathcal{C} : p_c(0) > 0\},$$

and assume there exists a unique

$$a \in \mathcal{C}_+ \qquad \text{with} \qquad p_a(0) > p_c(0) \quad \forall c \in \mathcal{C}_+ \setminus \{a\}.$$

Then:

1. *$\rho(t)$ is strictly increasing on $\{0 < \rho < 1\}$ and satisfies $\rho(t) \uparrow 1$.*
2. *Every $c \in \mathcal{C} \setminus \mathcal{C}_+$ remains zero for all t, and for each $c \in \mathcal{C}_+ \setminus \{a\}$,*

$$\frac{p_c(t)}{p_a(t)} \to 0.$$

   Thus the normalized correct-class composition converges to e_a on its support.

3. *Consequently,*

$$p(t) \to e_a.$$

Proof. As in Corollary B.1, any coordinate that starts at zero stays zero, so it suffices to work on the positive support.

Let

$$s_c := \frac{p_c}{\rho} \quad (c \in \mathcal{C}), \qquad r_i := \frac{p_i}{1-\rho} \quad (i \in \mathcal{I},\ \rho < 1),$$

and define

$$S_2 := \sum_{c \in \mathcal{C}} s_c^2, \qquad R_2 := \sum_{i \in \mathcal{I}} r_i^2.$$

Using Proposition B.2 and the induced probability dynamics, one computes

$$\dot{\rho} = h_G(\rho)\, \rho^2 (1-\rho)^2 \big(S_2 + R_2\big). \tag{20}$$

Since $h_G(\rho) > 0$ and $S_2 \geq 1/|\mathcal{C}_+|$, the right-hand side is strictly positive whenever $0 < \rho < 1$. Thus ρ is increasing and has a limit $L \in (0, 1]$. If $L < 1$, continuity and positivity of h_G imply that

$$m_L := \min_{\rho \in [L/2, (1+L)/2]} h_G(\rho) > 0.$$

For all large t we then have

$$\rho(t) \geq \frac{L}{2}, \qquad 1 - \rho(t) \geq \frac{1-L}{2}, \qquad h_G(\rho(t)) \geq m_L, \qquad S_2(t) \geq \frac{1}{|\mathcal{C}_+|},$$

so (20) yields

$$\dot{\rho}(t) \geq m_L \Big(\frac{L}{2}\Big)^2 \Big(\frac{1-L}{2}\Big)^2 \frac{1}{|\mathcal{C}_+|} > 0$$

for all sufficiently large t, contradicting convergence to L. Hence $\rho(t) \uparrow 1$.

Now let $u := 1 - \rho$. Since h_G is continuous on the compact interval $[0, 1]$, there exists $h_{G,\max} < \infty$ with $h_G(\rho) \leq h_{G,\max}$ for all ρ. Also $\rho \leq 1$ and $S_2 + R_2 \leq 2$. Therefore

$$\dot{u} = -h_G(\rho)\rho^2 u^2 (S_2 + R_2) \geq -2h_{G,\max} u^2.$$

It follows that

$$\frac{d}{dt}\frac{1}{u(t)} = -\frac{\dot{u}(t)}{u(t)^2} \leq 2h_{G,\max},$$

so

$$u(t) \geq \frac{1}{u(0)^{-1} + 2h_{G,\max} t}. \tag{21}$$

Hence $\int_0^\infty (1-\rho(t))\,dt = \infty$.

Because $\rho(t) \to 1$ and $h_G(1) = \sqrt{G-1} > 0$, there exists $T < \infty$ and $c_0 > 0$ such that

$$\rho(t) h_G(\rho(t)) \geq c_0 \qquad \text{for all } t \geq T.$$

Therefore the effective time

$$\tau(t) := \int_0^t \rho(s)(1-\rho(s)) h_G(\rho(s))\,ds$$

also diverges.

Now fix $c \in \mathcal{C}_+ \setminus \{a\}$ and set $q_c := p_c/p_a$. By Theorem B.2,

$$\frac{d}{dt} \log q_c = \alpha_C(\rho)(p_c - p_a) = \rho(1-\rho) h_G(\rho)(s_c - s_a).$$

Passing to the effective time $d\tau = \rho(1-\rho) h_G(\rho)\,dt$ gives

$$\frac{d}{d\tau} \log q_c = s_c - s_a = s_a(q_c - 1), \qquad \frac{dq_c}{d\tau} = q_c s_a (q_c - 1).$$

Since $q_c(0) < 1$, uniqueness implies $q_c(\tau) < 1$ for all τ, so q_c is strictly decreasing. As a remains the largest correct trace on the support $\mathcal{C}_+$,

$$s_a(\tau) \geq \frac{1}{|\mathcal{C}_+|}.$$

Also $1 - q_c(\tau) \geq 1 - q_c(0)$. Therefore

$$\frac{d}{d\tau} \log q_c = -s_a(1-q_c) \leq -\frac{1-q_c(0)}{|\mathcal{C}_+|}.$$

Integrating in τ shows $q_c(\tau) \leq q_c(0) e^{-\kappa_c \tau}$ for some $\kappa_c > 0$, hence $q_c(t) \to 0$ because $\tau(t) \to \infty$. Thus the correct-class composition converges to e_a, and since $\rho(t) \to 1$, we conclude $p(t) \to e_a$. □

Interpretation. The exact mean field of outcome-level GRPO is already diversity-collapsing. No neutral-drift argument is needed: the deterministic Euclidean-logit batch-gradient mean field itself amplifies whichever correct trace is currently larger. Finite-batch stochasticity can matter quantitatively, but the basic within-class selection mechanism is deterministic.

B.3 DPO: reference-ratio regression under a one-sided exchangeable-pair surrogate

Context. Direct Preference Optimization (DPO) is trained on a fixed dataset of preferred–dispreferred pairs and optimizes a logistic loss in the reference-relative log-probability margin. To obtain a closed-form dynamical system for pairwise correct-trace ratios, we study a *one-sided* exchangeable-pair surrogate. At each time t:

- every correct trace in $\mathcal{C}$ enters symmetrically on the preferred side, with a common preferred-side marginal weight $\lambda_t^+ > 0$;
- those correct traces do not appear on the dispreferred side; and
- conditional on the identity of the preferred trace, the dispreferred trace is drawn from a common loser environment ν_t^- that does not depend on that identity.

This surrogate isolates preferred-side reference-margin geometry. It should be read as a mechanistic diagnostic, not as a theorem about arbitrary offline preference datasets. A more faithful two-sided treatment is given in Subsection B.4.

Assume the reference policy has full support on the traces under consideration, and define the reference-relative margins

$$m_y(t) := \log p_y(t) - \log p_{\text{base},y}.$$

Proposition B.3 (One-sided DPO surrogate score field)**.** *Under the one-sided exchangeable-pair surrogate described above, the expected score of any correct trace $c \in \mathcal{C}$ is a scalar function of its own reference-relative margin:*

$$\psi_c^{\mathrm{DPO}}(p(t)) = g_t\big(m_c(t)\big),$$

where

$$g_t(m) := \lambda_t^+ \beta_{\mathrm{DPO}}\, \mathbb{E}_{\ell\sim\nu_t^-}\left[\sigma\Big(\beta_{\mathrm{DPO}}\big(m_\ell(t) - m\big)\Big)\right].$$

For each fixed t, the function g_t is C^1 and strictly decreasing:

$$g_t'(m) < 0 \qquad \textit{for all } m \in \mathbb{R}.$$

Proof. For a preferred–dispreferred pair (w, ℓ), the DPO objective is

$$\ell_{w,\ell} = \log\sigma\Big(\beta_{\mathrm{DPO}}\big[(\log p_w - \log p_{\mathrm{base},w}) - (\log p_\ell - \log p_{\mathrm{base},\ell})\big]\Big).$$

Since

$$\log p_w - \log p_\ell = \theta_w - \theta_\ell,$$

the softmax normalizer cancels in the pairwise log-ratio, and differentiation with respect to the trace-level logits gives

$$\nabla_\theta \ell_{w,\ell} = \beta_{\mathrm{DPO}}\,\sigma\Big(\beta_{\mathrm{DPO}}(m_\ell - m_w)\Big)(e_w - e_\ell).$$

Hence a single pair contributes

$$+\beta_{\mathrm{DPO}}\sigma\Big(\beta_{\mathrm{DPO}}(m_\ell - m_w)\Big)$$

to the winner logit and the opposite quantity to the loser logit.

Under the one-sided surrogate, each correct trace $c \in \mathcal{C}$ contributes only through its preferred-side appearances. Those appearances have common marginal weight λ_t^+, and conditional on $w = c$ the loser is drawn from the common environment ν_t^-. Therefore the expected score of c is exactly the displayed function $g_t(m_c(t))$.

Differentiating under the expectation yields

$$g_t'(m) = -\lambda_t^+ \beta_{\mathrm{DPO}}^2 \mathbb{E}_{\ell\sim\nu_t^-}\left[\sigma\Big(\beta_{\mathrm{DPO}}(m_\ell - m)\Big)\Big(1 - \sigma\big(\beta_{\mathrm{DPO}}(m_\ell - m)\big)\Big)\right] < 0,$$

since $\lambda_t^+ > 0$ and $\sigma(u)(1-\sigma(u)) > 0$ for all finite u. □

Theorem B.3 (Exact pairwise correct-correct log-ratio dynamics for the one-sided DPO surrogate)**.** *Let $a, b \in \mathcal{C}$ and assume $p_a(t), p_b(t) > 0$. Define*

$$z_{ab}(t) := \log\frac{p_a(t)}{p_b(t)}, \qquad z_{ab}^{\mathrm{base}} := \log\frac{p_{\mathrm{base},a}}{p_{\mathrm{base},b}}.$$

Then for each t there exists $\xi_{ab}(t)$ between $m_a(t)$ and $m_b(t)$ such that

$$\frac{d}{dt} z_{ab}(t) = g_t'\big(\xi_{ab}(t)\big)\Big(z_{ab}(t) - z_{ab}^{\mathrm{base}}\Big) - \varepsilon z_{ab}(t). \tag{22}$$

Proof. Apply (15) with the one-sided DPO surrogate score field:

$$\dot{z}_{ab} = g_t(m_a) - g_t(m_b) - \varepsilon z_{ab}.$$

For fixed t, the function g_t is C^1, so the mean value theorem yields a point $\xi_{ab}(t)$ between $m_a(t)$ and $m_b(t)$ such that

$$g_t(m_a) - g_t(m_b) = g_t'(\xi_{ab})(m_a - m_b).$$

Finally,

$$m_a - m_b = (\log p_a - \log p_{\mathrm{base},a}) - (\log p_b - \log p_{\mathrm{base},b}) = z_{ab} - z_{ab}^{\mathrm{base}},$$

which proves (22). □

Corollary B.3 (Monotone regression toward the reference ratio)**.** *In the unregularized case $\varepsilon = 0$, the one-sided DPO surrogate always pulls pairwise correct-trace ratios toward the base-policy ratio:*

$$z_{ab}(t) > z_{ab}^{\text{base}} \Longrightarrow \dot{z}_{ab}(t) < 0, \qquad z_{ab}(t) < z_{ab}^{\text{base}} \Longrightarrow \dot{z}_{ab}(t) > 0.$$

Moreover, if all winner and loser margins encountered along the trajectory remain in a compact interval $I \subset \mathbb{R}$ and

$$\lambda_t^+ \geq \underline{\lambda} > 0 \qquad \textit{for all } t,$$

then there exists $\kappa_I > 0$ such that

$$\left|z_{ab}(t) - z_{ab}^{\text{base}}\right| \leq e^{-\kappa_I t}\left|z_{ab}(0) - z_{ab}^{\text{base}}\right|.$$

This compactness hypothesis on margins is automatic on any trimmed simplex on which the reference policy has full support on the traces under consideration.

Proof. When $\varepsilon = 0$, Theorem B.3 gives

$$\frac{d}{dt}\big(z_{ab} - z_{ab}^{\text{base}}\big) = g_t'(\xi_{ab})\big(z_{ab} - z_{ab}^{\text{base}}\big).$$

Since $g_t'(\xi_{ab}) < 0$, the derivative always has the opposite sign from the displacement, proving monotone regression.

Now assume all winner and loser margins lie in a compact interval I. Then the sigmoid arguments

$$\beta_{\text{DPO}}(m_\ell - m)$$

lie in the compact interval

$$J := \beta_{\text{DPO}}(I - I).$$

Hence there exists $c_I > 0$ such that

$$\sigma(u)(1 - \sigma(u)) \geq c_I \qquad \forall u \in J.$$

Therefore

$$-g_t'(\xi_{ab}(t)) \geq \underline{\lambda}\,\beta_{\text{DPO}}^2 c_I =: \kappa_I > 0.$$

Setting $y(t) := z_{ab}(t) - z_{ab}^{\text{base}}$ gives

$$\frac{d}{dt}y(t)^2 = 2g_t'(\xi_{ab}(t))y(t)^2 \leq -2\kappa_I y(t)^2,$$

and Grönwall's inequality yields the stated exponential bound. □

Corollary B.4 (Exact stationary attenuation under entropy)**.** *Let $p^\star$ be any interior stationary point of the one-sided exchangeable-pair surrogate with $\varepsilon > 0$. Let $\nu_\star^-$ be the loser environment evaluated at $p^\star$, and let $g_\star$ denote the corresponding scalar score function. For any two correct traces $a, b \in \mathcal{C}$, there exists $\xi_{ab}^\star$ between $m_a^\star$ and $m_b^\star$ such that*

$$z_{ab}^\star = \frac{a_{ab}^\star}{a_{ab}^\star + \varepsilon}\, z_{ab}^{\text{base}}, \qquad a_{ab}^\star := -g_\star'(\xi_{ab}^\star) > 0.$$

Thus entropy attenuates inherited reference bias but does not create a semantic balancing force.

Proof. At a stationary point, $\dot{z}_{ab} = 0$. Evaluating Theorem B.3 at $p^\star$ gives

$$0 = g_\star'(\xi_{ab}^\star)\big(z_{ab}^\star - z_{ab}^{\text{base}}\big) - \varepsilon z_{ab}^\star.$$

Writing $a_{ab}^\star := -g_\star'(\xi_{ab}^\star) > 0$ and rearranging yields

$$(a_{ab}^\star + \varepsilon)z_{ab}^\star = a_{ab}^\star z_{ab}^{\text{base}},$$

which is the claimed formula. □

B.4 Robustness to replay buffers, clipping, KL tethers, and two-sided pair data

The simplified models above isolate the core within-class geometry of three scalar training families. We now strengthen that analysis in directions that are closer to common practice:

1. replayed accepted data in STaR-style self-training;
2. alternative centered group weightings, local PPO/GRPO clipping, and KL tethers in GRPO-style updates (Schulman et al., 2017; Ouyang et al., 2022); and
3. two-sided pair data in DPO, allowing a trace to appear on either the preferred or the dispreferred side.

The common conclusion is unchanged. These implementation details can rescale the within-class amplification, add regression toward replay-buffer ratios, or add regression toward reference-relative pairwise ratios. What they do *not* add is any force that depends on semantic distances among correct traces.

Lemma B.1 (Pairwise log-ratio identity with KL and entropy)**.** *Let $p(\cdot)$ evolve on the simplex according to*

$$\dot{p}_i = p_i\big(\psi_i(p) - \bar{\psi}(p)\big) - \beta_{\mathrm{KL}}\, p_i \left(\log \frac{p_i}{p_{\mathrm{base},i}} - \sum_j p_j \log \frac{p_j}{p_{\mathrm{base},j}} \right) - \varepsilon p_i\big(\log p_i - \langle \log p \rangle\big), \tag{23}$$

where

$$\bar{\psi}(p) := \sum_j p_j \psi_j(p), \qquad \langle \log p \rangle := \sum_j p_j \log p_j.$$

For any pair of traces a, b with $p_a(t), p_b(t) > 0$, define

$$z_{ab}(t) := \log \frac{p_a(t)}{p_b(t)}, \qquad z_{ab}^{\mathrm{base}} := \log \frac{p_{\mathrm{base},a}}{p_{\mathrm{base},b}}.$$

Then

$$\frac{d}{dt} z_{ab}(t) = \psi_a(p(t)) - \psi_b(p(t)) - \beta_{\mathrm{KL}}\big(z_{ab}(t) - z_{ab}^{\mathrm{base}}\big) - \varepsilon z_{ab}(t). \tag{24}$$

Proof. Divide (23) by p_i and subtract the b equation from the a equation. The common averages $\bar{\psi}(p)$, $\sum_j p_j \log \frac{p_j}{p_{\mathrm{base},j}}$, and $\langle \log p \rangle$ cancel. Since

$$\log \frac{p_a / p_{\mathrm{base},a}}{p_b / p_{\mathrm{base},b}} = z_{ab} - z_{ab}^{\mathrm{base}},$$

the claim follows. □

STaR with replayed accepted data. A common practical variant of STaR-style self-training mixes newly accepted traces with replayed accepted traces from earlier rounds. The next proposition gives the exact pairwise dynamics of that mechanism in the trace-level logit model.

Proposition B.4 (STaR with a replayed accepted-target mixture)**.** *Fix a replay law*

$$r \in \Delta^{|\mathcal{S}_T|-1}, \qquad \mathrm{supp}(r) \subseteq \mathcal{C},$$

and a mixing weight $\xi \in [0, 1]$. Define the mixed accepted target by

$$q_c^{\mathrm{mix}}(p) = (1-\xi)\frac{p_c}{\rho(p)} + \xi r_c \quad (c \in \mathcal{C}), \qquad q_i^{\mathrm{mix}}(p) = 0 \quad (i \in \mathcal{I}),$$

whenever $\rho(p) > 0$. If the model takes an infinitesimal Euclidean logit-ascent step on the log-likelihood against q^{mix}, then

$$\psi_i^{\mathrm{mix}}(p) = q_i^{\mathrm{mix}}(p) - p_i.$$

Hence for any $a, b \in \mathcal{C}$ with $p_a(t), p_b(t) > 0$,

$$\frac{d}{dt} \log \frac{p_a(t)}{p_b(t)} = (1-\xi)\frac{1-\rho(p(t))}{\rho(p(t))}\big(p_a(t) - p_b(t)\big) - \xi\Big(\big(p_a(t) - p_b(t)\big) - (r_a - r_b)\Big) - \varepsilon \log \frac{p_a(t)}{p_b(t)}. \tag{25}$$

Proof. For multinomial logits, gradient ascent on the log-likelihood against target law q^{mix} gives

$$\dot{\theta}_k = q_k^{\mathrm{mix}}(p) - p_k.$$

Thus, for $c \in \mathcal{C}$,

$$\psi_c^{\mathrm{mix}}(p) = (1-\xi)\frac{p_c}{\rho(p)} + \xi r_c - p_c,$$

and for $i \in \mathcal{I}$,

$$\psi_i^{\mathrm{mix}}(p) = 0 - p_i = -p_i.$$

Now apply the generic pairwise identity (15). For $a, b \in \mathcal{C}$,

$$\begin{aligned}\psi_a^{\mathrm{mix}}(p) - \psi_b^{\mathrm{mix}}(p) &= (1-\xi)\frac{p_a - p_b}{\rho} + \xi(r_a - r_b) - (p_a - p_b) \\ &= (1-\xi)\frac{1-\rho}{\rho}(p_a - p_b) - \xi\Big((p_a - p_b) - (r_a - r_b)\Big).\end{aligned}$$

Substituting this into (15) gives (25). □

Remark B.3 (Interpretation for replayed STaR targets)**.** *Equation* (25) *splits the within-correct dynamics into two pieces:*

1. *the original online self-amplification term*

$$(1-\xi)\frac{1-\rho}{\rho}(p_a - p_b),$$

   which favors the currently larger correct trace; and

2. *the replay-regression term*

$$-\xi\Big((p_a - p_b) - (r_a - r_b)\Big),$$

   which pulls probability differences toward stored replay-buffer differences.

Thus replay can slow or partially undo fixation, but only through past frequency ratios. *It still does not create a semantic balancing force.*

If the replay law is itself updated by an exponential moving average of newly accepted traces,

$$\dot{r}_c = \kappa\left(\frac{p_c}{\rho(p)} - r_c\right) \qquad (c \in \mathcal{C}),$$

then the same pairwise identity holds pointwise with time-dependent $r = r(t)$, and

$$\frac{d}{dt}(r_a - r_b) = \kappa\left(\frac{p_a - p_b}{\rho(p)} - (r_a - r_b)\right).$$

So even dynamically updated replay remains a memory term over accepted-frequency ratios rather than a semantic interaction.

Outcome-symmetric group-weighted GRPO. Real GRPO-style implementations vary in how they normalize rewards, whether they use population or sample standard deviations, whether they multiply advantages by composition-dependent coefficients, and whether they add local clipping or KL damping. The next proposition shows that a large class of such variants still induces the same rank-one within-correct geometry.

Proposition B.5 (Outcome-symmetric centered group updates)**.** *Fix group size $G \geq 2$. Let*

$$w_C, w_I : \{0, 1, \ldots, G\} \to \mathbb{R}$$

be two scalar weight functions satisfying the pointwise centeredness condition

$$t\, w_C(t) + (G-t)\, w_I(t) = 0 \qquad \text{for all } t = 0, 1, \ldots, G. \tag{26}$$

For a sampled group $Y_{1:G} \sim p^{\otimes G}$, let

$$T := \sum_{m=1}^{G} \mathbf{1}_{\{Y_m \in \mathcal{C}\}}$$

be its number of correct traces, and assign each sampled trace the weight

$$\widehat{W}_m = \begin{cases} w_C(T), & Y_m \in \mathcal{C}, \\ w_I(T), & Y_m \in \mathcal{I}. \end{cases}$$

Consider the one-step tabular surrogate

$$\widehat{\mathcal{L}}_w(\theta; Y_{1:G}) := \frac{1}{G} \sum_{m=1}^{G} \widehat{W}_m \log p_\theta(Y_m).$$

Then the exact Euclidean-logit mean field of $\widehat{\mathcal{L}}_w$ is

$$\psi_i^w(p) = \begin{cases} p_i \, A_C^w(\rho(p)), & i \in \mathcal{C}, \\ p_i \, A_I^w(\rho(p)), & i \in \mathcal{I}, \end{cases}$$

where

$$A_C^w(\rho) := \mathbb{E}_{S \sim \mathrm{Binom}(G-1,\rho)}\big[w_C(S+1)\big], \tag{27}$$

$$A_I^w(\rho) := \mathbb{E}_{S \sim \mathrm{Binom}(G-1,\rho)}\big[w_I(S)\big]. \tag{28}$$

Moreover,

$$\rho \, A_C^w(\rho) + (1-\rho) \, A_I^w(\rho) = 0 \qquad \forall \rho \in [0,1]. \tag{29}$$

If, in addition,

$$w_C(t) > 0 \quad \textit{for all } t = 1, \dots, G-1, \qquad w_C(G) = 0,$$

then

$$A_C^w(\rho) > 0 \qquad \forall \rho \in (0,1).$$

Proof. For one sampled group,

$$g_k(\theta; Y_{1:G}) := \partial_{\theta_k} \widehat{\mathcal{L}}_w(\theta; Y_{1:G}) = \frac{1}{G} \sum_{m=1}^{G} \widehat{W}_m \big(\mathbf{1}\{Y_m = k\} - p_k\big).$$

By (26),

$$\sum_{m=1}^{G} \widehat{W}_m = T \, w_C(T) + (G - T) \, w_I(T) = 0$$

pointwise, so the baseline term vanishes exactly:

$$g_k(\theta; Y_{1:G}) = \frac{1}{G} \sum_{m=1}^{G} \widehat{W}_m \mathbf{1}\{Y_m = k\}.$$

Taking expectations and using exchangeability within the group gives

$$\psi_k^w(p) := \mathbb{E}[g_k(\theta; Y_{1:G})] = \mathbb{E}[\widehat{W}_1 \mathbf{1}\{Y_1 = k\}] = p_k \, \mathbb{E}[\widehat{W}_1 \mid Y_1 = k].$$

If $k \in \mathcal{C}$, then conditional on $Y_1 = k$, the number of additional correct traces among the remaining $G-1$ samples is

$$S \sim \mathrm{Binom}(G-1, \rho),$$

and the total number of correct traces is $T = S + 1$. Hence

$$\mathbb{E}[\widehat{W}_1 \mid Y_1 = k] = A_C^w(\rho).$$

Similarly, if $k \in \mathcal{I}$, then $T = S$ with the same $S \sim \mathrm{Binom}(G-1, \rho)$, so

$$\mathbb{E}[\widehat{W}_1 \mid Y_1 = k] = A_I^w(\rho).$$

This proves the formula for ψ^w.

For (29), note that $\sum_{m=1}^G \widehat{W}_m = 0$ pointwise implies $\mathbb{E}[\widehat{W}_1] = 0$ by exchangeability, while

$$\mathbb{E}[\widehat{W}_1] = \rho\,\mathbb{E}[\widehat{W}_1 \mid Y_1 \in \mathcal{C}] + (1-\rho)\,\mathbb{E}[\widehat{W}_1 \mid Y_1 \in \mathcal{I}] = \rho A_C^w(\rho) + (1-\rho)A_I^w(\rho).$$

Thus (29) holds.

Finally, if $w_C(t) > 0$ on $t = 1, \ldots, G-1$ and $w_C(G) = 0$, then for every $\rho \in (0,1)$ the binomial law of S assigns positive probability to at least one value $s \in \{0, \ldots, G-2\}$, for which $w_C(s+1) > 0$. Since the remaining terms are nonnegative, $A_C^w(\rho) > 0$. □

Remark B.4 (If centeredness is not exact)**.** *If* (26) *is dropped, define*

$$B^w(\rho) := \frac{1}{G}\,\mathbb{E}_{T\sim\mathrm{Binom}(G,\rho)}\big[T\,w_C(T) + (G-T)\,w_I(T)\big].$$

Then the exact mean field becomes

$$\psi_i^w(p) = \begin{cases} p_i\big(A_C^w(\rho) - B^w(\rho)\big), & i \in \mathcal{C}, \\ p_i\big(A_I^w(\rho) - B^w(\rho)\big), & i \in \mathcal{I}. \end{cases}$$

So even without perfect centeredness, the correct-correct geometry remains rank one:

$$\psi_a^w(p) - \psi_b^w(p) = \big(A_C^w(\rho) - B^w(\rho)\big)\,(p_a - p_b).$$

The modification changes only a scalar coefficient; it still does not introduce any semantic coupling among correct traces.

Theorem B.4 (Robust correct-correct log-ratio dynamics for group-weighted GRPO)**.** *Assume the centered setting of Proposition B.5, and consider the augmented probability-space flow* (23) *with score field ψ^w from that proposition. Then for any $a, b \in \mathcal{C}$ with $p_a(t), p_b(t) > 0$,*

$$\frac{d}{dt}\log\frac{p_a(t)}{p_b(t)} = A_C^w\big(\rho(p(t))\big)\big(p_a(t) - p_b(t)\big) - \beta_{\mathrm{KL}}\left(\log\frac{p_a(t)}{p_b(t)} - \log\frac{p_{\mathrm{base},a}}{p_{\mathrm{base},b}}\right) - \varepsilon\log\frac{p_a(t)}{p_b(t)}. \tag{30}$$

In particular, if $A_C^w(\rho) \geq 0$, the task-dependent term still amplifies whichever correct trace is currently larger, while the KL and entropy terms only add reference-ratio regression and damping.

Proof. By Proposition B.5,

$$\psi_a^w(p) - \psi_b^w(p) = A_C^w(\rho)\,(p_a - p_b) \qquad (a, b \in \mathcal{C}).$$

Now apply Lemma B.1. □

Corollary B.5 (Local clipping does not change the first-order GRPO mean field)**.** *Fix $\delta \in (0,1)$ and old logits θ_{old}. For a sampled group $Y_{1:G}$ and centered outcome-symmetric weights $\widehat{W}_m$ as above, define the PPO/GRPO-style clipped surrogate*

$$\widehat{\mathcal{L}}_{\mathrm{clip}}(\theta; Y_{1:G}) := \frac{1}{G}\sum_{m=1}^G \min\Big(r_m(\theta)\widehat{W}_m,\ \mathrm{clip}(r_m(\theta), 1-\delta, 1+\delta)\widehat{W}_m\Big),$$

where

$$r_m(\theta) := \frac{p_\theta(Y_m)}{p_{\theta_{\mathrm{old}}}(Y_m)}.$$

Then

$$\nabla_\theta\widehat{\mathcal{L}}_{\mathrm{clip}}(\theta_{\mathrm{old}}; Y_{1:G}) = \nabla_\theta\widehat{\mathcal{L}}_w(\theta_{\mathrm{old}}; Y_{1:G}).$$

Consequently, the exact first-order local mean field *at the sampling policy is unchanged by clipping, and all conclusions of Proposition B.5 and Theorem B.4 apply verbatim to the local clipped field.*

Proof. For each sampled trace Y_m, we have

$$r_m(\theta_{\text{old}}) = 1.$$

Since 1 lies in the interior of the clipping interval $[1-\delta, 1+\delta]$, continuity of $r_m(\theta)$ implies that for θ in a neighborhood of θ_{old},

$$\text{clip}(r_m(\theta), 1-\delta, 1+\delta) = r_m(\theta).$$

Thus, in that neighborhood, the two arguments of the min coincide and equal $r_m(\theta)\widehat{W}_m$. Hence the mth summand of $\widehat{\mathcal{L}}_{\text{clip}}$ agrees locally with the unclipped summand

$$r_m(\theta)\widehat{W}_m.$$

Differentiating at $\theta = \theta_{\text{old}}$ therefore yields the same gradient as the unclipped surrogate. Summing over m gives the claim. □

Remark B.5 (What the GRPO robustness result covers)**.** *Proposition B.5 and Corollary B.5 cover, within the one-step tabular trace model:*

1. *population-SD and sample-SD normalization;*
2. *any common positive rescaling of the centered within-group weights;*
3. *composition-dependent centered reweightings of correct and incorrect samples; and*
4. *the first-order local field of PPO/GRPO clipping.*

Multiple optimizer steps away from the sampling policy introduce higher-order corrections in the policy displacement, but they do not create a new semantic interaction term among correct traces.

DPO on a general pair graph. The original DPO subsection studied a one-sided exchangeable surrogate chosen to isolate preferred-side reference-margin geometry. The next proposition gives the exact expected trace-level score under an arbitrary ordered-pair law; the theorem after it then shows that the earlier monotone-regression conclusion survives in a more faithful two-sided exchangeable block model.

Proposition B.6 (Exact DPO score under an arbitrary ordered-pair law)**.** *Let Π_t be any probability law on ordered pairs*

$$(w, \ell) \in \mathcal{S}_T \times \mathcal{S}_T,$$

where w is the preferred trace and ℓ the dispreferred trace. Define the winner and loser marginals

$$\pi_t^+(y) := \sum_{\ell} \Pi_t(y, \ell), \qquad \pi_t^-(y) := \sum_{w} \Pi_t(w, y),$$

and, whenever the corresponding marginal is positive, the conditional loser and winner environments

$$\nu_{t,y}^-(\ell) := \frac{\Pi_t(y, \ell)}{\pi_t^+(y)}, \qquad \nu_{t,y}^+(w) := \frac{\Pi_t(w, y)}{\pi_t^-(y)}.$$

Then the exact expected Euclidean-logit DPO score of trace y is

$$\psi_y^{\text{DPO}}(p(t)) = \beta_{\text{DPO}}\, \pi_t^+(y)\, \mathbb{E}_{\ell \sim \nu_{t,y}^-}\Big[\sigma\Big(\beta_{\text{DPO}}\big(m_\ell(t) - m_y(t)\big)\Big)\Big] - \beta_{\text{DPO}}\, \pi_t^-(y)\, \mathbb{E}_{w \sim \nu_{t,y}^+}\Big[\sigma\Big(\beta_{\text{DPO}}\big(m_y(t) - m_w(t)\big)\Big)\Big]. \tag{31}$$

Proof. For one pair (w, ℓ), the DPO objective is

$$\ell_{w,\ell} = \log \sigma\Big(\beta_{\text{DPO}}\big[(\log p_w - \log p_{\text{base},w}) - (\log p_\ell - \log p_{\text{base},\ell})\big]\Big).$$

Since

$$\log p_w - \log p_\ell = \theta_w - \theta_\ell,$$

the softmax normalizer cancels in the pairwise difference, and differentiation with respect to the trace-level logits gives

$$\nabla_\theta \ell_{w,\ell} = \beta_{\mathrm{DPO}}\, \sigma\Big(\beta_{\mathrm{DPO}}(m_\ell - m_w)\Big)\,(e_w - e_\ell).$$

Taking the yth coordinate and averaging over $(w,\ell) \sim \Pi_t$ yields

$$\psi_y^{\mathrm{DPO}}(p(t)) = \beta_{\mathrm{DPO}} \sum_\ell \Pi_t(y,\ell)\sigma\Big(\beta_{\mathrm{DPO}}(m_\ell - m_y)\Big) - \beta_{\mathrm{DPO}} \sum_w \Pi_t(w,y)\sigma\Big(\beta_{\mathrm{DPO}}(m_y - m_w)\Big).$$

Grouping terms by the marginals $\pi_t^\pm$ gives (31). □

Remark B.6 (Pair-graph locality)**.** *Equation* (31) *is the exact trace-level score for arbitrary pair data. It shows that DPO is* local to the comparison graph*: the update on trace y depends only on the ordered pairs in which y participates and on the scalar reference-relative margins of the traces connected to y through those pairs.*

What does not *appear is any term involving semantic distances among correct traces. More faithful pair data can encode comparison-specific biases through the pair graph, but they still do not create the semantic repulsion mechanism supplied by the creativity kernel.*

Theorem B.5 (Two-sided exchangeable correct block in DPO)**.** *Let $\mathcal{C}_{\mathrm{sym}} \subseteq \mathcal{C}$ be a set of correct traces. Assume that for each time t there exist nonnegative scalars λ_t^+, λ_t^- and probability laws ν_t^-, ν_t^+ on $\mathcal{S}_T$ such that for every $c \in \mathcal{C}_{\mathrm{sym}}$,*

$$\pi_t^+(c) = \lambda_t^+, \qquad \nu_{t,c}^- = \nu_t^-,$$

and

$$\pi_t^-(c) = \lambda_t^-, \qquad \nu_{t,c}^+ = \nu_t^+.$$

In words: every trace in $\mathcal{C}_{\mathrm{sym}}$ has the same preferred-side marginal and common conditional loser environment, and the same dispreferred-side marginal and common conditional winner environment.

Then there is a scalar function $f_t : \mathbb{R} \to \mathbb{R}$ such that for every $c \in \mathcal{C}_{\mathrm{sym}}$,

$$\psi_c^{\mathrm{DPO}}(p(t)) = f_t(m_c(t)),$$

namely

$$f_t(m) := \lambda_t^+ \beta_{\mathrm{DPO}}\, \mathbb{E}_{\ell\sim\nu_t^-}\left[\sigma\Big(\beta_{\mathrm{DPO}}(m_\ell(t) - m)\Big)\right] - \lambda_t^- \beta_{\mathrm{DPO}}\, \mathbb{E}_{w\sim\nu_t^+}\left[\sigma\Big(\beta_{\mathrm{DPO}}(m - m_w(t))\Big)\right]. \tag{32}$$

For each fixed t, the function f_t is C^1 and strictly decreasing whenever $\lambda_t^+ + \lambda_t^- > 0$:

$$\begin{aligned} f_t'(m) = &-\lambda_t^+ \beta_{\mathrm{DPO}}^2 \mathbb{E}_{\ell\sim\nu_t^-}\left[\sigma\Big(\beta_{\mathrm{DPO}}(m_\ell - m)\Big)\Big(1 - \sigma\big(\beta_{\mathrm{DPO}}(m_\ell - m)\big)\Big)\right] \\ &- \lambda_t^- \beta_{\mathrm{DPO}}^2 \mathbb{E}_{w\sim\nu_t^+}\left[\sigma\Big(\beta_{\mathrm{DPO}}(m - m_w)\Big)\Big(1 - \sigma\big(\beta_{\mathrm{DPO}}(m - m_w)\big)\Big)\right] < 0. \end{aligned} \tag{33}$$

Consequently, for any $a, b \in \mathcal{C}_{\mathrm{sym}}$ with $p_a(t), p_b(t) > 0$, there exists $\xi_{ab}(t)$ between $m_a(t)$ and $m_b(t)$ such that

$$\frac{d}{dt} z_{ab}(t) = f_t'(\xi_{ab}(t))\Big(z_{ab}(t) - z_{ab}^{\mathrm{base}}\Big) - \varepsilon z_{ab}(t). \tag{34}$$

In particular, in the unregularized case $\varepsilon = 0$,

$$z_{ab}(t) > z_{ab}^{\mathrm{base}} \Longrightarrow \dot{z}_{ab}(t) < 0, \qquad z_{ab}(t) < z_{ab}^{\mathrm{base}} \Longrightarrow \dot{z}_{ab}(t) > 0.$$

Proof. By Proposition B.6 and the exchangeability assumptions,

$$\psi_c^{\mathrm{DPO}}(p(t)) = \lambda_t^+ \beta_{\mathrm{DPO}}\, \mathbb{E}_{\ell\sim\nu_t^-}\left[\sigma\Big(\beta_{\mathrm{DPO}}(m_\ell(t) - m_c(t))\Big)\right] - \lambda_t^- \beta_{\mathrm{DPO}}\, \mathbb{E}_{w\sim\nu_t^+}\left[\sigma\Big(\beta_{\mathrm{DPO}}(m_c(t) - m_w(t))\Big)\right],$$

which is exactly (32) with $m = m_c(t)$.

Differentiating under the expectation gives (33). Every term is nonpositive, and if $\lambda_t^+ + \lambda_t^- > 0$, at least one term is strictly negative because $\sigma(u)(1 - \sigma(u)) > 0$ for all finite u.

Now fix $a, b \in \mathcal{C}_{\mathrm{sym}}$. By (15),

$$\dot{z}_{ab}(t) = \psi_a^{\mathrm{DPO}}(p(t)) - \psi_b^{\mathrm{DPO}}(p(t)) - \varepsilon z_{ab}(t) = f_t(m_a(t)) - f_t(m_b(t)) - \varepsilon z_{ab}(t).$$

Since f_t is C^1, the mean value theorem yields a point $\xi_{ab}(t)$ between $m_a(t)$ and $m_b(t)$ such that

$$f_t(m_a(t)) - f_t(m_b(t)) = f_t'(\xi_{ab}(t))\big(m_a(t) - m_b(t)\big).$$

Finally,

$$m_a - m_b = (\log p_a - \log p_{\mathrm{base},a}) - (\log p_b - \log p_{\mathrm{base},b}) = z_{ab} - z_{ab}^{\mathrm{base}},$$

which proves (34). The final sign statement follows from $f_t'(\xi_{ab}(t)) < 0$ when $\varepsilon = 0$. □

Corollary B.6 (Compact-margin exponential regression for two-sided exchangeable DPO)**.** *Assume the setting of Theorem B.5 with $\varepsilon = 0$, and suppose there exist a compact interval $I \subset \mathbb{R}$ and a constant $\underline{\lambda} > 0$ such that:*

1. *every margin $m_c(t)$ for $c \in \mathcal{C}_{\mathrm{sym}}$ lies in I along the trajectory;*
2. *every margin $m_y(t)$ of every trace y in the supports of ν_t^- and ν_t^+ also lies in I along the trajectory; and*
3. *$\lambda_t^+ + \lambda_t^- \geq \underline{\lambda}$ for all t.*

Then there exists $\kappa_I > 0$ such that for all $a, b \in \mathcal{C}_{\mathrm{sym}}$,

$$\left|z_{ab}(t) - z_{ab}^{\mathrm{base}}\right| \leq e^{-\kappa_I t}\left|z_{ab}(0) - z_{ab}^{\mathrm{base}}\right|.$$

Proof. Since all relevant margins stay in I, every sigmoid argument appearing in (33) lies in the compact interval

$$J := \beta_{\mathrm{DPO}}(I - I).$$

Hence there exists $c_I > 0$ such that

$$\sigma(u)(1 - \sigma(u)) \geq c_I \qquad \forall u \in J.$$

Therefore (33) yields

$$-f_t'(m) \geq (\lambda_t^+ + \lambda_t^-)\beta_{\mathrm{DPO}}^2 c_I \geq \underline{\lambda}\,\beta_{\mathrm{DPO}}^2 c_I =: \kappa_I \qquad \forall m \in I.$$

Setting

$$y(t) := z_{ab}(t) - z_{ab}^{\mathrm{base}},$$

Theorem B.5 gives

$$\dot{y}(t) = f_t'(\xi_{ab}(t))\, y(t),$$

so

$$\frac{d}{dt}y(t)^2 = 2f_t'(\xi_{ab}(t))\, y(t)^2 \leq -2\kappa_I y(t)^2.$$

Grönwall's inequality yields the claim. □

Remark B.7 (How the robust DPO theorem relates to the one-sided surrogate)**.** *The one-sided exchangeable-pair surrogate of Subsection B.3 is the special case of Theorem B.5 in which*

$$\lambda_t^- = 0.$$

So the more faithful two-sided theorem strictly extends the earlier one-sided analysis: it allows a trace to appear on either side of the pair while preserving the same qualitative conclusion of pairwise regression in reference-relative margin coordinates.

Remark B.8 (What scalar objectives do and do not do, after the robustness layer)**.** *This appendix now identifies the same structural limitation across both the simplified baselines and their more faithful robust variants.*

1. *STaR-style rejection fine-tuning amplifies whichever correct trace is currently most common among newly accepted samples. Replayed accepted data can add regression toward replay-buffer ratios, but only through stored frequencies.*

2. *Outcome-level GRPO, and more generally centered outcome-symmetric group-weighted score-function batch updates, induce a rank-one within-correct mean field of the form*

$$\dot{z}_{ab} \propto p_a - p_b,$$

possibly with additional KL or entropy damping. Alternative normalizations and local clipping change coefficients, not the absence of semantic coupling.

3. *DPO on a general pair graph is pair-graph local reference regularization. Under exchangeable correct blocks it regresses pairwise correct-trace ratios toward reference-relative values; more general pair graphs can encode comparison-specific biases, but still do not create semantic repulsion among correct traces.*

In all of these cases, scalar mechanisms can rescale, damp, or tether pairwise ratios, but they do not reward semantically distinct correct traces for being distinct.

Conclusion. Theorem 4.1 follows at the level proved in this appendix because the training signals analyzed here never contain an explicit relational term over correct traces. Replay buffers contribute only memory over past frequency ratios; clipping and KL tethers contribute only local damping or reference regression; pairwise preference data contributes only comparison-graph local regularization in scalar margin coordinates. None of these mechanisms supplies the missing *semantic repulsion* that would keep a portfolio of genuinely different correct strategies alive. That missing geometric ingredient is precisely what motivates the creativity kernel introduced in Appendix C.

C ESCAPING COLLAPSE VIA THE CREATIVITY KERNEL

The scalar-only analyses of Appendix B show a structural limitation of objectives that depend only on a one-dimensional notion of quality: they can improve correctness while still collapsing onto a narrow subset of semantically redundant solutions. The missing ingredient is *relational geometry*. To preserve a portfolio of genuinely different correct strategies, the objective must distinguish between "many strings that say the same thing" and "many strings that solve the task in different ways."

In DCR, that additional structure enters through a gated creativity kernel. Throughout this appendix we work in the binary-utility regime

$$U_i = \mathbf{1}_{\{i \in \mathcal{C}\}},$$

where both the correct set $\mathcal{C}$ and the incorrect set $\mathcal{I}$ are nonempty. To isolate the interaction between correctness and diversity, we set

$$\beta_{\mathrm{KL}} = 0.$$

Accordingly, the effective entropy coefficient is

$$A := \varepsilon + \lambda\alpha > 0,$$

and the regularized trace-level objective from Appendix A specializes to

$$\widetilde{J}(p) = \rho(p) - \lambda\beta\, p^\top K_{\mathrm{eff}} p + AH[p], \qquad \rho(p) := \sum_{c \in \mathcal{C}} p_c,$$

with score field

$$\phi_i(p) := U_i - 2\lambda\beta(K_{\mathrm{eff}}p)_i.$$

The corresponding Shahshahani gradient flow is therefore

$$\dot{p}_i = p_i\big(\phi_i(p) - \bar{\phi}(p)\big) - Ap_i\big(\log p_i - \langle \log p \rangle\big),$$

where

$$\bar{\phi}(p) := \sum_j p_j \phi_j(p), \qquad \langle \log p \rangle := \sum_j p_j \log p_j.$$

This appendix has two goals. First, we characterize the unique equilibrium exactly and quantify when incorrect traces are strongly suppressed. Second, we show how the creativity kernel acts at the level of *semantic strategies*: it repels distinct correct concepts while entropy homogenizes redundant surface forms.

C.1 The equilibrium geometry: correctness versus redundancy

A generic similarity kernel over all traces is unsafe: it can reward the model for finding "diverse ways of being wrong." The first design principle of DCR is therefore to apply the kernel only to verified-correct traces.

Definition C.1 (Gated effective kernel). Let $R \in \mathbb{R}^{S\times S}$ be the diagonal verifier matrix

$$R_{ii} = \mathbf{1}_{\{i\in\mathcal{C}\}}.$$

Given a symmetric positive semidefinite semantic similarity matrix K_{sem}, define the effective creativity kernel by

$$K_{\text{eff}} := RK_{\text{sem}}R \succeq 0.$$

Equivalently, every row and column corresponding to an incorrect trace is zero. In particular,

$$(K_{\text{eff}}p)_i = 0 \qquad \text{for every } i \in \mathcal{I}.$$

Thus incorrect traces feel *no diversity penalty*. In the no-KL binary-utility regime considered here, their relative masses are governed only by the common entropic barrier and the simplex constraint; with a KL tether, prior terms also enter (Remark C.2.1 below).

Because $K_{\text{eff}} \succeq 0$ and $A > 0$, the objective $\widetilde{J}$ is A-strongly concave on the affine simplex by Appendix A (Lemma A.5). Hence it admits a unique interior maximizer $p^\star \in \operatorname{int}\Delta^{S-1}$. By the global convergence theorem from Appendix A (Theorem A.7), the DCR flow converges to $p^\star$ from every interior initialization.

Theorem C.2 (Exact KKT equilibrium identities). At the unique equilibrium $p^\star$, the balance between correctness and semantic redundancy is characterized by the following identities.

1. **Exact equalization on the incorrect set.** For any $i, j \in \mathcal{I}$,
$$p_i^\star = p_j^\star.$$
2. **Incorrect-to-correct ratio.** For any $i \in \mathcal{I}$ and any $c \in \mathcal{C}$,
$$\frac{p_i^\star}{p_c^\star} = \exp\left(-\frac{1-2\lambda\beta(K_{\text{eff}}p^\star)_c}{A}\right).$$
3. **Correct-to-correct log-ratio.** For any $a, b \in \mathcal{C}$,
$$\log\frac{p_a^\star}{p_b^\star} = \frac{2\lambda\beta}{A}\Big((K_{\text{eff}}p^\star)_b - (K_{\text{eff}}p^\star)_a\Big).$$

Proof. At the interior maximizer, the first-order simplex optimality condition says that the variational derivative is constant across coordinates: there exists $\mu \in \mathbb{R}$ such that

$$U_k - 2\lambda\beta(K_{\text{eff}}p^\star)_k - A(1+\log p_k^\star) = \mu \qquad \forall k \in \{1,\dots,S\}.$$

If $i \in \mathcal{I}$, then $U_i = 0$ and $(K_{\text{eff}}p^\star)_i = 0$, so

$$-A(1+\log p_i^\star) = \mu.$$

This is the same equation for every incorrect coordinate, which proves item 1.

If $c \in \mathcal{C}$, then

$$1 - 2\lambda\beta(K_{\text{eff}}p^\star)_c - A(1+\log p_c^\star) = \mu.$$

Subtracting the incorrect equation from the correct one yields

$$\log\frac{p_i^\star}{p_c^\star} = -\frac{1-2\lambda\beta(K_{\text{eff}}p^\star)_c}{A},$$

which is item 2. Subtracting the KKT equations for two correct traces $a, b \in \mathcal{C}$ eliminates both the multiplier μ and the common utility term 1, giving

$$-A\big(\log p_a^\star - \log p_b^\star\big) = -2\lambda\beta\Big((K_{\text{eff}}p^\star)_a - (K_{\text{eff}}p^\star)_b\Big),$$

which is exactly item 3. □

Remark C.2.1 (Reintroducing a nonuniform base policy). If a KL tether with coefficient $\beta_{\mathrm{KL}} > 0$ and a full-support base policy p_{base} is reintroduced, then the total entropy-like coefficient becomes

$$A_{\mathrm{tot}} := A + \beta_{\mathrm{KL}}.$$

For an incorrect trace $i \in \mathcal{I}$, the KKT equation is then

$$\beta_{\mathrm{KL}} \log p_{\mathrm{base},i} - A_{\mathrm{tot}}(1 + \log p_i^\star) = \mu.$$

Hence the conditional profile on the incorrect set is

$$p_i^\star = \rho_I^\star \frac{(p_{\mathrm{base},i})^{\beta_{\mathrm{KL}}/A_{\mathrm{tot}}}}{\sum_{j\in\mathcal{I}} (p_{\mathrm{base},j})^{\beta_{\mathrm{KL}}/A_{\mathrm{tot}}}}, \qquad \rho_I^\star := \sum_{j\in\mathcal{I}} p_j^\star.$$

Thus a nonuniform base policy breaks exact incorrect equalization: on the incorrect sub-simplex, the equilibrium profile is determined by a power law of the prior. Correct traces still receive prior-dependent terms through their own KKT equations, but only the incorrect block is governed purely by prior regression because the gated kernel vanishes there.

Corollary C.3 (Quantitative safety margin and incorrect-mass bound). For each correct trace $c \in \mathcal{C}$, define its equilibrium correctness margin by

$$m_c^\star := 1 - 2\lambda\beta(K_{\mathrm{eff}} p^\star)_c.$$

Then for every $i \in \mathcal{I}$ and $c \in \mathcal{C}$,

$$\frac{p_i^\star}{p_c^\star} = \exp\left(-\frac{m_c^\star}{A}\right).$$

In particular:

1. $p_i^\star < p_c^\star$ if and only if $m_c^\star > 0$;
2. if $m_c^\star \geq \eta > 0$ for all $c \in \mathcal{C}$, then
$$\frac{p_i^\star}{p_c^\star} \leq e^{-\eta/A} \qquad \forall i \in \mathcal{I},\ \forall c \in \mathcal{C}.$$

Write $M := |\mathcal{C}|$, $N := |\mathcal{I}|$, and

$$m_{\min}^\star := \min_{c\in\mathcal{C}} m_c^\star.$$

If $u^\star$ denotes the common mass assigned to each incorrect trace, then

$$u^\star = \frac{1}{N + \sum_{c\in\mathcal{C}} e^{m_c^\star/A}}, \qquad \rho_I^\star = Nu^\star = \frac{N}{N + \sum_{c\in\mathcal{C}} e^{m_c^\star/A}} \leq \frac{N}{N + Me^{m_{\min}^\star/A}}.$$

Proof. The ratio formula is just Theorem C.2, item 2, rewritten using the definition of $m_c^\star$. Let $u^\star$ denote the common mass of each incorrect trace. Then for every $c \in \mathcal{C}$,

$$p_c^\star = u^\star e^{m_c^\star/A}.$$

Summing over all coordinates gives

$$1 = Nu^\star + \sum_{c\in\mathcal{C}} p_c^\star = u^\star\left(N + \sum_{c\in\mathcal{C}} e^{m_c^\star/A}\right),$$

which yields the formula for $u^\star$ and hence for $\rho_I^\star = Nu^\star$. The upper bound follows from

$$\sum_{c\in\mathcal{C}} e^{m_c^\star/A} \geq Me^{m_{\min}^\star/A}.$$

□

Remark C.4 (Constructive feasibility of a two-level equilibrium). Assume $\lambda\beta > 0$. Let $M := |\mathcal{C}|$ and $N := |\mathcal{I}|$. Fix a target two-level equilibrium of the form

$$p_c^\star = p_C := \frac{1 - N\delta_\star}{M} \quad (c \in \mathcal{C}), \qquad p_i^\star = \delta_\star \quad (i \in \mathcal{I}),$$

with $p_C > \delta_\star$.

Consider the gated rank-one kernel that vanishes outside the correct-correct block and equals $\kappa_{CC}\mathbf{1}\mathbf{1}^\top$ on that block. Then

$$(K_{\text{eff}}p^\star)_c = \kappa_{CC}(1 - N\delta_\star) \qquad \forall c \in \mathcal{C}.$$

Matching the ratio identity from Corollary C.3 gives the exact choice

$$\kappa_{CC} = \frac{1 - A\log(p_C/\delta_\star)}{2\lambda\beta(1 - N\delta_\star)}.$$

Hence whenever

$$A\log\frac{p_C}{\delta_\star} \le 1,$$

this choice satisfies $\kappa_{CC} \ge 0$, so the resulting kernel is PSD and realizes the prescribed two-level equilibrium. This provides a concrete witness that the safety-margin constraints are not vacuous.

C.2 From traces to concepts: rigorous coarse-graining

The KKT identities already show that the creativity kernel changes equilibrium mass allocation across correct traces. The next question is *where* that mass moves. At this level, the relevant redistribution is between semantic strategies rather than raw token strings. We now make that statement precise by aggregating the microscopic trace dynamics onto a partition of the trace space into semantic lumps.

Definition C.5 (Semantic lumps and aggregation operator). Let the trace space be partitioned into L nonempty, disjoint semantic lumps

$$\mathcal{L}_1, \dots, \mathcal{L}_L.$$

Traces within the same lump represent the same underlying strategy up to surface-form variation; traces in different lumps represent genuinely distinct strategies.

Let $\mathbf{M} \in \{0,1\}^{L\times S}$ be the indicator matrix of this partition,

$$\mathbf{M}_{k,i} = \mathbf{1}_{\{i\in\mathcal{L}_k\}},$$

and define the mass of lump k by

$$q_k(p) := (\mathbf{M}p)_k = \sum_{i\in\mathcal{L}_k} p_i.$$

Lemma C.6 (Exact lump-mass ODE). Define

$$m_k(p) := \sum_{i\in\mathcal{L}_k} p_i \log p_i, \qquad \bar{h}(p) := \sum_{j=1}^{S} p_j \log p_j.$$

Under the DCR flow,

$$\dot{q}_k(t) = \sum_{i\in\mathcal{L}_k} p_i(t)\big(\phi_i(p(t)) - \bar{\phi}(p(t))\big) - A\Big(m_k(p(t)) - q_k(p(t))\bar{h}(p(t))\Big),$$

where $\bar{\phi}(p) := \sum_j p_j\phi_j(p)$.

Proof. By linearity,

$$\dot{q}_k = \sum_{i\in\mathcal{L}_k} \dot{p}_i.$$

Substituting the DCR ODE

$$\dot{p}_i = p_i(\phi_i - \bar{\phi}) - Ap_i(\log p_i - \bar{h})$$

and summing over $i \in \mathcal{L}_k$ gives

$$\dot{q}_k = \sum_{i\in\mathcal{L}_k} p_i(\phi_i - \bar{\phi}) - A\left(\sum_{i\in\mathcal{L}_k} p_i \log p_i - \bar{h}\sum_{i\in\mathcal{L}_k} p_i\right),$$

which is exactly the claimed identity. □

To obtain a closed macroscopic ODE, we must express both the selective term and the entropic term in lump-level variables. The selective term closes under structural assumptions on the kernel; the entropic term already admits an exact decomposition.

Theorem C.7 (Exact entropic coarse-graining identity and closure error). For $p \in \operatorname{int}\Delta^{S-1}$, every lump mass $q_k(p)$ is strictly positive. Define the conditional entropy inside lump k by

$$H_k(p) := -\sum_{i\in\mathcal{L}_k} \frac{p_i}{q_k(p)} \log \frac{p_i}{q_k(p)},$$

and the average within-lump entropy by

$$\bar{H}(p) := \sum_{\ell=1}^{L} q_\ell(p) H_\ell(p).$$

Also define the coarse-grained log-average

$$\langle \log q \rangle_q := \sum_{\ell=1}^{L} q_\ell(p) \log q_\ell(p).$$

Then

$$m_k(p) - q_k(p)\bar{h}(p) = q_k(p)\Big(\log q_k(p) - \langle \log q\rangle_q\Big) - q_k(p)\Big(H_k(p) - \bar{H}(p)\Big).$$

Consequently, the entropic contribution to the lump ODE is exactly

$$-Aq_k(p)\Big(\log q_k(p) - \langle \log q\rangle_q\Big) + Aq_k(p)\Big(H_k(p) - \bar{H}(p)\Big).$$

The first term is the closed coarse-grained entropy flow; the second term is the exact closure error.

If

$$s_{\max} := \max_{1\le\ell\le L} |\mathcal{L}_\ell|,$$

then

$$\big|q_k(p)\big(H_k(p) - \bar{H}(p)\big)\big| \le q_k(p)\log s_{\max}.$$

Proof. Let

$$w_i := \frac{p_i}{q_k(p)} \qquad (i \in \mathcal{L}_k).$$

Then $\sum_{i\in\mathcal{L}_k} w_i = 1$, and by definition of H_k,

$$m_k(p) = \sum_{i\in\mathcal{L}_k} p_i \log p_i = q_k(p)\log q_k(p) - q_k(p)H_k(p).$$

Summing over all lumps yields

$$\bar{h}(p) = \sum_{\ell=1}^{L} m_\ell(p) = \sum_{\ell=1}^{L} q_\ell(p)\log q_\ell(p) - \sum_{\ell=1}^{L} q_\ell(p)H_\ell(p) = \langle \log q\rangle_q - \bar{H}(p).$$

Substituting this identity into $m_k - q_k\bar{h}$ gives the claimed decomposition.

For the bound, note that

$$0 \le H_k(p) \le \log|\mathcal{L}_k| \le \log s_{\max}, \qquad 0 \le \bar{H}(p) \le \log s_{\max}.$$

Hence

$$|H_k(p) - \bar{H}(p)| \le \log s_{\max},$$

and multiplying by $q_k(p)$ gives the result. □

Interpretation. Theorem C.7 cleanly separates two effects. The term

$$-Aq_k\big(\log q_k - \langle \log q\rangle_q\big)$$

is exactly the entropy flow one would obtain after coarse-graining to the lump simplex. The only obstruction to perfect closure is the within-lump entropy mismatch $H_k - \bar{H}$. If all lumps carry the same internal syntactic entropy, then the entropic part of the lumped dynamics closes exactly.

We now identify the selective part of the dynamics under a simple but revealing kernel model.

Theorem C.8 (Inter-lump repulsion and intra-lump homogenization). Assume the semantic kernel is block-constant across semantic lumps:

$$K_{\text{sem}}(a,b) = \begin{cases} k_{\text{in}}, & a,b \text{ belong to the same lump,} \\ k_{\text{out}}, & a,b \text{ belong to different lumps,} \end{cases} \qquad k_{\text{in}} > k_{\text{out}} \ge 0.$$

For each lump m, define its correct mass by

$$r_m(p) := \sum_{i \in \mathcal{L}_m \cap \mathcal{C}} p_i, \qquad \rho(p) := \sum_{c \in \mathcal{C}} p_c.$$

Then:

1. if $a \in \mathcal{L}_m \cap \mathcal{C}$ and $b \in \mathcal{L}_n \cap \mathcal{C}$ belong to different lumps ($m \ne n$),

$$\phi_a(p) - \phi_b(p) = -2\lambda\beta\,(k_{\text{in}} - k_{\text{out}})\big(r_m(p) - r_n(p)\big);$$

2. if $a, b \in \mathcal{L}_m \cap \mathcal{C}$ belong to the same lump, then

$$\phi_a(p) = \phi_b(p).$$

   Consequently, along any DCR trajectory $t \mapsto p(t) \in \operatorname{int}\Delta^{S-1}$, if

$$z_{ab}(t) := \log\frac{p_a(t)}{p_b(t)},$$

   then

$$\dot{z}_{ab}(t) = -Az_{ab}(t), \qquad z_{ab}(t) = z_{ab}(0)e^{-At}.$$

Proof. Fix $a \in \mathcal{L}_m \cap \mathcal{C}$. Because the kernel is gated, only correct traces contribute to $(K_{\text{eff}}p)_a$. Hence

$$(K_{\text{eff}}p)_a = k_{\text{in}} r_m(p) + k_{\text{out}}\big(\rho(p) - r_m(p)\big) = (k_{\text{in}} - k_{\text{out}})r_m(p) + k_{\text{out}}\rho(p).$$

The same formula holds for $b \in \mathcal{L}_n \cap \mathcal{C}$. Subtracting gives

$$(K_{\text{eff}}p)_a - (K_{\text{eff}}p)_b = (k_{\text{in}} - k_{\text{out}})\big(r_m(p) - r_n(p)\big),$$

and therefore

$$\phi_a(p) - \phi_b(p) = -2\lambda\beta\Big((K_{\text{eff}}p)_a - (K_{\text{eff}}p)_b\Big),$$

which proves item 1.

If $a, b \in \mathcal{L}_m \cap \mathcal{C}$ lie in the same lump, then the a-th and b-th rows of K_{eff} coincide, so $(K_{\text{eff}}p)_a = (K_{\text{eff}}p)_b$ and therefore $\phi_a(p) = \phi_b(p)$. Along any interior trajectory, the universal log-ratio identity from Appendix A then gives

$$\dot{z}_{ab}(t) = \phi_a(p(t)) - \phi_b(p(t)) - Az_{ab}(t) = -Az_{ab}(t),$$

whose solution is $z_{ab}(t) = z_{ab}(0)e^{-At}$. □

Interpretation. The theorem separates two roles that scalar objectives conflate:

- if one *correct concept* becomes overrepresented relative to another, the creativity kernel lowers its score and pushes mass back toward the minority concept;
- if two correct traces are merely redundant surface forms of the same concept, the kernel exerts no differential force between them, and entropy alone equalizes their relative masses.

This is exactly the qualitative behavior DCR is designed to induce: repulsion across distinct strategies, homogenization within the same strategy.

The block-constant model is useful because it is fully explicit, but the next result shows that the underlying principle is much more general.

Theorem C.9 (Support-function identity). For any kernel matrix K_{eff} (in particular, for the gated PSD kernels considered in DCR) and any two correct traces $a, b \in \mathcal{C}$,

$$\sup_{p \in \Delta^{S-1}} \left|\phi_a(p) - \phi_b(p)\right| = 2\lambda\beta \left\|(K_{\text{eff}})_{a\cdot} - (K_{\text{eff}})_{b\cdot}\right\|_\infty.$$

Proof. For correct traces $a, b \in \mathcal{C}$, the utility terms cancel, so

$$\phi_a(p) - \phi_b(p) = -2\lambda\beta \left((K_{\text{eff}})_{a\cdot} - (K_{\text{eff}})_{b\cdot}\right)^\top p.$$

Thus we are maximizing the absolute value of a linear functional over the simplex. By linear programming, the maximum is attained at a vertex e_j, and therefore

$$\sup_{p \in \Delta^{S-1}} \left|\phi_a(p) - \phi_b(p)\right| = 2\lambda\beta \max_j \left|(K_{\text{eff}})_{aj} - (K_{\text{eff}})_{bj}\right| = 2\lambda\beta \left\|(K_{\text{eff}})_{a\cdot} - (K_{\text{eff}})_{b\cdot}\right\|_\infty.$$

□

Interpretation. Theorem C.9 makes the geometry explicit: two correct traces are pushed apart only to the extent that the kernel judges them to be meaningfully different. If their kernel rows are nearly identical, the diversity pressure between them is small and entropy dominates. That is precisely the separation of roles that scalar objectives lack.

C.3 Practical kernel design and sufficient tuning rules

The results above answer the qualitative question—what kind of equilibrium DCR induces. We now turn to the quantitative question—how to pick the kernel and hyperparameters so that the equilibrium is both safe and useful.

Two points are worth separating clearly.

1. *Existence, uniqueness, and global convergence* already follow from the Appendix A conditions $A > 0$ and $K_{\text{eff}} \succeq 0$.
2. The design rules below are *sufficient conditions* for stronger operational guarantees: explicit margins against incorrect traces, interpretable within-lump versus across-lump behavior, and conservative hyperparameter choices.

All statements in this subsection are given in the no-KL setting $\beta_{\text{KL}} = 0$ used throughout this appendix. If a KL tether is reintroduced with a *uniform* base policy, then A is replaced by $A + \beta_{\text{KL}}$. For a *nonuniform* base policy, the full KKT system must be used; Remark C.2.1 gives the corresponding incorrect-profile geometry.

1. Use kernels that are actually PSD. The convergence and strict-concavity theory from Appendix A requires $K_{\text{eff}} \succeq 0$. Safe choices include:

- the Gram kernel $K_{\text{sem}}(i,j) = e_i^\top e_j$ for embeddings $e_i \in \mathbb{R}^d$ (PSD by construction; if the embeddings are normalized, the entries lie in $[-1,1]$);
- the shifted cosine kernel
$$K_{\text{sem}}(i,j) = \frac{1 + e_i^\top e_j}{2}$$
for normalized embeddings (PSD, with entries in $[0,1]$);
- the Gaussian RBF kernel
$$K_{\text{sem}}(i,j) = \exp\bigl(-\gamma \|e_i - e_j\|_2^2\bigr)$$
(PSD, with entries in $(0,1]$).

See Pillonetto et al. (2022) for the RKHS and PSD-kernel background underlying these constructions. By contrast, entrywise truncations such as $\max(0, e_i^\top e_j)$ are *not* PSD in general and should not be used when the PSD hypothesis is required.

Gating preserves positive semidefiniteness because

$$x^\top K_{\text{eff}} x = x^\top R K_{\text{sem}} R x = (Rx)^\top K_{\text{sem}} (Rx) \geq 0.$$

Gating typically destroys strict positive definiteness by introducing null directions on incorrect traces, but that is harmless here: the entropy term already guarantees strict concavity of the full objective.

2. Gate the kernel: $K_{\text{eff}} = R K_{\text{sem}} R$. This is essential to the intended semantics of the objective. An ungated kernel is still mathematically tractable, but it pushes mass around *all* traces—including incorrect ones. In other words, it can reward the model for being diverse in ways that are semantically wrong.

Gating removes that failure mode. Once K_{eff} is gated, the diversity term acts only on verified-correct traces. The relative profile inside the incorrect block is then shaped only by entropy (and, if present, the KL tether), although the total incorrect mass still couples to the correct block through the simplex constraint and the correctness utility.

3. Impose a uniform margin bound on diversity pressure. Corollary C.3 shows that the relevant quantity is the equilibrium correctness margin

$$m_c^\star = 1 - 2\lambda\beta (K_{\text{eff}} p^\star)_c.$$

To lower-bound this margin uniformly, use the operator norm

$$\|K_{\text{eff}}\|_{1\to\infty} := \sup_{\|x\|_1 = 1} \|K_{\text{eff}} x\|_\infty = \max_{i,j} |(K_{\text{eff}})_{ij}|.$$

Since every $p \in \Delta^{S-1}$ satisfies $\|p\|_1 = 1$,

$$|(K_{\text{eff}} p)_c| \leq \|K_{\text{eff}}\|_{1\to\infty} \qquad \forall c,\ \forall p \in \Delta^{S-1}.$$

Therefore the dimension-free sufficient condition

$$\boxed{\lambda\beta < \frac{1}{2\|K_{\text{eff}}\|_{1\to\infty}}}$$

implies the uniform lower bound

$$m_c^\star \geq \eta_K := 1 - 2\lambda\beta \|K_{\text{eff}}\|_{1\to\infty} > 0 \qquad \forall c \in \mathcal{C}.$$

Applying Corollary C.3 then gives

$$\frac{p_i^\star}{p_c^\star} \leq e^{-\eta_K / A} \qquad \forall i \in \mathcal{I},\ \forall c \in \mathcal{C},$$

and

$$\rho_I^\star \le \frac{N}{N + Me^{\eta_K/A}}, \qquad M := |\mathcal{C}|, \quad N := |\mathcal{I}|.$$

For common kernels with $|(K_{\text{eff}})_{ij}| \le 1$—including normalized dot-product kernels, shifted cosine kernels, and Gaussian RBF kernels—the sufficient rule

$$\lambda\beta < \tfrac{1}{2}$$

is already a universal sufficient condition. For example, $\lambda\beta = 0.25$ leaves a nontrivial global margin buffer $\eta_K \ge 0.5$.

4. Tune A for breadth versus sharpening. Once $\lambda\beta$ is fixed, the entropy coefficient A controls how broadly probability mass is spread once the redundancy penalty is in place.

First, Corollary C.3 shows that smaller A strengthens suppression of incorrect traces:

$$\rho_I^\star \le \frac{N}{N + Me^{\eta_K/A}}.$$

For fixed $\eta_K > 0$, this upper bound decreases monotonically as A decreases.

Second, under the same-lump idealization of Theorem C.8, A is exactly the within-lump homogenization rate:

$$\dot{z}_{ab}(t) = -Az_{ab}(t) \qquad (a, b \text{ in the same correct lump}).$$

Thus larger A makes redundant surface forms equalize faster.

Third, for two correct traces $a, b \in \mathcal{C}$, Theorem C.2 gives

$$\log \frac{p_a^\star}{p_b^\star} = \frac{2\lambda\beta}{A}\Big((K_{\text{eff}}p^\star)_b - (K_{\text{eff}}p^\star)_a\Big).$$

So for a *fixed* kernel-induced score gap, increasing A attenuates the corresponding equilibrium log-ratio. However, because $p^\star$ itself also depends on A, there is no universal monotonic law for every possible inter-lump separation metric.

Practical reading of the trade-off. Smaller A sharpens correctness margins and pushes incorrect mass down more aggressively; larger A spreads mass more broadly and homogenizes redundant syntax more quickly. The kernel still determines *where* mass is pushed apart. In that sense,

$$K_{\text{eff}} \text{ chooses the geometry of diversity,} \qquad \lambda\beta \text{ chooses its strength,} \qquad A \text{ chooses its breadth.}$$

Taken together, this appendix gives the following picture. Under the basic DCR conditions $A > 0$ and $K_{\text{eff}} \succeq 0$, the flow has a unique globally attracting equilibrium. The gating construction ensures that diversity pressure is applied only where it is semantically meaningful—among verified-correct traces. The KKT identities then quantify exactly when incorrect traces are suppressed, while the coarse-graining analysis shows how the kernel redistributes mass across *concepts* rather than merely across strings. Under the sufficient tuning rules above, this equilibrium additionally enjoys explicit safety margins and controlled incorrect mass. When the kernel geometry moreover aligns with semantic strategy structure—for example as in Theorem C.8, or more generally when within-strategy rows are nearly equal and across-strategy rows differ—the same framework also yields repulsion between distinct correct strategies and rapid homogenization of redundant within-strategy surface forms.

D EXACT PARAMETRIC GEOMETRY OF THE LOGIT SPACE

This appendix is narrow and structural. Sections A to C analyze the DCR objective directly on a finite trace simplex. Here we isolate the additional curvature introduced by the *parameterization* that maps token logits to a trace distribution. This lets us separate two distinct sources of conditioning:

- the intrinsic trace-level geometry of the DCR objective, captured by the moduli of the trace-space score field; and

- the extra smoothness cost incurred when that objective is composed with a softmax-based autoregressive model.

The main conclusions are:

1. a single softmax block has dimension-free Jacobian and Hessian bounds with exact sharp constants;
2. composing such blocks autoregressively contributes only an explicit linear factor in the horizon T to the *logit-to-trace parameterization itself*; and
3. any worse dependence on T must come from the trace-level objective moduli or from the neural map from weights to sequence logits, not from repeated softmax composition alone.

D.0. Exact-length sequence space and scope

Let $\mathcal{V}$ be the token vocabulary, with cardinality

$$V := |\mathcal{V}|.$$

For the geometry in this appendix, we work on the exact-length sequence space

$$\mathcal{Y}_T := \mathcal{V}^T, \qquad N := |\mathcal{Y}_T| = V^T.$$

This choice is deliberate: it makes the map from autoregressive token laws to a probability vector on traces literally a map into the full simplex Δ^{N-1}, so identities such as

$$\sum_{y \in \mathcal{Y}_T} p^\Theta(y) = 1, \qquad (Dp^\Theta)^\top \mathbf{1} = 0, \qquad \sum_{y \in \mathcal{Y}_T} \nabla^2_\Theta p^\Theta(y) = 0$$

hold exactly, with no bookkeeping exceptions.

If the main paper represents shorter traces by EOS padding, then $\mathcal{Y}_T$ is the ambient padded sequence space. We do *not* impose absorbing-EOS constraints inside the parameterization here. This makes the parameterization more general. In the common padded-EOS setting, valid-trace restrictions or absorbing-EOS constraints just fix some output coordinates or parameter blocks; the derivative bounds below do not worsen, and we do not rely on that refinement here.

Sections A to C apply to any finite trace space. In the present appendix, the trace space is $\mathcal{Y}_T$, and the DCR objective is viewed as a function

$$\widetilde{J} : \Delta^{N-1} \to \mathbb{R}.$$

We proceed in two stages.

1. We first analyze a *single* softmax block on the centered logit space, proving exact Jacobian and Hessian bounds that do not depend on the vocabulary size.
2. We then compose these blocks autoregressively and combine the resulting parameterization bounds with trace-level moduli for the DCR objective.

D.1 Single-block softmax geometry

Let

$$T_V := \mathbf{1}^\perp = \{\theta \in \mathbb{R}^V : \langle \mathbf{1}, \theta \rangle = 0\}$$

be the centered logit space. The softmax map is

$$p_\theta = \mathrm{softmax}(\theta), \qquad p_\theta(i) = \frac{e^{\theta_i}}{\sum_{j=1}^V e^{\theta_j}}.$$

Because

$$\mathrm{softmax}(\theta + c\mathbf{1}) = \mathrm{softmax}(\theta) \qquad \forall c \in \mathbb{R},$$

the parameterization has a one-dimensional gauge redundancy, and restricting to T_V removes it. On T_V, the softmax map is a real-analytic diffeomorphism onto the relative interior

$$\mathrm{ri}(\Delta^{V-1}) = \{p \in \Delta^{V-1} : p_i > 0 \ \forall i\},$$

with inverse (Amari and Nagaoka, 2007)

$$\theta = \Pi_{\mathbf{1}^\perp} \log p, \qquad \Pi_{\mathbf{1}^\perp} := I - \frac{1}{V}\mathbf{1}\mathbf{1}^\top.$$

Its Jacobian is

$$J_\theta := \nabla_\theta p_\theta = \mathrm{diag}(p_\theta) - p_\theta p_\theta^\top.$$

Theorem D.1 (Loewner bounds and operator norm). For every $v \in T_V$,

$$v^\top J_\theta v = \sum_{i=1}^{V} p_\theta(i) v_i^2 - \Big(\sum_{i=1}^{V} p_\theta(i) v_i\Big)^2 = \mathrm{Var}_{i\sim p_\theta}(v_i).$$

If

$$p_{\min} := \min_i p_\theta(i),$$

then on the tangent space T_V,

$$\boxed{p_{\min} I_{T_V} \preceq J_\theta|_{T_V} \preceq \frac{1}{2} I_{T_V}, \qquad \|J_\theta|_{T_V}\|_{\mathrm{op}} \le \frac{1}{2}.}$$

Proof. The quadratic-form identity is immediate from the definition of J_θ:

$$v^\top J_\theta v = \sum_{i=1}^{V} p_\theta(i) v_i^2 - \Big(\sum_{i=1}^{V} p_\theta(i) v_i\Big)^2.$$

This is exactly the variance of the scalar random variable $i \mapsto v_i$ under the law p_θ.

For the upper bound, Popoviciu's inequality (see Lim and McCann, 2022, for discussion and extensions) gives

$$\mathrm{Var}_{p_\theta}(v_i) \le \frac{1}{4}\big(\max_i v_i - \min_i v_i\big)^2.$$

Set

$$M := \max_i v_i, \qquad m := \min_i v_i.$$

Because $v \in T_V$, its coordinates sum to zero, so $M \ge 0 \ge m$. Therefore

$$\|v\|_2^2 \ge M^2 + m^2 \ge \frac{1}{2}(M - m)^2,$$

and hence

$$v^\top J_\theta v = \mathrm{Var}_{p_\theta}(v_i) \le \frac{1}{4}(M - m)^2 \le \frac{1}{2}\|v\|_2^2.$$

This proves

$$J_\theta|_{T_V} \preceq \frac{1}{2} I_{T_V}.$$

For the lower bound, write

$$p_\theta = p_{\min}\mathbf{1} + q, \qquad q_i \ge 0.$$

Since $\mathbf{1}^\top v = 0$ for $v \in T_V$,

$$\sum_{i=1}^{V} p_\theta(i) v_i = q^\top v.$$

Therefore

$$v^\top J_\theta v = p_{\min}\|v\|_2^2 + \sum_{i=1}^{V} q_i v_i^2 - (q^\top v)^2.$$

By Cauchy–Schwarz,

$$(q^\top v)^2 \le \Big(\sum_i q_i\Big)\Big(\sum_i q_i v_i^2\Big) \le \sum_i q_i v_i^2,$$

because $\sum_i q_i = 1 - V p_{\min} \le 1$. Hence the bracketed term is nonnegative, and

$$v^\top J_\theta v \ge p_{\min}\|v\|_2^2.$$

This proves

$$p_{\min} I_{T_V} \preceq J_\theta|_{T_V}.$$

The operator-norm bound is the upper Loewner bound rewritten. □

Remark D.1. *The constant $1/2$ is sharp. It is attained when $V = 2$ at the uniform binary law, and for $V > 2$ it is approached by centered logits whose softmax mass concentrates on two coordinates with probabilities $1/2$ and $1/2$.*

D.2 Exact Hessian suprema for one softmax block

For $k, \ell \in \{1, \dots, V\}$, define the second-derivative slice

$$H_{k\ell}(\theta) \in \mathbb{R}^V, \qquad [H_{k\ell}(\theta)]_i := \partial_{\theta_\ell}\partial_{\theta_k} p_\theta(i).$$

Thus $H_{k\ell}(\theta)$ is the output-direction vector obtained by differentiating column k of the Jacobian with respect to coordinate ℓ.

Theorem D.2 (Exact dimension-free Hessian suprema). For every $V \ge 2$,

$$\boxed{\sup_{\theta,k,\ell} \|H_{k\ell}(\theta)\|_1 = \frac{1}{3\sqrt{3}}, \qquad \sup_{\theta,k,\ell} \|H_{k\ell}(\theta)\|_2 = \frac{1}{\sqrt{54}}.}$$

Proof. Differentiating $J_{ik}(\theta) = p_i(\delta_{ik} - p_k)$ gives

$$[H_{k\ell}(\theta)]_i = p_i\Big[(\delta_{i\ell} - p_\ell)(\delta_{ik} - p_k) - p_k(\delta_{k\ell} - p_\ell)\Big].$$

This expression is equivariant under simultaneous permutation of coordinates, so the norms depend only on whether $k = \ell$ or $k \ne \ell$. It therefore suffices to analyze one diagonal slice and one off-diagonal slice.

Step 1: diagonal slice. Take $(k, \ell) = (1, 1)$ and write $a := p_1$. Then

$$[H_{11}]_1 = a(1-a)(1-2a), \qquad [H_{11}]_j = -a(1-2a)p_j \quad (j \ne 1).$$

Hence

$$\|H_{11}\|_1 = |a(1-a)(1-2a)| + a|1-2a|\sum_{j\ne 1} p_j = 2|a(1-a)(1-2a)|.$$

By symmetry it suffices to maximize this on $a \in [0, 1/2]$, where it becomes

$$f(a) = 2a(1-a)(1-2a) = 2(a - 3a^2 + 2a^3).$$

Its derivative is

$$f'(a) = 2(1 - 6a + 6a^2),$$

which vanishes at

$$a_\star = \frac{1}{2} \pm \frac{1}{2\sqrt{3}}.$$

The maximizer in $[0, 1/2]$ is $a_\star = \frac{1}{2} - \frac{1}{2\sqrt{3}}$, and substitution yields

$$\sup_\theta \|H_{11}(\theta)\|_1 = \frac{1}{3\sqrt{3}}.$$

Step 2: off-diagonal slice. Take $(k, \ell) = (1, 2)$ and write

$$a := p_1, \qquad b := p_2, \qquad r := 1 - a - b.$$

A direct computation gives

$$H_{12} = ab\,(2a - 1,\ 2b - 1,\ 2p_3, \ldots, 2p_V),$$

so

$$\|H_{12}\|_1 = ab\Big(|2a - 1| + |2b - 1| + 2r\Big).$$

There are three cases.

If $a, b \le 1/2$, then

$$\|H_{12}\|_1 = ab\big((1 - 2a) + (1 - 2b) + 2r\big) = 4abr \le \frac{4}{27} < \frac{1}{3\sqrt{3}}.$$

If $a \ge 1/2 \ge b$, then

$$\|H_{12}\|_1 = ab\big((2a - 1) + (1 - 2b) + 2r\big) = 2ab(1 - 2b).$$

Since $a \le 1 - b$,

$$\|H_{12}\|_1 \le 2b(1 - b)(1 - 2b) \le \frac{1}{3\sqrt{3}},$$

by the same one-variable maximization as in the diagonal case. The case $b \ge 1/2 \ge a$ is symmetric. Therefore

$$\sup_{\theta,k,\ell} \|H_{k\ell}(\theta)\|_1 = \frac{1}{3\sqrt{3}}.$$

Step 3: the ℓ_2 supremum. Because $\sum_i p_\theta(i) \equiv 1$, every second-derivative slice has zero sum:

$$\sum_{i=1}^{V} [H_{k\ell}(\theta)]_i = 0.$$

If $x \in \mathbb{R}^V$ satisfies $\sum_i x_i = 0$, define

$$P := \sum_{x_i > 0} x_i = \sum_{x_i < 0} (-x_i) = \frac{1}{2}\|x\|_1.$$

Then

$$\|x\|_2^2 = \sum_{x_i > 0} x_i^2 + \sum_{x_i < 0} x_i^2 \le P^2 + P^2 = \frac{1}{2}\|x\|_1^2,$$

so

$$\|x\|_2 \le \frac{1}{\sqrt{2}}\|x\|_1.$$

Applying this to $x = H_{k\ell}(\theta)$ gives

$$\|H_{k\ell}(\theta)\|_2 \le \frac{1}{\sqrt{2}}\|H_{k\ell}(\theta)\|_1 \le \frac{1}{\sqrt{2}} \cdot \frac{1}{3\sqrt{3}} = \frac{1}{\sqrt{54}}.$$

This bound is sharp: when $V = 2$, the diagonal family H_{11} has the form $(u, -u)$, so equality holds in $\|x\|_2 = \|x\|_1/\sqrt{2}$ at the maximizing value of a. For $V > 2$, the same value is approached by letting the residual mass on the remaining coordinates tend to zero. □

Corollary D.3 (Global Lipschitz continuity of the softmax Jacobian). For all $\theta_1, \theta_2 \in T_V$,

$$\boxed{\|J_{\theta_2} - J_{\theta_1}\|_{\mathrm{op}} \le \frac{1}{3\sqrt{3}}\|\theta_2 - \theta_1\|_1 \le \frac{\sqrt{V}}{3\sqrt{3}}\|\theta_2 - \theta_1\|_2.}$$

Proof. Let $h \in \mathbb{R}^V$. The directional derivative $DJ_\theta[h]$ is the matrix whose kth column is

$$(DJ_\theta[h])_{\cdot k} = \sum_{\ell=1}^{V} h_\ell H_{k\ell}(\theta).$$

Hence, by Theorem D.2,

$$\|(DJ_\theta[h])_{\cdot k}\|_1 \le \sum_{\ell=1}^{V} |h_\ell| \, \|H_{k\ell}(\theta)\|_1 \le \frac{1}{3\sqrt{3}}\|h\|_1.$$

Therefore

$$\|DJ_\theta[h]\|_1 \le \frac{1}{3\sqrt{3}}\|h\|_1.$$

Since $DJ_\theta[h]$ is symmetric, its spectral norm is bounded by its matrix 1-norm, so

$$\|DJ_\theta[h]\|_{\mathrm{op}} \le \frac{1}{3\sqrt{3}}\|h\|_1.$$

Integrating along the line segment from θ_1 to θ_2 yields

$$J_{\theta_2} - J_{\theta_1} = \int_0^1 DJ_{\theta_1 + s(\theta_2 - \theta_1)}[\theta_2 - \theta_1] \, \mathrm{d}s,$$

and thus

$$\|J_{\theta_2} - J_{\theta_1}\|_{\mathrm{op}} \le \frac{1}{3\sqrt{3}}\|\theta_2 - \theta_1\|_1.$$

The second inequality is the standard bound $\|x\|_1 \le \sqrt{V}\|x\|_2$. □

D.3 The smooth active regime and implementation clipping

To keep all derivatives classical, we work on a trimmed token-level regime. Fix

$$\underline{\delta} \in (0, 1/V),$$

and define the active set

$$\Theta_{\underline{\delta}} := \Big\{\theta \in T_V : \min_i \mathrm{softmax}(\theta)_i \ge \underline{\delta}\Big\}.$$

For each coordinate,

$$\log \mathrm{softmax}_i(\theta) = \theta_i - \log\Big(\sum_j e^{\theta_j}\Big)$$

is concave in θ, so each superlevel set $\{\theta : \mathrm{softmax}_i(\theta) \ge \underline{\delta}\}$ is convex. Hence $\Theta_{\underline{\delta}}$ is convex.

On this active region, the centered logit lift is exactly the identity. Indeed, if $p_\theta = \mathrm{softmax}(\theta)$, then

$$\log p_\theta = \theta - \log\Big(\sum_j e^{\theta_j}\Big)\mathbf{1},$$

so projecting onto $\mathbf{1}^\perp$ gives

$$\Pi_{\mathbf{1}^\perp} \log p_\theta = \theta.$$

Thus all derivatives used below are ordinary C^2 derivatives; no generalized or nonsmooth calculus is needed.

Implementation note (hard clipping is separate from the formal theory). In code, one may still use a clip-and-renormalize safeguard

$$\mathcal{C}_{\delta_\star}(p) := \frac{\max(p, \delta_\star)}{\|\max(p, \delta_\star)\|_1}.$$

If one wants the final renormalized vector to satisfy

$$\min_i \left(\mathcal{C}_{\delta_\star}(p)\right)_i \geq \underline{\delta},$$

it is enough to choose

$$\delta_\star = \frac{\underline{\delta}}{1 - (V-1)\underline{\delta}}.$$

Indeed,

$$\|\max(p, \delta_\star)\|_1 \leq 1 + (V-1)\delta_\star,$$

so every coordinate of the renormalized vector is at least

$$\frac{\delta_\star}{1 + (V-1)\delta_\star} = \underline{\delta}.$$

This clipping map is only a numerical safeguard; it plays no role in the formal smoothness bounds below.

D.4 Trace-level moduli and the gauge convention

Let

$$\Delta^{N-1}_{\delta_{\mathrm{tr}}} := \left\{ p \in \Delta^{N-1} : \min_{y \in \mathcal{Y}_T} p_y \geq \delta_{\mathrm{tr}} \right\}$$

be a trimmed trace simplex, where $N = V^T$. Assume the base trace law satisfies $p_{\mathrm{base}}(y) > 0$ for every $y \in \mathcal{Y}_T$.

Gauge convention. For the composed trace law p^Θ, one has

$$(Dp^\Theta)^\top \mathbf{1} = 0, \qquad \sum_{y \in \mathcal{Y}_T} \nabla^2_\Theta p^\Theta(y) = 0,$$

because $\sum_y p^\Theta(y) \equiv 1$. Consequently, if a trace-level score representative g is replaced by

$$g'(p) = g(p) + c(p)\mathbf{1}$$

for any scalar function $c(p)$, then the first- and second-order derivatives of the composed objective do not change:

$$(Dp^\Theta)^\top g'(p^\Theta) = (Dp^\Theta)^\top g(p^\Theta),$$

and

$$(Dp^\Theta)^\top \nabla g'(p^\Theta) Dp^\Theta + \sum_y g'_y(p^\Theta) \nabla^2_\Theta p^\Theta(y) = (Dp^\Theta)^\top \nabla g(p^\Theta) Dp^\Theta + \sum_y g_y(p^\Theta) \nabla^2_\Theta p^\Theta(y).$$

Thus, for parameter-space geometry, one may work with any ambient Euclidean representative of the simplex gradient, defined only up to an additive $\mathbf{1}$-multiple.

Lemma D.4 (Trace-level moduli of the DCR objective). Define the Euclidean trace-level score representative

$$g(p) := U - 2\lambda\beta K_{\mathrm{eff}} p + \beta_{\mathrm{KL}} \log p_{\mathrm{base}} - A(1 + \log p),$$

where

$$A := \lambda\alpha + \beta_{\mathrm{KL}} + \varepsilon.$$

Set

$$G_{\infty,\mathrm{tr}} := \sup_{p \in \Delta^{N-1}_{\delta_{\mathrm{tr}}}} \|g(p)\|_\infty, \qquad L_{\mathrm{tr}} := \sup_{p \in \Delta^{N-1}_{\delta_{\mathrm{tr}}}} \|\nabla g(p)\|_{1\to\infty}.$$

Then

$$\boxed{G_{\infty,\mathrm{tr}} \le \|U\|_\infty + 2\lambda\beta\|K_{\mathrm{eff}}\|_{1\to\infty} + \beta_{\mathrm{KL}}\|\log p_{\mathrm{base}}\|_\infty + A\big(1 - \log\delta_{\mathrm{tr}}\big),}$$

and

$$\boxed{L_{\mathrm{tr}} \le 2\lambda\beta\|K_{\mathrm{eff}}\|_{1\to\infty} + \frac{A}{\delta_{\mathrm{tr}}}.}$$

Here

$$\|M\|_{1\to\infty} := \sup_{\|x\|_1=1} \|Mx\|_\infty = \max_{i,j} |M_{ij}|.$$

Proof. From the definition of g,

$$\|g(p)\|_\infty \le \|U\|_\infty + 2\lambda\beta\|K_{\mathrm{eff}}p\|_\infty + \beta_{\mathrm{KL}}\|\log p_{\mathrm{base}}\|_\infty + A\|1 + \log p\|_\infty.$$

Now

$$\|K_{\mathrm{eff}}p\|_\infty \le \|K_{\mathrm{eff}}\|_{1\to\infty}\|p\|_1 = \|K_{\mathrm{eff}}\|_{1\to\infty},$$

and on $\Delta^{N-1}_{\delta_{\mathrm{tr}}}$,

$$|1 + \log p_y| \le 1 - \log\delta_{\mathrm{tr}}.$$

This proves the bound for $G_{\infty,\mathrm{tr}}$.

Differentiating g with respect to p gives

$$\nabla g(p) = -2\lambda\beta K_{\mathrm{eff}} - A\ \mathrm{diag}(1/p),$$

since the base-policy term does not depend on p. Therefore

$$\|\nabla g(p)\|_{1\to\infty} \le 2\lambda\beta\|K_{\mathrm{eff}}\|_{1\to\infty} + A\|\,\mathrm{diag}(1/p)\|_{1\to\infty}.$$

Because

$$\|\,\mathrm{diag}(1/p)\|_{1\to\infty} = \max_y \frac{1}{p_y} \le \frac{1}{\delta_{\mathrm{tr}}},$$

the bound for L_{tr} follows. □

Corollary D.4.1 (Uniform-base simplification). If the base trace law is uniform on $\mathcal{Y}_T$, then

$$\log p_{\mathrm{base}} = -\log N\,\mathbf{1}.$$

By the gauge convention, this additive $\mathbf{1}$-multiple can be dropped without changing any composed first- or second-order derivative. In that case one may work with the simplified representative

$$g(p) = U - 2\lambda\beta K_{\mathrm{eff}}p - A(1 + \log p),$$

and the bound on $G_{\infty,\mathrm{tr}}$ loses the term $\beta_{\mathrm{KL}}\|\log p_{\mathrm{base}}\|_\infty$.

Corollary D.4.2 (Lipschitz continuity of the DCR vector field). Define the DCR vector field on the trace simplex by

$$F(p) := p \odot \big(g(p) - \langle p, g(p)\rangle\mathbf{1}\big),$$

where $\odot$ denotes componentwise multiplication. Then on $\Delta^{N-1}_{\delta_{\mathrm{tr}}}$,

$$\boxed{\|F(p) - F(q)\|_1 \le \big(3G_{\infty,\mathrm{tr}} + 2L_{\mathrm{tr}}\big)\|p - q\|_1.}$$

Proof. Write

$$m(p) := \langle p, g(p)\rangle.$$

Then

$$F(p) - F(q) = (p - q) \odot \big(g(p) - m(p)\mathbf{1}\big) + q \odot \big(g(p) - g(q) - (m(p) - m(q))\mathbf{1}\big).$$

For the first term,

$$\|(p-q)\odot(g(p)-m(p)\mathbf{1})\|_1 \le \|p-q\|_1\,\|g(p)-m(p)\mathbf{1}\|_\infty.$$

Since $|m(p)| \le \|g(p)\|_\infty \le G_{\infty,\mathrm{tr}}$,

$$\|g(p)-m(p)\mathbf{1}\|_\infty \le 2G_{\infty,\mathrm{tr}},$$

so

$$\|(p-q)\odot(g(p)-m(p)\mathbf{1})\|_1 \le 2G_{\infty,\mathrm{tr}}\|p-q\|_1.$$

For the second term, convexity of $\Delta^{N-1}_{\delta_{\mathrm{tr}}}$ and the mean-value theorem imply

$$\|g(p)-g(q)\|_\infty \le L_{\mathrm{tr}}\|p-q\|_1.$$

Also,

$$|m(p)-m(q)| = |p^\top g(p)-q^\top g(q)| \le |(p-q)^\top g(p)| + |q^\top(g(p)-g(q))| \le \big(G_{\infty,\mathrm{tr}}+L_{\mathrm{tr}}\big)\|p-q\|_1.$$

Since $\|q\|_1 = 1$,

$$\begin{aligned}\|q\odot(g(p)-g(q)-(m(p)-m(q))\mathbf{1})\|_1 &\le \|g(p)-g(q)\|_\infty + |m(p)-m(q)| \\ &\le \big(G_{\infty,\mathrm{tr}}+2L_{\mathrm{tr}}\big)\|p-q\|_1.\end{aligned}$$

Summing the two estimates gives the claim. □

D.5 Autoregressive composition and sequence-level smoothness

Let

$$\mathcal{U}_T := \bigcup_{t=0}^{T-1}\mathcal{V}^t$$

be the set of prefixes of length strictly less than T. The tabular autoregressive parameter space is

$$\mathbf{\Theta}_T := \Big\{\Theta=(\theta_u)_{u\in\mathcal{U}_T} : \theta_u \in T_V\Big\}.$$

We equip $\mathbf{\Theta}_T$ with the block Euclidean norm

$$\|\Theta\|_2^2 := \sum_{u\in\mathcal{U}_T}\|\theta_u\|_2^2.$$

For $\Theta\in\mathbf{\Theta}_T$, define the local token laws

$$q_u := \mathrm{softmax}(\theta_u),$$

and the induced exact-length sequence law on $\mathcal{Y}_T = \mathcal{V}^T$ by

$$p^\Theta(y_{1:T}) := \prod_{t=1}^{T} q_{y_{<t}}(y_t).$$

Because we work on the full exact-length space $\mathcal{Y}_T$, this is a genuine probability vector in Δ^{N-1}, with no missing mass and no post-EOS corner cases.

Restrict attention to the active region

$$\mathbf{\Theta}_{\underline{\delta},T} := \Big\{\Theta\in\mathbf{\Theta}_T : \min_{u,i} q_u(i) \ge \underline{\delta}\Big\}.$$

This set is convex, because it is a Cartesian product of convex token-level active sets. Moreover, every sequence probability satisfies

$$p^\Theta(y) \ge \delta_{\mathrm{tr}} := \underline{\delta}^T, \qquad y\in\mathcal{Y}_T,$$

so $p^\Theta \in \Delta^{N-1}_{\delta_{\mathrm{tr}}}$.

Define the composed objective

$$\Phi_T(\Theta) := \widetilde{J}(p^\Theta).$$

Theorem D.5 (Autoregressive composition and sequence-level smoothness). On $\mathbf{\Theta}_{\underline{\delta},T}$, the map $\Theta \mapsto p^\Theta$ is C^2, and so is Φ_T. Moreover,

$$\boxed{\|\nabla_\Theta \Phi_T(\Theta)\|_2 \le \sqrt{2T}\, G_{\infty,\mathrm{tr}},}$$

and

$$\boxed{\|\nabla^2_\Theta \Phi_T(\Theta)\|_{\mathrm{op}} \le 2T\, L_{\mathrm{tr}} + \Big(2T + \frac{1}{2}\Big) G_{\infty,\mathrm{tr}} \;=:\; L_{\Theta,T}.}$$

Proof. Fix a sequence $y = (y_1, \dots, y_T) \in \mathcal{Y}_T$ and define its log-probability

$$\ell_y(\Theta) := \log p^\Theta(y) = \sum_{t=1}^{T} \log q_{y_{<t}}(y_t).$$

Exactly T prefix blocks appear in this sum, one for each time step. For the block $u = y_{<t}$,

$$\nabla_{\theta_u} \log q_u(y_t) = e_{y_t} - q_u.$$

Since both e_{y_t} and q_u are probability vectors,

$$\|e_{y_t} - q_u\|_2^2 \le 2.$$

Therefore the full gradient of ℓ_y has exactly T nonzero blocks and

$$\|\nabla_\Theta \ell_y\|_2^2 = \sum_{t=1}^{T} \|e_{y_t} - q_{y_{<t}}\|_2^2 \le 2T.$$

On each visited block,

$$\nabla^2_{\theta_u} \log q_u(y_t) = -J_{\theta_u},$$

so Theorem D.1 yields

$$\|\nabla^2_\Theta \ell_y\|_{\mathrm{op}} \le \frac{1}{2}.$$

Now

$$\nabla_\Theta p^\Theta(y) = p^\Theta(y)\, \nabla_\Theta \ell_y,$$

and

$$\nabla^2_\Theta p^\Theta(y) = p^\Theta(y)\Big(\nabla_\Theta \ell_y\, \nabla_\Theta \ell_y^\top + \nabla^2_\Theta \ell_y\Big).$$

Hence

$$\|\nabla_\Theta p^\Theta(y)\|_2 \le \sqrt{2T}\, p^\Theta(y), \qquad \|\nabla^2_\Theta p^\Theta(y)\|_{\mathrm{op}} \le \Big(2T + \frac{1}{2}\Big) p^\Theta(y).$$

Let Dp^Θ denote the Jacobian of the map $\Theta \mapsto p^\Theta$. For any unit vector h,

$$\|Dp^\Theta[h]\|_1 = \sum_{y \in \mathcal{Y}_T} |\langle \nabla_\Theta p^\Theta(y), h\rangle| \le \sum_y \|\nabla_\Theta p^\Theta(y)\|_2 \le \sqrt{2T} \sum_y p^\Theta(y) = \sqrt{2T}.$$

Therefore

$$\|Dp^\Theta\|_{2\to 1} \le \sqrt{2T}.$$

By the gauge convention of the previous subsection, the chain rule may be written using the representative g:

$$\nabla_\Theta \Phi_T(\Theta) = (Dp^\Theta)^\top g(p^\Theta),$$

and

$$\nabla^2_\Theta \Phi_T(\Theta) = (Dp^\Theta)^\top \nabla g(p^\Theta) Dp^\Theta + \sum_{y \in \mathcal{Y}_T} g_y(p^\Theta)\, \nabla^2_\Theta p^\Theta(y).$$

For the first term,

$$\begin{aligned}\|(Dp^\Theta)^\top \nabla g(p^\Theta) Dp^\Theta\|_{\mathrm{op}} &= \sup_{\|u\|_2=\|v\|_2=1} \big|\langle u, (Dp^\Theta)^\top \nabla g(p^\Theta) Dp^\Theta v\rangle\big| \\ &= \sup_{\|u\|_2=\|v\|_2=1} \big|\langle Dp^\Theta u, \nabla g(p^\Theta) Dp^\Theta v\rangle\big| \\ &\le \sup_{\|u\|_2=\|v\|_2=1} \|Dp^\Theta u\|_1 \, \|\nabla g(p^\Theta) Dp^\Theta v\|_\infty \\ &\le \|Dp^\Theta\|_{2\to 1}^2 \, \|\nabla g(p^\Theta)\|_{1\to\infty} \\ &\le 2T\, L_{\mathrm{tr}}.\end{aligned}$$

For the second term,

$$\Big\| \sum_y g_y(p^\Theta) \nabla^2_\Theta p^\Theta(y) \Big\|_{\mathrm{op}} \le G_{\infty,\mathrm{tr}} \sum_y \|\nabla^2_\Theta p^\Theta(y)\|_{\mathrm{op}} \le \Big(2T + \frac{1}{2}\Big) G_{\infty,\mathrm{tr}} \sum_y p^\Theta(y),$$

which is

$$\Big\| \sum_y g_y(p^\Theta) \nabla^2_\Theta p^\Theta(y) \Big\|_{\mathrm{op}} \le \Big(2T + \frac{1}{2}\Big) G_{\infty,\mathrm{tr}}.$$

Adding the two bounds proves the Hessian estimate.

For the gradient,

$$\|\nabla_\Theta \Phi_T(\Theta)\|_2 \le \|Dp^\Theta\|_{2\to 1} \, \|g(p^\Theta)\|_\infty \le \sqrt{2T}\, G_{\infty,\mathrm{tr}}.$$

This completes the proof. □

What Theorem D.5 does and does not say. The theorem isolates only the smoothness contributed by the *parameterization*. Because

$$\delta_{\mathrm{tr}} = \underline{\delta}^T,$$

the trace-level moduli $G_{\infty,\mathrm{tr}}$ and L_{tr} may themselves depend strongly on T. Thus Theorem D.5 rules out exponential blow-up coming from *repeated softmax composition alone*; it does *not* assert that the full optimization problem is uniformly linear in T.

Conservative step-size guarantee. The region $\mathbf{\Theta}_{\underline{\delta},T}$ is convex, and Φ_T is $L_{\Theta,T}$-smooth there. Therefore, for any $\Theta \in \mathbf{\Theta}_{\underline{\delta},T}$ and any step size $0 < \eta_\Theta < 2/L_{\Theta,T}$ such that the update remains in the active region,

$$\Phi_T\big(\Theta + \eta_\Theta \nabla_\Theta \Phi_T(\Theta)\big) \ge \Phi_T(\Theta) + \eta_\Theta \Big(1 - \frac{L_{\Theta,T}\eta_\Theta}{2}\Big) \|\nabla_\Theta \Phi_T(\Theta)\|_2^2.$$

In particular,

$$\boxed{\eta_\Theta \le \frac{1}{L_{\Theta,T}}}$$

is a conservative sufficient choice for monotone ascent.

D.6 From tabular sequence logits to neural-network weight space

In a neural model, the tabular sequence-logit variable Θ is generated by a parameterized backbone. Let $\mathcal{W} \subseteq \mathbb{R}^d$ be a convex parameter region, and let

$$F_T : \mathcal{W} \to \mathbf{\Theta}_T$$

be a C^2 sequence-logit map such that

$$F_T(\mathcal{W}) \subseteq \mathbf{\Theta}_{\underline{\delta},T}.$$

Define the weight-space objective

$$\Psi_T(W) := \Phi_T(F_T(W)).$$

Assume that on $\mathcal{W}$,

$$\|DF_T(W)\|_{\mathrm{op}} \le J_{W,T}, \qquad \|D^2 F_T(W)\|_{\mathrm{op}} \le H_{W,T},$$

where $\|D^2 F_T(W)\|_{\mathrm{op}}$ denotes the induced bilinear operator norm

$$\sup_{\|u\|_2=\|v\|_2=1} \|D^2 F_T(W)[u,v]\|_2.$$

Corollary D.6 (Mapping to neural-network weight space). Under the assumptions above, Ψ_T is C^2 on $\mathcal{W}$, and

$$\boxed{\|\nabla_W^2 \Psi_T(W)\|_{\mathrm{op}} \le J_{W,T}^2 L_{\Theta,T} + H_{W,T}\sqrt{2T}\, G_{\infty,\mathrm{tr}} \;=:\; L_{W,T}.}$$

Proof. For unit vectors $u, v \in \mathbb{R}^d$, the bilinear chain rule gives

$$\nabla_W^2 \Psi_T(W)[u,v] = \nabla_\Theta^2 \Phi_T(F_T(W))[DF_T(W)u, DF_T(W)v] + \big\langle \nabla_\Theta \Phi_T(F_T(W)), D^2 F_T(W)[u,v] \big\rangle.$$

Therefore

$$\begin{aligned}\big|\nabla_W^2 \Psi_T(W)[u,v]\big| \le{}& \|\nabla_\Theta^2 \Phi_T(F_T(W))\|_{\mathrm{op}}\, \|DF_T(W)u\|_2\, \|DF_T(W)v\|_2 \\ &+ \|\nabla_\Theta \Phi_T(F_T(W))\|_2\, \|D^2 F_T(W)[u,v]\|_2.\end{aligned}$$

Taking the supremum over unit u, v and applying Theorem D.5 yields

$$\|\nabla_W^2 \Psi_T(W)\|_{\mathrm{op}} \le J_{W,T}^2 L_{\Theta,T} + H_{W,T}\sqrt{2T}\, G_{\infty,\mathrm{tr}},$$

as claimed. □

On any such convex region $\mathcal{W}$, any step size $0 < \eta_W < 2/L_{W,T}$ such that

$$W + \eta_W \nabla_W \Psi_T(W) \in \mathcal{W}$$

satisfies the standard ascent lemma. In particular,

$$\boxed{\eta_W \le \frac{1}{L_{W,T}}}$$

is a conservative sufficient choice for monotone ascent.

Conclusion. A single softmax block has fully dimension-free curvature with exact sharp constants. Composing such blocks autoregressively contributes only an explicit linear factor in the horizon T to the parameterization itself. Thus no exponential ill-conditioning arises from repeated softmax composition alone. Any residual conditioning must come from the trace-level DCR geometry encoded in $G_{\infty,\mathrm{tr}}$ and L_{tr}, or from the architecture-dependent map F_T from weights to sequence logits.

E STOCHASTIC DYNAMICS OF POST-TRAINING ON THE SIMPLEX

The master DCR ODE of Appendix A describes the *expected* drift of training on the probability simplex. Real post-training uses finite batches of sampled traces, so the update is noisy. This appendix studies stochastic simplex surrogates that preserve the correct tangent covariance geometry while remaining analytically tractable.

Three distinctions matter up front. First, this appendix is *not* an exact weak-limit theorem for any one post-training optimizer. Instead, we work with *moment-matched* diffusions whose covariance is proportional to the multinomial covariance

$$J_p := \mathrm{diag}(p) - pp^\top.$$

Second, we keep three processes separate throughout: (i) an unreflected diffusion on the open simplex, used only up to its first boundary hit; (ii) its standard absorbing continuation on the closed simplex, used for classical Wright–Fisher fixation statements; and (iii) a normally reflected diffusion on a trimmed simplex, used to model numerical floors and interior confinement. Third, the GRPO discussion below is an *idealized* surrogate designed to isolate the effect of group-normalized variance depletion; it should not be read either as an optimizer-specific SDE limit theorem or as the exact deterministic GRPO mean field analyzed in Appendix B.

These distinctions matter. They prevent boundary fixation, numerical reflection, and local fluctuations around an interior equilibrium from being conflated.

E.1 A moment-matched simplex diffusion surrogate

Write

$$\Delta^{S-1} := \{p \in [0,1]^S : \sum_{i=1}^{S} p_i = 1\}, \qquad \operatorname{int}\Delta^{S-1} := \{p \in \Delta^{S-1} : p_i > 0 \ \forall i\},$$

and let

$$T_\Delta := \mathbf{1}^\perp = \{v \in \mathbb{R}^S : \langle \mathbf{1}, v\rangle = 0\}$$

be the simplex tangent space.

Fix a drift field

$$F : \operatorname{int}\Delta^{S-1} \to T_\Delta.$$

Define the volatility matrix

$$\Sigma_{ik}(p) := \delta_{ik}\sqrt{p_i} - p_i\sqrt{p_k}, \qquad p \in \operatorname{int}\Delta^{S-1}.$$

A direct computation gives

$$\Sigma(p)\Sigma(p)^\top = \operatorname{diag}(p) - pp^\top = J_p, \qquad \mathbf{1}^\top\Sigma(p) = 0.$$

Thus $\Sigma(p)$ injects tangent noise with exactly the multinomial covariance geometry.

Let

$$q : \operatorname{int}\Delta^{S-1} \to [0,\infty)$$

be a nonnegative modulation function, let $\gamma > 0$ be the overall stochastic scale, and define the first boundary hitting time

$$\tau_\partial := \inf\{t \geq 0 : \min_i p_i(t) = 0\}.$$

Theorem E.1 (Moment-matched simplex diffusion surrogate)**.** *Let (p_t) be any continuous semimartingale solution, up to time τ_∂, of the Itô SDE*

$$dp_t = F(p_t)\,dt + \sqrt{\gamma\, q(p_t)}\,\Sigma(p_t)\,dW_t, \qquad t < \tau_\partial, \tag{35}$$

where W_t is standard S-dimensional Brownian motion. Then:

1. *the simplex constraint is preserved,*

$$\sum_{i=1}^{S} p_i(t) = 1 \qquad \text{for all } t < \tau_\partial;$$

2. *the predictable quadratic variation is*

$$d\langle p\rangle_t = \gamma\, q(p_t)\, J_{p_t}\, dt; \tag{36}$$

3. *the choice $q \equiv 1$ yields the constant-modulation Wright–Fisher surrogate.*

Proof. Since $\mathbf{1}^\top F(p) = 0$ and $\mathbf{1}^\top\Sigma(p) = 0$, Itô's formula gives

$$d\Big(\sum_{i=1}^{S} p_i(t)\Big) = 0.$$

This proves mass conservation. The quadratic variation is

$$d\langle p\rangle_t = \gamma\, q(p_t)\,\Sigma(p_t)\Sigma(p_t)^\top\,dt = \gamma\, q(p_t)\, J_{p_t}\,dt,$$

which is exactly (36). The final claim is immediate. □

Conventions on continuation. Equation (35) is only used on the open simplex, hence only up to τ_∂. Whenever we later discuss *global* behavior on the closed simplex, we will say so explicitly. In particular: (i) classical Wright–Fisher fixation statements are made for the standard absorbing closed-simplex process; (ii) numerical-floor statements are made for a reflected process on a trimmed simplex.

GRPO variance-depletion proxy. Under binary outcome supervision and group size $G \geq 2$, group-standard-deviation normalization produces zero advantage whenever all sampled traces are correct or all are incorrect. A natural first-order modulation is the mixed-group probability

$$q_G(\rho) := 1 - \rho^G - (1-\rho)^G, \qquad \rho \in [0,1], \tag{37}$$

which is exactly the probability that an i.i.d. group contains at least one correct and at least one incorrect trace. We use $q \equiv 1$ for the idealized constant-modulation Wright–Fisher regime, and $q = q_G \circ \rho$ for the GRPO-inspired variance-depletion regime in Section E.4.

Corollary E.1 (Log-ratio SDE for the unreflected surrogate)**.** *Assume the drift has DCR form*

$$F_i(p) = p_i\Big(\phi_i(p) - \bar{\phi}(p) - A\big(\log p_i - \langle \log p \rangle\big)\Big),$$

with effective entropic weight $A > 0$, where

$$\bar{\phi}(p) := \sum_{j=1}^{S} p_j \phi_j(p), \qquad \langle \log p \rangle := \sum_{j=1}^{S} p_j \log p_j.$$

(In the pure scalar-objective limit of Appendix B, $A = \varepsilon$.) For any pair of traces (a, b), define

$$z_{ab}(t) := \log p_a(t) - \log p_b(t), \qquad \tau_{ab} := \inf\{t \geq 0 : p_a(t)p_b(t) = 0\}.$$

Then on $[0, \tau_{ab} \wedge \tau_\partial)$,

$$\begin{aligned} dz_{ab}(t) = &\Big(\phi_a(p_t) - \phi_b(p_t) - A z_{ab}(t)\Big)dt \\ &- \frac{\gamma q(p_t)}{2}\left(\frac{1-p_a(t)}{p_a(t)} - \frac{1-p_b(t)}{p_b(t)}\right)dt \qquad (38) \\ &+ \sqrt{\gamma q(p_t)}\left(\frac{dW_a(t)}{\sqrt{p_a(t)}} - \frac{dW_b(t)}{\sqrt{p_b(t)}}\right). \qquad (39) \end{aligned}$$

Proof. Apply Itô's formula to

$$f(p) = \log p_a - \log p_b$$

on the open simplex, stopped at $\tau_{ab} \wedge \tau_\partial$. The first-order terms reproduce the deterministic log-ratio drift from Appendix A:

$$\frac{F_a(p)}{p_a} - \frac{F_b(p)}{p_b} = \phi_a(p) - \phi_b(p) - A(\log p_a - \log p_b).$$

The second-order terms use

$$d\langle p_a, p_a\rangle_t = \gamma q(p_t) p_a(t)(1-p_a(t))\,dt, \qquad d\langle p_b, p_b\rangle_t = \gamma q(p_t) p_b(t)(1-p_b(t))\,dt,$$

so

$$-\frac{1}{2}\frac{d\langle p_a\rangle_t}{p_a(t)^2} + \frac{1}{2}\frac{d\langle p_b\rangle_t}{p_b(t)^2} = -\frac{\gamma q(p_t)}{2}\left(\frac{1-p_a(t)}{p_a(t)} - \frac{1-p_b(t)}{p_b(t)}\right)dt.$$

Finally,

$$\frac{1}{p_a}\Sigma_{a\cdot}(p) - \frac{1}{p_b}\Sigma_{b\cdot}(p) = \frac{e_a^\top}{\sqrt{p_a}} - \frac{e_b^\top}{\sqrt{p_b}},$$

which yields the martingale term in (38). □

Interpretation. Even when the diffusion is symmetric in probability space, the logarithm introduces an Itô correction that penalizes minority strategies as they approach the boundary. Under GRPO, that same geometric correction is multiplied by the state-dependent modulation $q_G(\rho)$, so it disappears when mixed groups become rare.

E.2 Reflection on a trimmed simplex and local boundary repulsion

Fix a computational floor

$$\delta_\star \in (0, 1/S), \qquad D_{\delta_\star} := \Delta^{S-1}_{\delta_\star} := \{p \in \Delta^{S-1} : \min_i p_i \geq \delta_\star\}.$$

This set lies in the affine hyperplane

$$H := \{p \in \mathbb{R}^S : \langle \mathbf{1}, p\rangle = 1\}.$$

All normals, tangent cones, and reflections in this subsection are understood *relative to* H. Equivalently, one may identify H isometrically with $\mathbb{R}^{S-1}$ by choosing an orthonormal basis of $\mathbf{1}^\perp$. We make that identification explicit in Theorem E.3.

Before introducing reflection, we first isolate the intrinsic one-coordinate repulsion produced by the entropic barrier.

Theorem E.2 (Local boundary repulsion for the unreflected surrogate)**.** *Fix a coordinate i and a boundary band $[\delta_\star, y_{\max}]$ with $\delta_\star < y_{\max} \leq 1$. Assume q and $\bar\phi - \phi_i$ admit continuous extensions to the closed band*

$$\mathcal{B}_{i,\mathrm{band}} := \{p \in \Delta^{S-1} : p_i \in [\delta_\star, y_{\max}]\},$$

and define the finite constants

$$\Gamma_{\mathrm{band}} := \inf_{p \in \mathcal{B}_{i,\mathrm{band}}} \big(\langle \log p\rangle - \log p_i\big),$$

$$M^{(i,\mathrm{band})}_{\mathrm{out}} := \sup_{p \in \mathcal{B}_{i,\mathrm{band}}} \big(\bar\phi(p) - \phi_i(p)\big), \qquad q^{(i,\mathrm{band})}_{\max} := \sup_{p \in \mathcal{B}_{i,\mathrm{band}}} q(p).$$

Assume $\Gamma_{\mathrm{band}} > 0$; for example, any band with $y_{\max} < 1/S$ satisfies this by continuity and compactness. If

$$A\,\Gamma_{\mathrm{band}} > M^{(i,\mathrm{band})}_{\mathrm{out}},$$

set

$$\mu := \delta_\star\Big(A\,\Gamma_{\mathrm{band}} - M^{(i,\mathrm{band})}_{\mathrm{out}}\Big) > 0,$$

and define

$$\sigma^2_{\max} := \gamma q^{(i,\mathrm{band})}_{\max} \sup_{x \in [\delta_\star, y_{\max}]} x(1-x).$$

Let

$$\tau_\delta := \inf\{t \geq 0 : p_i(t) = \delta_\star\}, \qquad \tau_y := \inf\{t \geq 0 : p_i(t) = y_{\max}\}.$$

Assume in addition that the initial condition satisfies

$$p_i(0) \in (\delta_\star, y_{\max}).$$

Then, with the convention $e^{-\infty} = 0$ when $\sigma^2_{\max} = 0$,

$$\mathbb{P}\big(\tau_\delta < \tau_y \wedge \tau_\partial\big) \leq \exp\left(-\frac{2\mu}{\sigma^2_{\max}}\big(p_i(0) - \delta_\star\big)\right). \tag{40}$$

In particular, if $\sigma^2_{\max} = 0$, then the process is deterministic on the band and the hitting probability is 0.

Proof. Set

$$\tau := \tau_\delta \wedge \tau_y \wedge \tau_\partial.$$

Because $p_i(0) \in (\delta_\star, y_{\max})$ and the paths are continuous, on the event $\{t < \tau\}$ one has $p_t \in \mathcal{B}_{i,\mathrm{band}}$. Hence

$$\begin{aligned} F_i(p_t) &= p_i(t)\Big(\phi_i(p_t) - \bar\phi(p_t) + A\big(\langle \log p_t\rangle - \log p_i(t)\big)\Big) \\ &\geq \delta_\star\Big(-M^{(i,\mathrm{band})}_{\mathrm{out}} + A\Gamma_{\mathrm{band}}\Big) = \mu. \end{aligned}$$

Also,

$$\frac{d}{dt}\langle p_i\rangle_t = \gamma q(p_t)p_i(t)(1-p_i(t)) \le \sigma_{\max}^2.$$

If $\sigma_{\max}^2 = 0$, then p_i is deterministic on the band and has drift at least $\mu > 0$, so it cannot hit $\delta_\star$ before leaving upward. This gives the last sentence.

Assume now $\sigma_{\max}^2 > 0$ and set

$$\lambda := \frac{2\mu}{\sigma_{\max}^2}, \qquad S(x) := \exp\big(-\lambda(x-\delta_\star)\big).$$

By Itô's formula,

$$dS\big(p_i(t\wedge\tau)\big) = S'\big(p_i\big)dp_i + \frac{1}{2}S''\big(p_i\big)d\langle p_i\rangle.$$

Using $S' = -\lambda S$ and $S'' = \lambda^2 S$, together with the drift and variance bounds above, gives

$$\mathcal{L}S(x) \le -\lambda\mu S(x) + \frac{1}{2}\lambda^2\sigma_{\max}^2 S(x) = 0 \qquad \text{for } x \in [\delta_\star, y_{\max}].$$

Hence

$$X_t := S\big(p_i(t\wedge\tau)\big)$$

is a bounded supermartingale, so for every $T > 0$,

$$\mathbb{E}[X_T] \le X_0 = S\big(p_i(0)\big).$$

Let

$$A := \{\tau_\delta < \tau_y \wedge \tau_\partial\}.$$

On the event A, one has $\tau = \tau_\delta < \infty$, and therefore $X_T = 1$ for all $T \ge \tau$. Thus

$$\mathbf{1}_A \le \liminf_{T\to\infty} X_T \qquad \text{almost surely.}$$

By Fatou's lemma,

$$\mathbb{P}(A) \le \mathbb{E}\Big[\liminf_{T\to\infty} X_T\Big] \le \liminf_{T\to\infty}\mathbb{E}[X_T] \le S\big(p_i(0)\big).$$

This is exactly (40). □

The previous theorem is deliberately local: it controls one-coordinate downward excursions inside a prescribed band. It does *not* by itself imply global ergodicity on the open simplex. To model global confinement with a numerical floor, we now switch to a reflected diffusion on the trimmed simplex.

Theorem E.3 (Reflected surrogate on the trimmed simplex)**.** *Let $D := D_{\delta_\star} \subset H$. Assume that the drift and diffusion coefficients*

$$b(p) := F(p), \qquad G(p) := \sqrt{\gamma q(p)}\,\Sigma(p)$$

are Lipschitz on D. Consider the reflected SDE on the hyperplane H

$$dP_t = b(P_t)\,dt + G(P_t)\,dW_t + dK_t, \qquad P_t \in D, \tag{41}$$

where $K_0 = 0$ is a bounded-variation process such that, pathwise,

$$dK_t \in N_D^{H,\mathrm{in}}(P_t)\,d|K|_t,$$

with $N_D^{H,\mathrm{in}}(p)$ denoting the relative inward normal cone *of D at p inside the hyperplane H. Then* (41) *admits a unique strong solution. Moreover, the associated Markov process on compact state space D has at least one invariant probability measure, which we denote by*

$$\Pi_{\gamma,\delta_\star,q}.$$

Proof. Fix the barycenter

$$p^\circ := \frac{1}{S}\mathbf{1} \in H,$$

and choose an orthonormal matrix $Q \in \mathbb{R}^{S\times(S-1)}$ whose columns form a basis of $\mathbf{1}^\perp$. Then every point in H can be written uniquely as

$$p = p^\circ + Qx, \qquad x \in \mathbb{R}^{S-1}.$$

Under this identification, the trimmed simplex becomes the compact convex polytope

$$\widehat{D} := \{x \in \mathbb{R}^{S-1} : p^\circ + Qx \in D\}.$$

Define the projected coefficients

$$\widehat{b}(x) := Q^\top b(p^\circ + Qx), \qquad \widehat{G}(x) := Q^\top G(p^\circ + Qx).$$

Because b and G are Lipschitz on D, the projected coefficients are Lipschitz on $\widehat{D}$. The relative inward normal cone of D inside H is carried by $Q^\top$ to the ordinary inward normal cone of $\widehat{D}$ in $\mathbb{R}^{S-1}$. Therefore the reflected SDE

$$dX_t = \widehat{b}(X_t)\,dt + \widehat{G}(X_t)\,dW_t + d\widehat{K}_t, \qquad X_t \in \widehat{D},$$

with normal reflection on the convex domain $\widehat{D} \subset \mathbb{R}^{S-1}$, falls under the standard Skorokhod theory for reflected SDEs on convex domains (Lions and Sznitman, 1984; Saisho, 1987). Hence it has a unique strong solution. Mapping back by

$$P_t := p^\circ + QX_t, \qquad K_t := Q\widehat{K}_t,$$

yields a unique strong solution of (41). Since D is compact and the reflected diffusion is Feller, the Krylov–Bogolyubov theorem gives at least one invariant probability measure on D (Ethier and Kurtz, 1986). □

For the DCR coefficients from Appendix A, and for continuous modulations such as $q \equiv 1$ or $q = q_G \circ \rho$, the hypotheses of Theorem E.3 are automatic on $D_{\delta_\star}$. The reason is straightforward: on a trimmed simplex every coordinate is bounded away from zero, so the coefficients are smooth and hence Lipschitz on the compact state space.

E.3 Heterozygosity, fixation, and local fluctuations

We now separate three effects that should not be conflated: (i) exact neutral fixation in the absorbing Wright–Fisher model, (ii) the opposing effect of numerical reflection on a trimmed simplex, and (iii) local Gaussian fluctuations around an interior DCR equilibrium.

Restrict attention in this subsection to the correct-trace simplex

$$\Delta^{m-1} := \big\{p \in [0,1]^m : \sum_{j=1}^m p_j = 1\big\}, \qquad m := |\mathcal{C}|,$$

and its trimmed version

$$\Delta^{m-1}_{\delta_\star} := \big\{p \in \Delta^{m-1} : \min_{1\le j\le m} p_j \ge \delta_\star\big\}, \qquad \delta_\star \in (0, 1/m).$$

Define the heterozygosity (equivalently, the Gini–Simpson diversity index)

$$H(p) := 1 - \sum_{j=1}^m p_j^2.$$

Then $H(p) = 0$ if and only if p is a simplex vertex.

Let $b \in \mathbf{1}^\perp$ be a microscopic bias vector, and write

$$m_b(p) := \sum_{j=1}^m b_j p_j.$$

Consider the reflected diffusion on $\Delta_{\delta_\star}^{m-1}$ with drift

$$F_i^b(p) := p_i\big(b_i - m_b(p)\big) - \varepsilon p_i\big(\log p_i - \langle \log p\rangle\big), \qquad 1 \le i \le m,$$

and covariance modulation q as in Theorem E.1. Here $\Sigma(p)$ denotes the $m \times m$ version of the moment-matched volatility matrix.

Theorem E.4 (Reflection opposes heterogeneity loss)**.** *For the reflected surrogate above,*

$$dH(P_t) = \Big(-\gamma q(P_t)H(P_t) - 2\mathcal{C}_b(P_t) + 2\varepsilon C_{\mathrm{ent}}(P_t)\Big)dt + dM_t + \langle \nabla H(P_t), dK_t\rangle, \tag{42}$$

where

$$\mathcal{C}_b(p) := \sum_{i=1}^m p_i^2 b_i - \Big(\sum_{i=1}^m p_i^2\Big)\Big(\sum_{j=1}^m p_j b_j\Big),$$

$$C_{\mathrm{ent}}(p) := \sum_{i=1}^m p_i^2\big(\log p_i - \langle \log p\rangle\big) \ge 0,$$

and M_t is a local martingale. Moreover,

$$\langle \nabla H(P_t), dK_t\rangle \ge 0 \qquad \textit{pathwise.} \tag{43}$$

Thus the reflection term contributes nonnegatively to the heterozygosity balance and acts locally against heterogeneity loss.

Proof. Since $\nabla H(p) = -2p$ and $\nabla^2 H = -2I$, Itô's formula gives

$$\frac{1}{2}\operatorname{Tr}\big(\gamma q(P_t) J_{P_t} \nabla^2 H\big) = -\gamma q(P_t) H(P_t).$$

Also,

$$\begin{aligned}\langle \nabla H(p), F^b(p)\rangle &= -2\sum_{i=1}^m p_i^2\big(b_i - m_b(p)\big) + 2\varepsilon \sum_{i=1}^m p_i^2\big(\log p_i - \langle \log p\rangle\big)\\ &= -2\mathcal{C}_b(p) + 2\varepsilon C_{\mathrm{ent}}(p).\end{aligned}$$

This yields (42).

It remains to prove the sign statements. First,

$$2C_{\mathrm{ent}}(p) = \sum_{i,j=1}^m p_i p_j (p_i - p_j)(\log p_i - \log p_j) \ge 0,$$

because $(u-v)(\log u - \log v) \ge 0$ for all $u, v > 0$. Second, on an active face $\{p_k = \delta_\star\}$ of the trimmed simplex, the relative inward normal is

$$n^{(k)} = e_k - \frac{1}{m}\mathbf{1}.$$

Therefore

$$\langle \nabla H(p), n^{(k)}\rangle = \big\langle -2p, e_k - \tfrac{1}{m}\mathbf{1}\big\rangle = 2\Big(\frac{1}{m} - \delta_\star\Big) > 0.$$

At a point with several active faces, the reflection increment dK_t lies in the cone generated by the corresponding inward normals, so the same inequality yields (43). □

Corollary E.2 (Neutral Wright–Fisher fixation on the closed simplex)**.** *Consider the classical neutral Wright–Fisher diffusion on the closed simplex Δ^{m-1}, equivalently the absorbing continuation of the unreflected constant-modulation surrogate with*

$$q \equiv 1, \qquad b = 0, \qquad \varepsilon = 0.$$

Then

$$dH_t = -\gamma H_t\, dt + dM_t, \qquad \mathbb{E}[H_t] = e^{-\gamma t} H_0. \tag{44}$$

Consequently,

$$p_t \to e_K \qquad \textit{almost surely for some random vertex } e_K.$$

Proof. With $b = 0$, $\varepsilon = 0$, and no reflection, Theorem E.4 reduces to

$$dH_t = -\gamma H_t\, dt + dM_t.$$

Thus $(H_t)_{t\geq 0}$ is a bounded nonnegative supermartingale, so it converges almost surely and in L^1 to some random variable $H_\infty \geq 0$. Taking expectations in (44) gives

$$\mathbb{E}[H_t] = e^{-\gamma t} H_0 \to 0,$$

so necessarily $\mathbb{E}[H_\infty] = 0$ and therefore $H_\infty = 0$ almost surely.

It remains to upgrade $H_t \to 0$ to vertex convergence of p_t. Fix $\eta \in (0, 1/2)$ and define the disjoint vertex neighborhoods

$$U_k^\eta := \{p \in \Delta^{m-1} : p_k > 1 - \eta\}, \qquad 1 \leq k \leq m.$$

If $p \notin \bigcup_{k=1}^m U_k^\eta$, then every coordinate satisfies $p_k \leq 1 - \eta$, and hence

$$\sum_{k=1}^m p_k^2 \leq (1-\eta)^2 + \eta^2, \qquad afortiori \qquad H(p) \geq 2\eta(1-\eta).$$

Since $H_t \to 0$ almost surely, there exists an almost surely finite random time T_η such that

$$H_t < 2\eta(1-\eta) \qquad \text{for all } t \geq T_\eta.$$

Thus, for all sufficiently large t, the process lies in the disjoint union $\bigcup_k U_k^\eta$. By path continuity, it cannot move from one component U_i^η to another U_j^η without crossing the complement of that union, where $H \geq 2\eta(1-\eta)$. Hence there exists a random index K such that eventually

$$p_t \in U_K^\eta.$$

Taking $\eta = 1/4$, we have $p_K(t) \geq 3/4$ for all large t, and then

$$\begin{aligned} H_t &= 1 - \sum_{j=1}^m p_j(t)^2 \\ &\geq 1 - p_K(t)^2 - \big(1 - p_K(t)\big)^2 \\ &= 2p_K(t)\big(1 - p_K(t)\big) \\ &\geq \frac{3}{2}\big(1 - p_K(t)\big). \end{aligned}$$

Therefore $1 - p_K(t) \leq \frac{2}{3} H_t \to 0$, which implies $p_K(t) \to 1$ and $p_j(t) \to 0$ for $j \neq K$. So $p_t \to e_K$ almost surely. □

Remark E.1 (Local fluctuation scale near an interior DCR equilibrium)**.** *Let $p^\star \in \operatorname{int} \Delta^{S-1}$ be the unique DCR equilibrium from Appendix A, and set $q_\star := q(p^\star)$. Because $p^\star$ is interior, no reflection is active in a sufficiently small neighborhood of $p^\star$.*

Let

$$A_\star := DF(p^\star)\big|_{T_\Delta}$$

be the Jacobian of the deterministic drift restricted to the tangent space, and define

$$B_\star := -A_\star.$$

Also write

$$G(p) := \sqrt{\gamma q(p)}\, \Sigma(p), \qquad G_\star := G(p^\star) = \sqrt{\gamma q_\star}\, \Sigma_\star, \qquad \Sigma_\star := \Sigma(p^\star).$$

The exact first-order expansion of the unreflected surrogate around $p^\star$ contains the additional state-dependent noise term

$$\big(DG(p^\star)[\xi_t]\big)\, dW_t.$$

If one freezes the diffusion coefficient at $p^\star$ and retains only the linearized drift, one obtains the leading Ornstein–Uhlenbeck approximation

$$d\xi_t = A_\star \xi_t\, dt + G_\star\, dW_t = -B_\star \xi_t\, dt + \sqrt{\gamma q_\star}\, \Sigma_\star\, dW_t.$$

Its stationary covariance $\mathcal{C}$ solves the Lyapunov equation

$$B_\star \mathcal{C} + \mathcal{C} B_\star^\top = \gamma q_\star Q_\star, \qquad Q_\star := J_{p^\star}.$$

Hence the local fluctuation scale is controlled by the ratio $\gamma q_\star/\kappa$, where κ is the smallest positive contraction rate of the linearized deterministic drift (for example, the smallest positive eigenvalue of the symmetric part of $B_\star$ on T_Δ). This recovers the usual heuristic: stable diversity requires the effective diffusion scale to remain small relative to deterministic contraction toward the equilibrium.

E.4 GRPO-style variance depletion and conditional stochastic freezing

The constant-modulation Wright–Fisher model of Corollary E.2 retains a nonvanishing stochastic forcing up to absorption/fixation on the closed simplex. GRPO changes this picture because group-standard-deviation normalization turns the stochastic signal off when sampled groups become nearly pure-correct or nearly pure-incorrect (Shao et al., 2024).

To isolate that variance-depletion mechanism *alone*, we do *not* use the exact deterministic GRPO mean field from Appendix B. Appendix B already showed that the exact mean field contains deterministic within-class amplification and obeys the class-mass identity

$$\dot\rho = h_G(\rho)\, \rho^2 (1-\rho)^2 \big(S_2 + R_2\big).$$

That exact flow is not the object of study here. Instead, we introduce a simpler class-symmetric two-block surrogate that preserves the sign pattern of correct-versus-incorrect class-mass drift and the same mixed-group variance shutdown factor q_G, while deliberately suppressing the exact within-class amplification proved in Appendix B.

Let

$$\rho(p) := \sum_{c \in \mathcal{C}} p_c$$

be the total correct mass, and let q_G be defined by (37). Let $h_G : [0,1] \to (0,\infty)$ be the positive scalar function from Appendix B. We recall its endpoint values as

$$h_G(0) := h_G(1) := \sqrt{G-1}.$$

Part 1 of Theorem E.5 verifies that this definition is continuous. We combine this modulation with the two-block GRPO surrogate drift

$$F_i(p) = \begin{cases} p_i(1-\rho) h_G(\rho), & i \in \mathcal{C}, \\ -p_i \rho\, h_G(\rho), & i \in \mathcal{I}, \end{cases} \tag{45}$$

where the endpoint values above are used when $\rho \in \{0,1\}$. We emphasize again that (45) is an *idealized surrogate*: the point of the theorem below is to separate the geometry of variance depletion from the optimizer-specific details of any concrete implementation.

Theorem E.5 (Variance depletion and conditional stochastic freezing in a two-block GRPO surrogate)**.** *Assume binary outcome supervision and group size $G \geq 2$. Then:*

1. *The mixed-group modulation satisfies*

$$q_G(\rho) \sim G(1-\rho) \qquad \text{as } \rho \uparrow 1.$$

Moreover, the endpoint extension of h_G is continuous on $[0,1]$, with

$$h_G(0) = h_G(1) = \sqrt{G-1}.$$

2. *Consider any continuous semimartingale $(p_t)_{t\geq 0}$ on the closed simplex Δ^{S-1} with decomposition*

$$p_i(t) = p_i(0) + A_i(t) + M_i(t), \qquad A_i(t) := \int_0^t F_i(p_s)\, ds,$$

where F is the two-block GRPO surrogate drift (45), each M_i is a continuous local martingale, and the predictable quadratic variation satisfies

$$d\langle p\rangle_t = \gamma q_G(\rho_t) J_{p_t}\, dt. \tag{46}$$

Then each drift term is monotone in time:

$$A_i \text{ is nondecreasing for } i \in \mathcal{C}, \qquad A_i \text{ is nonincreasing for } i \in \mathcal{I}.$$

3. *The martingale quadratic variation satisfies*

$$d\langle M_i\rangle_t = \gamma q_G(\rho_t) p_i(t)(1-p_i(t))\, dt \leq \frac{\gamma}{4} q_G(\rho_t)\, dt. \tag{47}$$

Consequently, on any event where

$$\int_0^\infty q_G(\rho_t)\, dt < \infty, \tag{48}$$

every coordinate $p_i(t)$ converges almost surely on that event.

4. *In particular, along the deterministic class-mass trajectory of the same two-block surrogate with initial condition $\rho_0 \in (0,1)$,*

$$\dot\rho = \rho(1-\rho) h_G(\rho), \tag{49}$$

one has

$$\rho_t \uparrow 1,$$

and

$$\int_0^\infty q_G(\rho_t)\, dt = \int_{\rho_0}^1 \frac{q_G(\rho)}{\rho(1-\rho)h_G(\rho)}\, d\rho < \infty. \tag{50}$$

Thus the deterministic two-block mastery regime has a finite cumulative variance budget.

Proof. For the asymptotic of q_G, write $\rho = 1-\varepsilon$. Then

$$q_G(1-\varepsilon) = 1 - (1-\varepsilon)^G - \varepsilon^G = G\varepsilon + O(\varepsilon^2), \qquad \varepsilon \downarrow 0,$$

which proves $q_G(\rho) \sim G(1-\rho)$.

Next we verify the endpoint values of h_G. By the correct-trace formula from Appendix B,

$$(1-\rho)h_G(\rho) = \mathbb{E}_{S\sim \mathrm{Binom}(G-1,\rho)}\left[\sqrt{\frac{G-1-S}{S+1}}\right].$$

As $\rho \downarrow 0$, the binomial law concentrates on $S=0$, so

$$(1-\rho)h_G(\rho) \to \sqrt{G-1}.$$

Since $1-\rho \to 1$, this yields

$$h_G(\rho) \to \sqrt{G-1} \qquad \text{as } \rho \downarrow 0.$$

Now set $\rho = 1-\varepsilon$. The $S = G-2$ term contributes

$$\mathbb{P}(S=G-2)\sqrt{\frac{1}{G-1}} = (G-1)(1-\varepsilon)^{G-2}\varepsilon \cdot \frac{1}{\sqrt{G-1}} = \sqrt{G-1}\,\varepsilon + O(\varepsilon^2),$$

while all terms with $S \le G-3$ are $O(\varepsilon^2)$. Therefore

$$(1-\rho)h_G(\rho) = \sqrt{G-1}(1-\rho) + O\big((1-\rho)^2\big),$$

so

$$h_G(\rho) \to \sqrt{G-1} \qquad \text{as } \rho \uparrow 1.$$

This proves continuity of the endpoint extension on $[0,1]$.

The monotonicity of A_i is immediate from the sign of the two-block drift (45). For the quadratic variation, (46) gives

$$d\langle M_i\rangle_t = d\langle p_i\rangle_t = \gamma q_G(\rho_t)\,[J_{p_t}]_{ii}\,dt = \gamma q_G(\rho_t) p_i(t)(1-p_i(t))\,dt,$$

which is exactly (47).

Assume now that (48) holds. Then (47) implies $\langle M_i\rangle_\infty < \infty$, so $M_i(t)$ converges almost surely by the local-martingale convergence theorem. Since $p_i(t) \in [0,1]$ for all t, the identity

$$A_i(t) = p_i(t) - p_i(0) - M_i(t)$$

shows that $A_i(t)$ is bounded. Because A_i is monotone, it must converge. Hence $p_i(t) = p_i(0) + A_i(t) + M_i(t)$ converges almost surely for every coordinate on the event (48).

Finally, along the deterministic mastery trajectory (49), one has $\dot\rho > 0$ on $(0,1)$, so ρ_t is increasing and bounded above by 1. Hence $\rho_t \to L \in [\rho_0, 1]$. If $L < 1$, then continuity of the right-hand side of (49) gives

$$\dot\rho_t \to L(1-L)h_G(L) > 0,$$

contradicting convergence of ρ_t to L. Therefore $\rho_t \uparrow 1$.

Along this same trajectory,

$$dt = \frac{d\rho}{\rho(1-\rho)h_G(\rho)},$$

so

$$\int_0^\infty q_G(\rho_t)\,dt = \int_{\rho_0}^1 \frac{q_G(\rho)}{\rho(1-\rho)h_G(\rho)}\,d\rho.$$

Near $\rho = 1$, the numerator satisfies $q_G(\rho) \sim G(1-\rho)$, while $h_G(\rho) \to \sqrt{G-1} > 0$. Therefore the integrand remains bounded near 1, and the integral is finite. This proves (50). □

Interpretation. Theorem E.5 does *not* rule out rare boundary hits before mastery, and it does *not* assert that every stochastic trajectory of the two-block surrogate automatically satisfies the finite-budget condition (48). What it shows is conditional and precise: on any path for which the cumulative variance budget $\int_0^\infty q_G(\rho_t)\,dt$ is finite, the martingale part has finite quadratic variation and the coordinates converge. Along the deterministic mastery trajectory of the same two-block surrogate, this finite-budget condition holds exactly. This is the clean mathematical sense in which variance depletion can replace persistent Wright–Fisher-type stochastic forcing by convergence to a path-dependent limiting portfolio. This finite-budget conclusion is specific to the two-block surrogate (45); it is not a statement about the exact GRPO mean field from Appendix B.

Edge cases. If $\rho_0 = 1$, then $q_G(\rho_t) = 0$ and the process is already frozen. If $\rho_0 = 0$, then neither the two-block drift nor the covariance term creates correct mass, so the correct set is never entered. The relevant regime is therefore $\rho_0 \in (0,1)$.

Bibliographic note. The reflected-diffusion statement above is the standard Skorokhod construction on a convex domain, applied after identifying the simplex hyperplane with $\mathbb{R}^{S-1}$; classical references are Lions and Sznitman (1984) and Saisho (1987). The GRPO-specific discussion in Section E.4 uses a two-block, moment-matched surrogate motivated by group-standard-deviation normalization, not an exact covariance formula or exact deterministic drift for any one implementation.

Phenomenon	Representative papers	DCR lens
Entropy sensitivity	Cui et al. (2025); Cheng et al. (2026)	is consistent with entropic damping needing to offset selective score gaps before the trajectory enters a near-boundary low-entropy regime
Large-k degradation	Yue et al. (2025); Dang et al. (2025); He et al. (2025); Zhao et al. (2025)	scalar objectives can sharpen the distribution inside the correct set and can fail to preserve rare correct strategies
Temperature fails to restore diversity	Yun et al. (2025); Murthy et al. (2025)	post-hoc decoding rescales logits but need not undo training-induced pairwise ratio changes
Entropy-shaped updates help but remain incomplete	Cheng et al. (2026); Cui et al. (2025)	entropy slows concentration, but without a kernel it does not distinguish paraphrases from distinct strategies
Reward–entropy coupling	Cui et al. (2025)	reward gains and entropy loss can arise from the same concentration dynamics under scalar post-training

Table 2: **From theory to recent observations.** The cited papers are not assumed by our analysis; they are external empirical phenomena that the DCR framework helps organize mechanistically.

F CONNECTING THEORY TO PRACTICE

This appendix gives a trace-space account of how the mechanisms isolated in the main text and Appendices A–E relate to several empirical phenomena reported in recent work on RLVR and alignment. The goal is not to claim a one-to-one identification between our stylized trace-level models and any single large-scale implementation. Rather, the point is mechanistic: the same pairwise-ratio geometry that drives our theorems offers a lens for interpreting why several apparently different empirical failure modes recur across settings.

Scope. These comparisons are structural. Our theory works on finite trace spaces and simplified trace-level surrogates, whereas the cited works use token-level objectives, large parametric models, and concrete optimizers. The comparison is therefore at the level of distributional effect: which ratios are amplified, damped, tethered, or semantically repelled.

F.1 Why entropy control can look knife-edge

Recent empirical work, especially Cui et al. (2025), shows that plain entropy control can be highly coefficient-sensitive: small coefficients barely change training, while slightly larger ones can over-broaden the policy or destabilize learning. Cheng et al. (2026) further show that plain entropy regularization can induce unstable entropy dynamics, motivating more structured entropy-shaped updates. Our framework offers a mechanistic lens on this behavior through the universal pairwise log-ratio law. Under scalar training,

$$\dot{z}_{ij} = \Delta\psi_{ij}(p) - \varepsilon z_{ij},$$

and under DCR,

$$\dot{z}_{ij} = \Delta\phi_{ij}(p) - A z_{ij}.$$

In both cases, entropy enters as linear damping against task-induced score differences. If the effective entropic weight is too small relative to the score gaps encountered along training, damping does not act strongly enough before the trajectory moves into a low-entropy near-boundary regime. On trimmed simplices, Theorem A.4 makes this competition explicit through the comparison between inward entropic drift and maximal outward score pressure: when the barrier dominates that outward pressure, the flow remains inside the chosen trimmed simplex. The result is not a claim of a literal discontinuous phase transition for the exact regularized flow, which retains an interior maximizer whenever $A > 0$, but it is consistent with the threshold-like sensitivity to coefficient choices seen empirically.

F.2 Large-k degradation, distribution sharpening, and tail erosion

A recurring empirical pattern is that RLVR improves pass@1 while harming large-k performance. Yue et al. (2025) and Dang et al. (2025) report that RLVR-trained models can outperform their base models at small k yet underperform them at large k, and He et al. (2025) identify in GRPO a rank bias that reinforces already probable correct proofs while neglecting rare ones. This is consistent with the structural failure mode analyzed by Theorem 4.1. Large-k performance depends on the tail of correct traces, but STaR and exact mean-field GRPO amplify whichever correct trace is already larger, while DPO regresses ratios toward reference-relative values. None of these scalar objectives creates a force that protects rare but semantically distinct correct strategies for being distinct. Hence pass@1 can rise even as the tail needed for large-k sampling is depleted. The same concentration picture is also consistent with the amplification-of-pretraining-distributions perspective of Zhao et al. (2025). Appendix G, especially the toy suite and Figure 4, illustrates the same geometry across the synthetic studies in the appendix: STaR fixates, the DPO-style toy surrogate regresses pairwise ratios toward reference-relative values, the GRPO-style two-block variance-depletion surrogate shows slower path-dependent concentration, and DCR alone retains support across multiple correct clusters.

F.3 Why post-hoc temperature is not a substitute for online diversity preservation

Recent work finds that diversity collapse can persist even under high-temperature sampling (Yun et al., 2025), and that prompt-based perturbations can have a larger effect on conceptual diversity than temperature manipulations alone (Murthy et al., 2025). DCR helps explain why post-hoc temperature is limited. Temperature rescales logits at inference time, but the training dynamics change the pairwise log-ratios themselves. Once post-training has driven many correct–correct or correct–incorrect ratios far from their base values, test-time temperature can flatten the resulting distribution only around those new ratios; it cannot in general reconstruct the pre-collapse portfolio of strategies. In this framework, preserving diversity is primarily an online training problem, not merely a decoding problem.

F.4 Why entropy-shaped updates help, but do not solve semantic redundancy

Cheng et al. (2026) show that augmenting the advantage with an entropy-based term improves pass@K, especially at large K. At the level of mechanism, this aligns with the entropic damping term in our probability-space dynamics: entropy opposes rapid concentration and can keep more of the tail alive. The same logic helps interpret covariance-aware entropy control methods such as those studied by Cui et al. (2025). Our analysis also clarifies the limit of these approaches. Entropy is blind to whether two traces are paraphrases or genuinely different strategies. It spreads mass, but it does not add semantic repulsion. This is why Section 6 and Appendix C go beyond entropy-only control and add a gated kernel term that acts directly on strategic redundancy among verified-correct traces.

F.5 Why reward and entropy often track each other

Cui et al. (2025) report a tight empirical relation between reward and entropy throughout training. This coupling is natural at the trace-space level. In binary-verification regimes, reward is largely controlled by the total correct mass $\rho(p)$, while entropy is controlled by both the correct/incorrect split and the concentration of mass within the correct set. Under scalar post-training, these quantities are driven by the same concentration mechanism: probability moves toward a smaller set of high-scoring traces. We therefore interpret tightly linked reward gains and entropy loss as a consequence of the same sharpening dynamics, while remaining agnostic about any universal closed-form curve. What this theory does *not* claim is that one universal functional relation must hold across all algorithms, models, and tasks; the exact empirical fit is problem-dependent.

F.6 Design lessons for future algorithms

Viewed through DCR, recent positive results from methods that reward rare correct solutions or encourage exploratory outcomes can be interpreted as moving training away from pure scalar sharpening by preserving or reintroducing support for underrepresented successful traces (He et al., 2025; Song et al., 2025). This theory suggests three concrete design rules. First, keep a strictly positive interior barrier so that pairwise ratios remain well defined and near-boundary collapse is controllable. Second, if diversity pressure is used, gate it to verified-

correct traces so that exploration is not paid for with incorrect mass. Third, build the kernel on strategy-level features rather than surface wording, so that the regularizer repels redundant concepts rather than merely spreading paraphrases. Under the formal conditions of Theorem A.7, together with the gated-kernel setup analyzed in Appendix C, the DCR flow has a unique globally attracting interior equilibrium. Appendix G shows a broad safe region, while the held-out real-math study indicates that gated and ungated variants are tied at reporting precision in that fixed-bank regime.

Takeaway. These comparisons suggest that DCR does more than restate that diversity matters. It identifies the missing relational geometry behind several empirical failure modes and helps explain why fixes based only on entropy or KL can help but remain incomplete. The practical implication is that post-training objectives should be designed around the distribution *inside the correct set*, not only around scalar correctness.

G EXPERIMENTAL RESULTS

This appendix collects the empirical evidence for Distributional Creative Reasoning (DCR). It combines a theory-faithful toy finite-simplex suite, supporting synthetic diagnostics, a held-out same-support real-math study on MATH-500 (Lightman et al., 2023), a controlled symbolic strategy study on Game of 24 and Countdown, and the ReasoningTrap diagnostic (Jang et al., 2025). The implementation and configuration files are available at `https://github.com/maxruizluyten/creative_reasoning_release`. The appendix proceeds in stages: the shared protocol fixes scope and reproducibility, the synthetic sections test the mechanism directly, held-out MATH-500 provides the main real-task result, the symbolic section gives an exact-label comparison, and ReasoningTrap serves as a secondary robustness check.

G.1 Shared Protocol and Reproducibility

Hardware, seeds, and scope. All reportable runs use one `NVIDIA RTX 6000` with 49 GB VRAM and the fixed seed set $\{101, 202, 303, 404, 505\}$. The non-toy sections are *fixed-bank same-support reweighting studies*, not end-to-end online post-training runs: for each prompt, every method optimizes a distribution over the same finite bank of sampled traces and is evaluated on that shared support. The non-toy claims therefore concern redistribution over *discovered* candidates, not discovery of new strategies outside the sampled bank. All reported paired intervals use 1000-resample percentile bootstraps, with prompts or problem instances as the resampling unit rather than individual traces.

Common generator and prompting. All real-text generation experiments use `Qwen/Qwen2.5-Math-1.5B` (Qwen Team, 2024) as the common base generator. The shared completion template instructs the model to solve the problem step by step and place the final answer once at the end as `\boxed{...}`. For MATH-500 and ReasoningTrap bank generation, decoding uses `temperature=0.8`, `top_p=0.95`, `do_sample=True`, and `max_new_tokens=1024`. For the symbolic tasks, generation uses the same `temperature=0.8`, `top_p=0.95`, and `do_sample=True`, with `max_new_tokens=256`; when public corpora are available, those corpora are reused directly rather than regenerated. The symbolic corpora come from the external `llm-reasoning-uft` release of Ni et al. (2025), so the Game of 24 and Countdown comparisons are anchored to a public source rather than to a private regenerated bank.

Same-support bank construction. For each prompt x and each seed, a finite trace bank

$$\mathcal{S}_x = \{\tau_1, \ldots, \tau_{n_x}\}$$

is constructed once and then shared by every method. The bank builder draws 32 raw traces, normalizes and deduplicates them, and if fewer than 8 unique traces remain, draws additional batches of 8 raw traces up to a hard cap of 64 total raw samples.

Verification and strategy labels. The synthetic studies use oracle correctness and oracle strategy labels by construction. The symbolic studies use exact symbolic verifiers: a trace is correct if and only if its final expression uses the supplied numbers legally and evaluates exactly to the target, and strategy labels are canonicalized SymPy expression trees (Meurer et al., 2017). The real-math and ReasoningTrap studies use boxed-answer extraction plus symbolic or numeric normalization, together with clustering of verified-correct reasoning bodies

using `Qwen/Qwen3-Embedding-0.6B` (Qwen Team, 2025). The final clustering threshold is selected once on a 50-prompt calibration subset by maximizing penalized silhouette (Rousseeuw, 1987),

$$\text{score}(\tau) = \text{mean silhouette}(\tau) \times \text{scorable bank fraction}(\tau),$$

over $\tau \in \{0.05, 0.075, 0.10, 0.125, 0.15, 0.175, 0.20\}$. The selected rule uses $\tau = 0.05$.

Reported metrics. For each prompt, let q_x denote the induced distribution over verified correct strategy labels after aggregating trace-level mass inside the prompt's bank.

- *Effective number of strategies:*
$$N_{\text{eff}}(x) := \exp\big(H[q_x]\big).$$
- *Semantic coverage at eight draws:*
$$\text{cov@8}(x) := \sum_c \Big(1 - (1 - q_{x,c})^8\Big),$$
the expected number of distinct verified-correct strategies observed in eight i.i.d. draws from the optimized distribution.
- *Safety margin:*
$$\min_{c \in \mathcal{C}_x} \Big(1 - 2\lambda\beta (K_{\text{eff}} p)_c\Big),$$
reported after optimizing the prompt-level distribution.
- *DCR KKT residual:* the maximum absolute stationarity residual after subtracting the best simplex Lagrange multiplier from the prompt-level first-order conditions of the DCR objective above.

For non-DCR baselines, the DCR KKT residual is reported only as a diagnostic distance to DCR stationarity, so it is not expected to vanish. All tables report prompt averages of these metrics unless stated otherwise.

Optimization and baselines. All same-support DCR runs optimize the binary-utility, no-KL objective from Appendix C,

$$\widetilde{J}[p] = \sum_i U_i p_i + AH[p] - \lambda\beta\, p^\top K_{\text{eff}} p,$$

with exponentiated mirror ascent from the uniform initialization, step size $\eta = 0.10$, at most 2000 steps, and early stopping when $\|p^{(t+1)} - p^{(t)}\|_1 < 10^{-8}$. We use $A = 0.05$ by default and choose $\lambda\beta$ through the safety-normalized scale

$$s := 2\lambda\beta \|K_{\text{eff}}\|_{1\to\infty},$$

with default $s = 0.25$. By Corollary C.3, this leaves the conservative global margin buffer $\eta_K = 1 - s$. The same-support comparison set is Utility, Entropy, DCR (Gated), DCR (Ungated), and DCR (Shuffled). In the toy suite we additionally report RFT, exact mean-field GRPO, and one-sided DPO.

Code and assets. The project source, environment files, experiment configurations, tests, and audit commands are available in `maxruizluyten/creative_reasoning_release`. Exact paper reproduction additionally depends on frozen release bundles and large external artifacts outside the code-only snapshot.

Project and reused-asset licenses. The project source is ours, and the reused external assets are public releases that we do not redistribute inside this repository. The corresponding public source pages list the following terms:

- `creative_reasoning_release` (project source): MIT license.
- `Qwen/Qwen2.5-Math-1.5B` (common base generator): Apache-2.0, as listed on the Hugging Face model card.
- `Qwen/Qwen3-Embedding-0.6B` (embedding model for clustering verified correct traces): Apache-2.0, as listed on the Hugging Face model card.
- `ReasoningTrap/ReasoningTrap` (evaluation pipeline): MIT license.

- `ReasoningTrap/AIME` (ConditionedMath AIME benchmark): MIT, as listed on the Hugging Face dataset page.
- `ReasoningTrap/MATH500` (ConditionedMath MATH500 benchmark): the public Hugging Face dataset page and README were available, but no explicit license was stated on the cited page.
- `twni2016/llm-reasoning-uft` (symbolic corpora source for Game of 24 and Countdown): Apache-2.0. That repository also links external dataset files, which we reuse without redistributing them here.
- `HuggingFaceH4/MATH-500` (held-out benchmark source page): the cited Hugging Face page names `openai/prm800k` as the source of the selected math splits. The Hugging Face MATH-500 page itself did not state a standalone license, while the linked `openai/prm800k` repository is MIT-licensed.

All reused assets here are public model, code, or benchmark releases rather than new human-subject data collected for this paper, and we do not redistribute personally identifying content.

G.2 Synthetic Mechanism Studies

The synthetic studies play two distinct roles. The toy suite is the main theory-faithful mechanism study: it directly tests the universal log-ratio identity, global convergence, and the geometry of the DCR equilibrium. The additional synthetic panels are supporting diagnostics that broaden the view of the same geometry.

Table 3: **Toy finite-simplex summary.** The toy suite isolates the semantic-coverage effect cleanly. DCR (Gated) raises coverage from 2.8540 (UTILITY) and 2.8533 (ENTROPY) to 2.8786 while keeping pass@8 = 1.0 and driving incorrect mass to numerical zero. The shuffled control is weaker, which shows that the gain depends on semantic alignment rather than generic flattening of the trace distribution.

Method	N_{eff}	cov@8	Incorrect Mass	pass@8	Safety Margin
DCR (Gated)	2.9923	2.8786	0.0000	1.0000	0.9125
DCR (Shuffled)	2.7889	2.7572	0.0000	1.0000	0.9527
DCR (Ungated)	2.9923	2.8786	0.0000	1.0000	0.9125
DPO	2.7416	2.7014	0.0222	1.0000	—
ENTROPY	2.9512	2.8533	0.0000	1.0000	1.0000
GRPO	2.9512	2.7975	0.0971	1.0000	—
RFT	2.9512	2.7635	0.1446	1.0000	—
UTILITY	2.9522	2.8540	0.0000	1.0000	1.0000

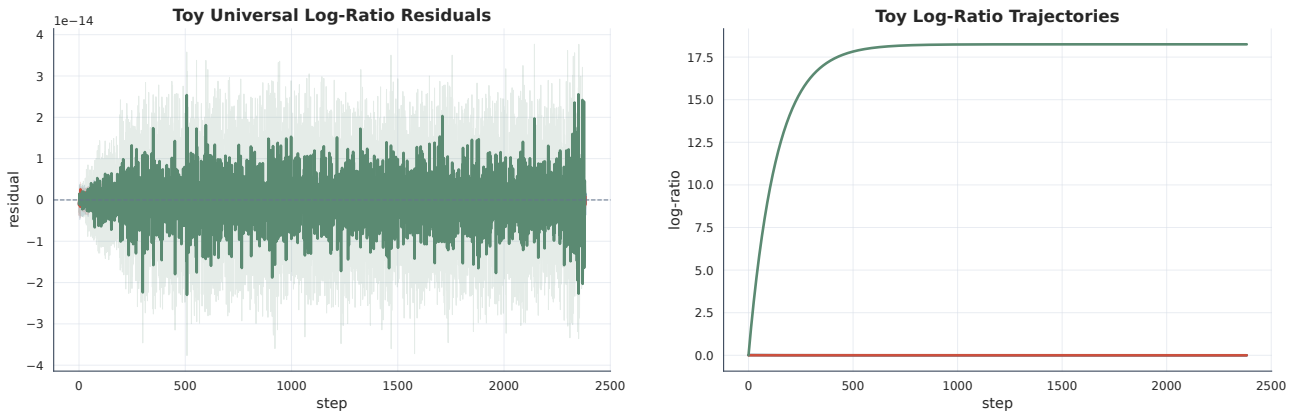

Figure 1: **Toy log-ratio diagnostics.** *Left:* residuals for the universal log-ratio identity of Theorem A.2. *Right:* representative same-lump, cross-lump, and correct-vs-incorrect log-ratio trajectories. At print scale, two trajectories lie nearly on top of the horizontal axis, so the panel is intended as a qualitative diagnostic rather than a precise color-decoding plot. The visible separation is consistent with the intended "homogenization within a strategy, repulsion across strategies" geometry.

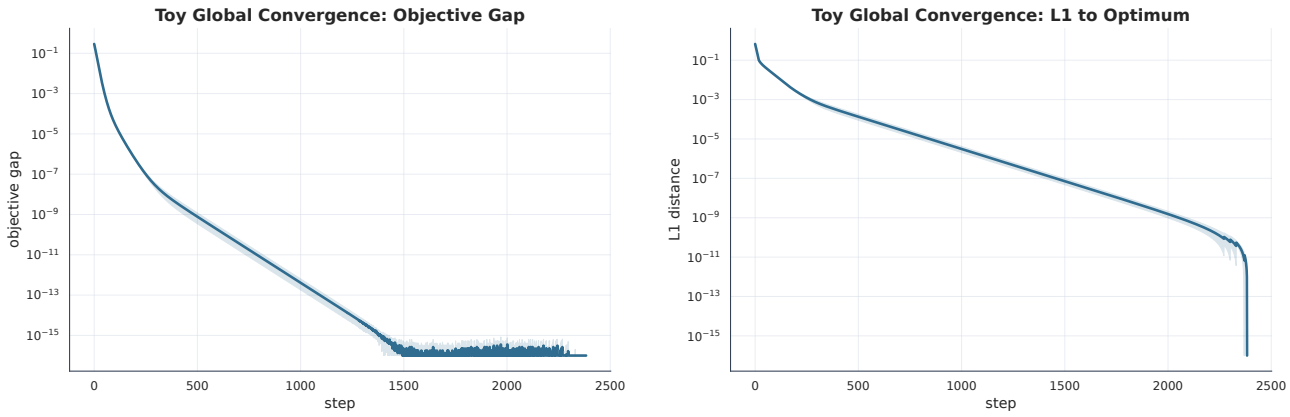

Figure 2: **Toy global-convergence diagnostics.** *Left:* semilog objective gap $\widetilde{J}(p^\star) - \widetilde{J}(p_t)$. *Right:* ℓ_1-distance to the optimizer. Both panels decay cleanly, providing the numerical counterpart of the strict-concavity and global-convergence theory in Theorem A.7.

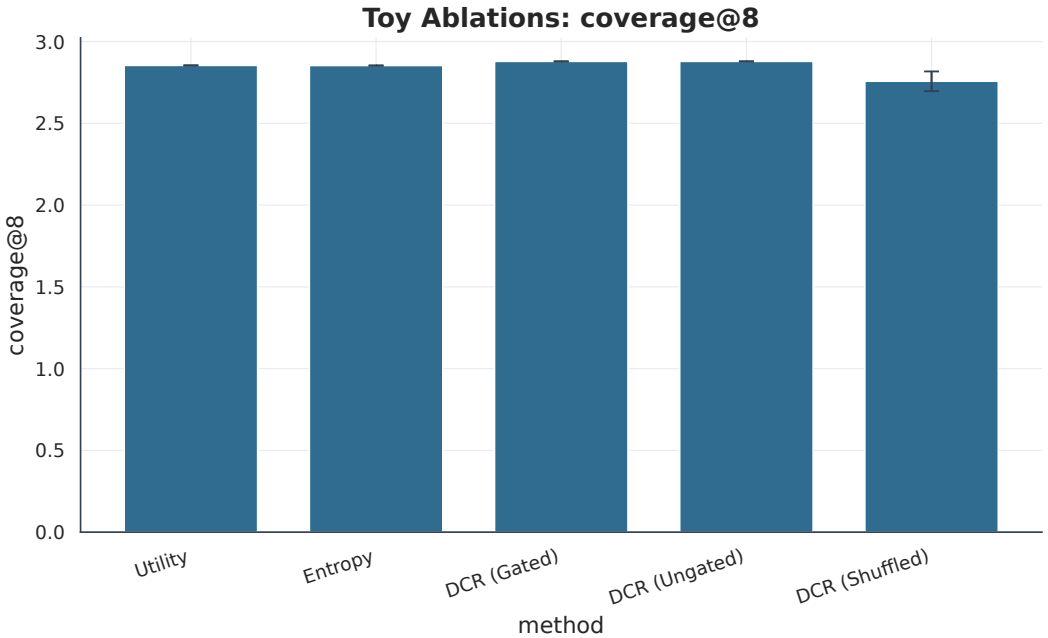

Figure 3: **Toy ablation summary.** Aligned DCR attains the best coverage trade-off, the shuffled control is weaker, and the scalar baselines remain more vulnerable to redundant concentration even when pass@8 is unchanged. In this toy kernel, the ungated and gated variants coincide numerically because the kernel already vanishes on the incorrect block.

Additional synthetic diagnostics. The appendix also includes a broader family of synthetic panels that examine the same mechanisms from complementary angles. The scalar-objective dynamics in Figure 4 and the phase diagrams in Figure 8 give the broadest synthetic view of the geometry. The overlay, alignment, ablation, and safety plots provide complementary diagnostics for the same regime.

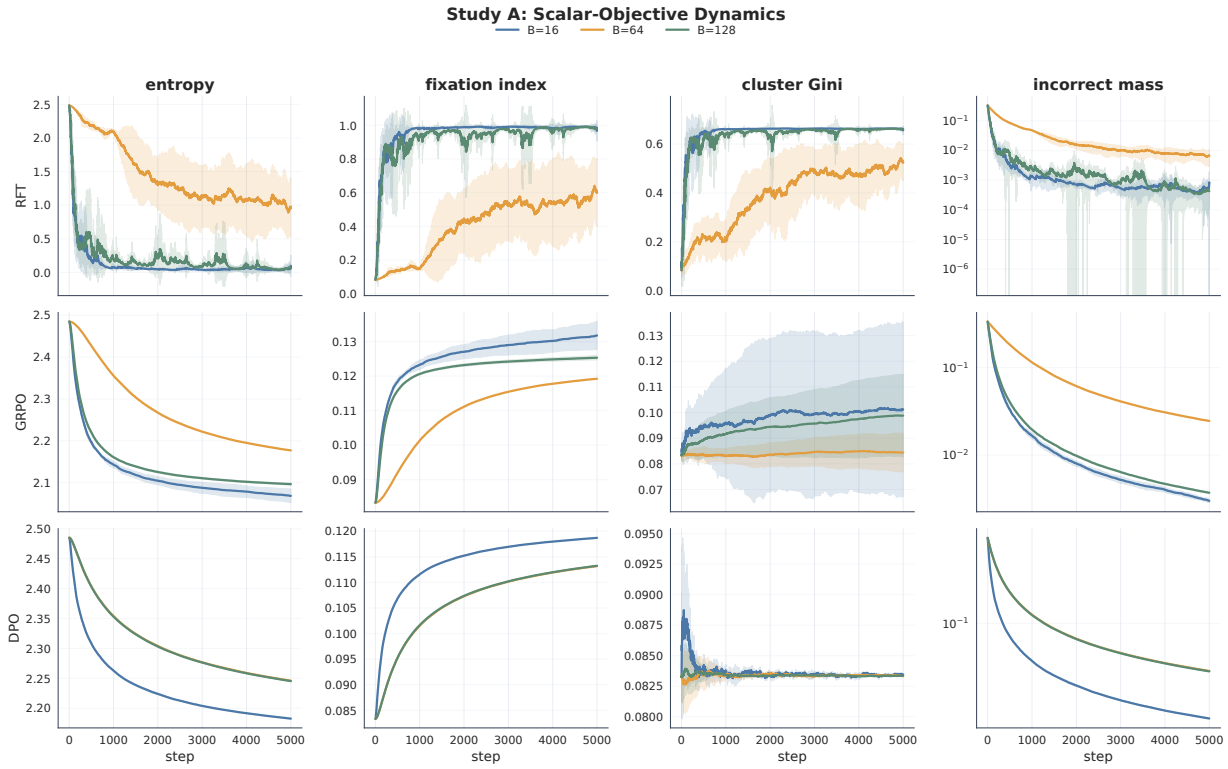

Figure 4: **Additional synthetic diagnostic: scalar dynamics.** Across representative synthetic settings, the scalar baselines drift toward concentration with distinct patterns: RFT fixates fastest, GRPO shows slower path-dependent concentration, and DPO regresses toward its reference-relative state.

Synthetic summary. The synthetic evidence supports three points. First, the toy suite directly renders the DCR mechanism visible: pairwise log-ratio geometry, global convergence, and semantic-coverage gains all appear in the intended regime. Second, the toy ablation shows that the gain is semantic rather than a generic preference for flatter distributions. Third, the supporting synthetic panels show that the same geometry occupies a broad safe region rather than a single fragile operating point.

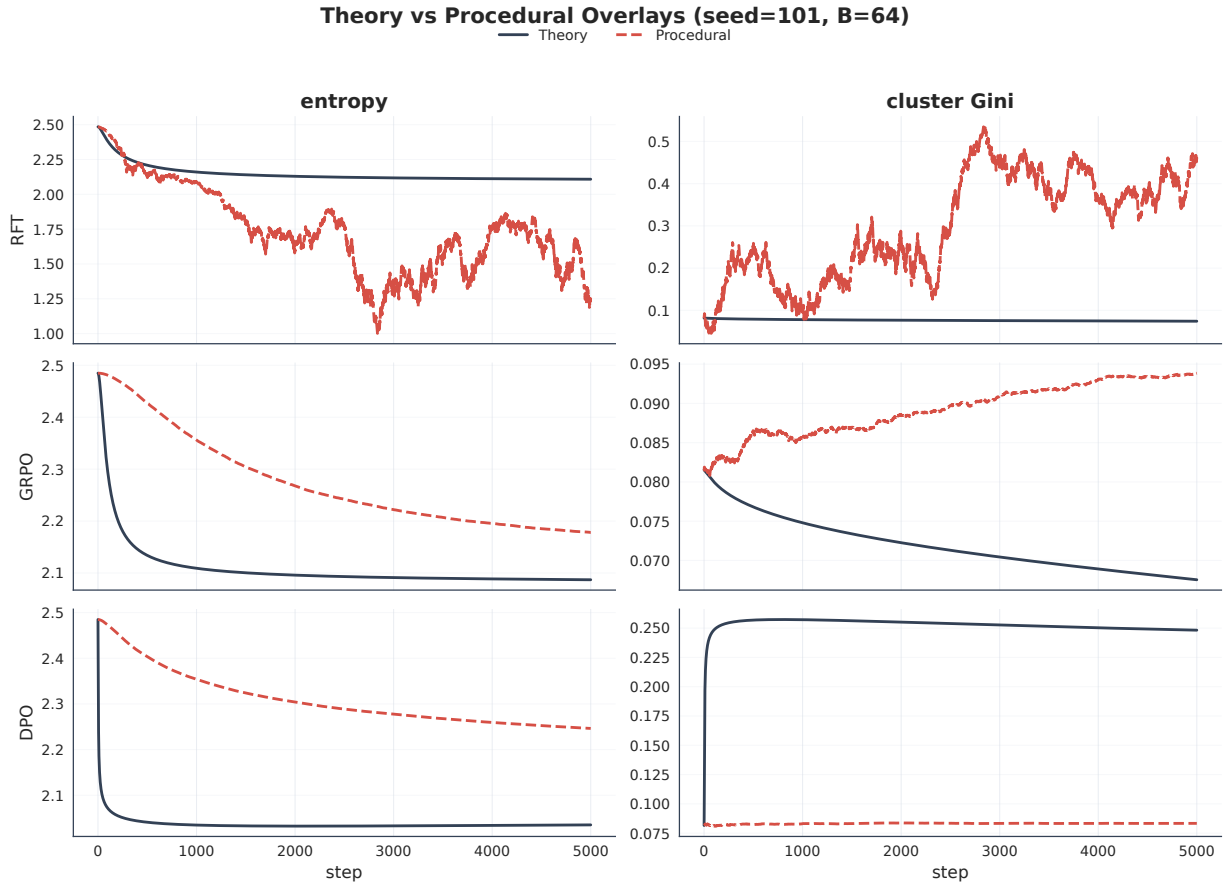

Figure 5: **Additional synthetic diagnostic: theory-versus-procedural overlays.** The procedural surrogate preserves the same event ordering as the idealized flow, so the panel remains informative about ordering even when the trajectories do not match pointwise.

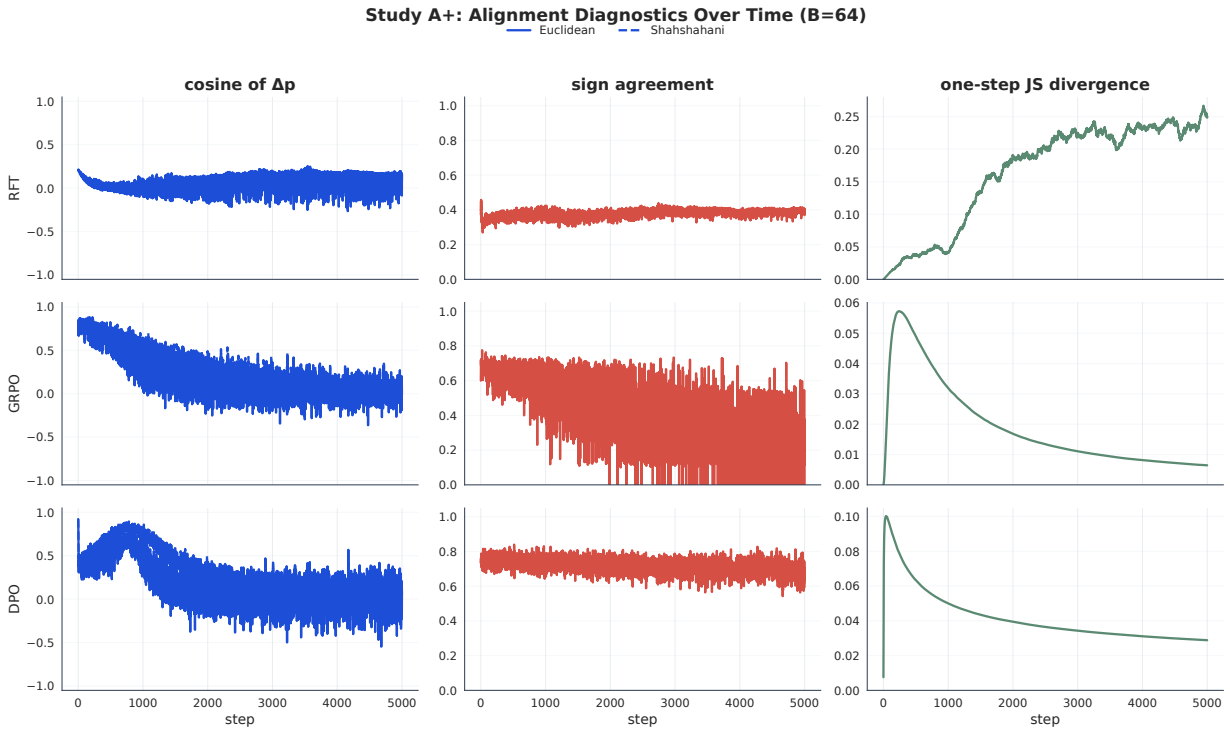

Figure 6: **Additional synthetic diagnostic: alignment-over-time trajectories.** The procedural dynamics track the theoretical targets directionally over time, so the surrogate preserves the intended drift geometry rather than reversing it. Columns report cosine of Δp, sign agreement, and one-step JS divergence. In the cosine column, the solid blue trace is the Euclidean reference and the dashed blue trace is the Shahshahani reference.

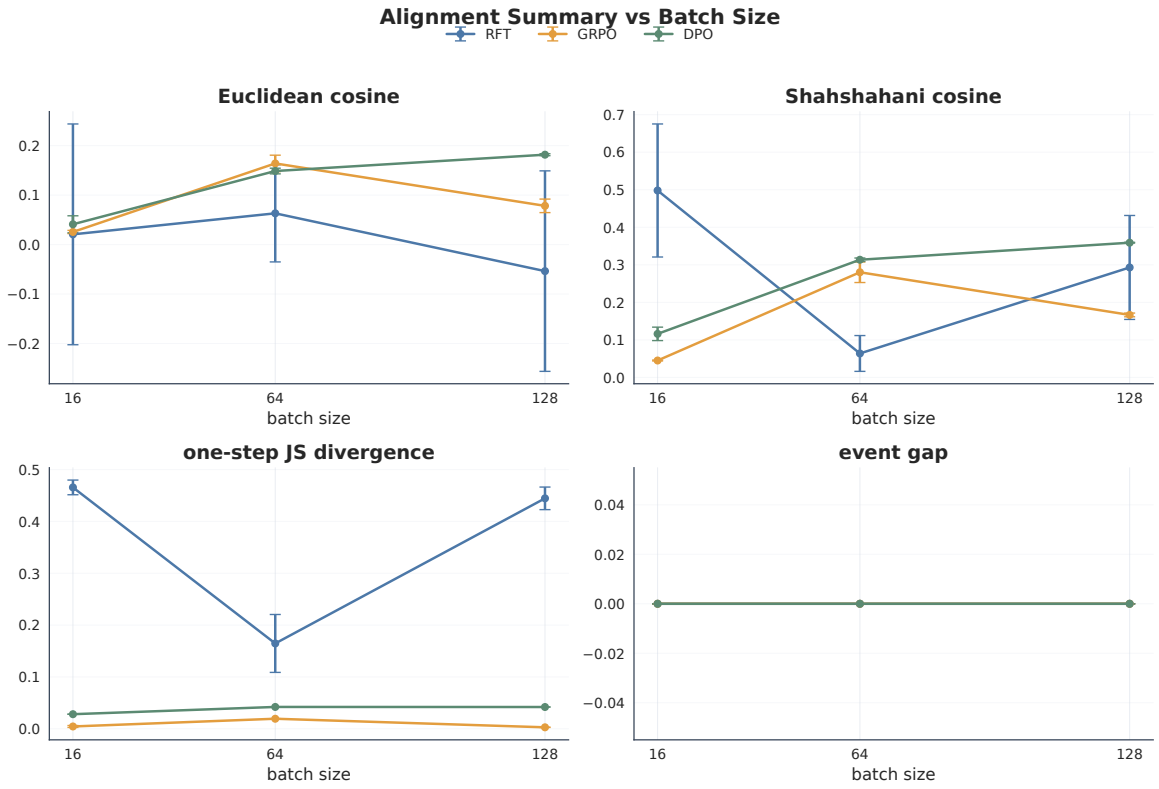

Figure 7: **Additional synthetic diagnostic: alignment summary.** Across batch sizes, the procedural surrogate preserves the qualitative method ranking.

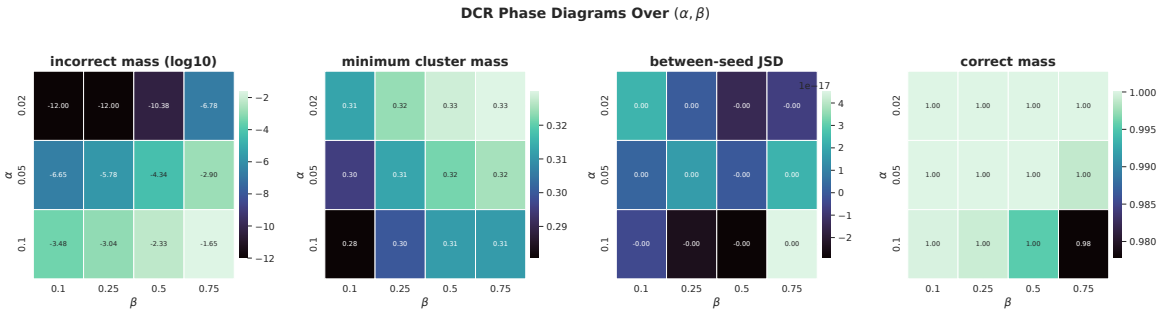

Figure 8: **Additional synthetic diagnostic: phase diagrams.** The main takeaway is a broad DCR band with essentially zero incorrect mass, high correct mass, and near-zero between-seed JSD, rather than a single tuned operating point.

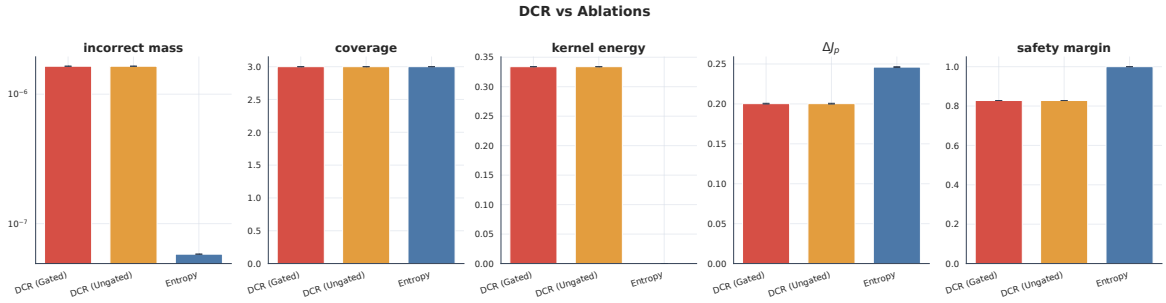

Figure 9: **Additional synthetic diagnostic: ablation comparison.** At the selected center point, all three methods keep incorrect mass numerically negligible and remain in the positive-safety regime. Relative to the entropy-only baseline, the DCR variants induce nonzero kernel energy while leaving coverage nearly unchanged.

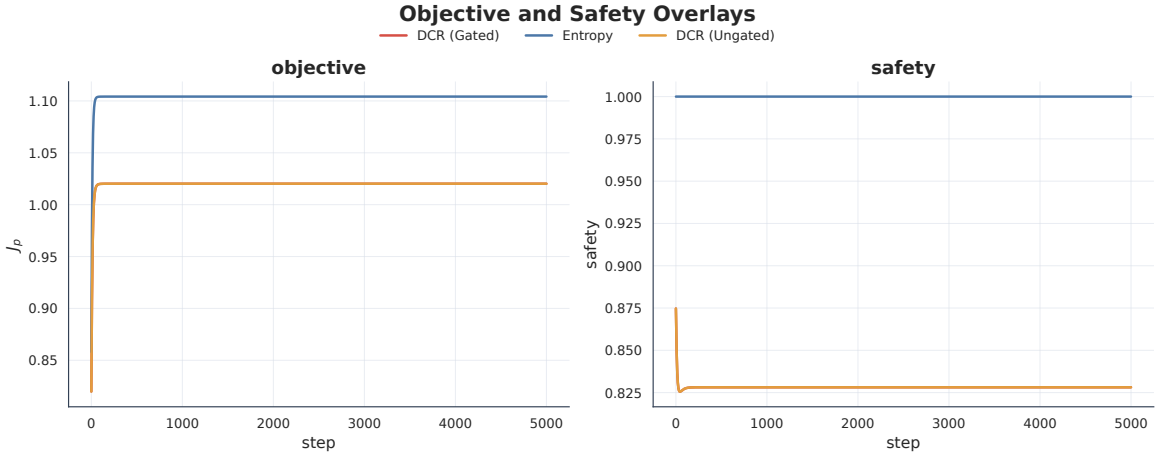

Figure 10: **Additional synthetic diagnostic: objective and safety overlay.** At the selected center point, ENTROPY reaches the highest plotted method-specific score and the largest safety margin, while the DCR variants remain safely inside the positive-margin regime. Because the left panel traces each method under its own score rather than a shared cross-method objective, this figure is best read as a local optimization-and-safety diagnostic.

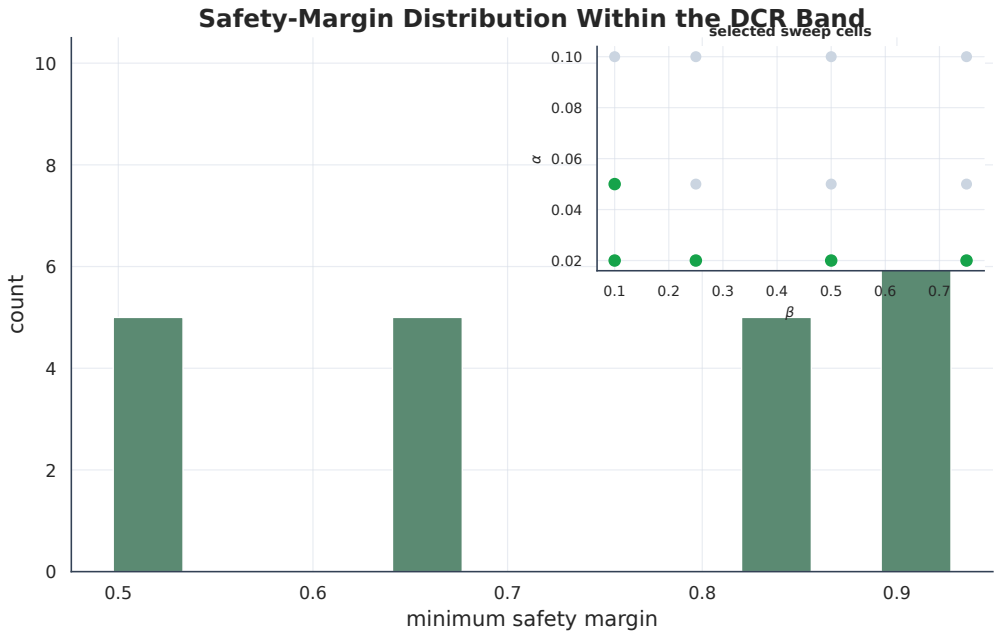

Figure 11: **Additional synthetic diagnostic: safety-margin distribution.** The full selected DCR band stays strictly positive in safety, with minimum margin approximately 0.497, so the synthetic gains are not achieved by crossing into net encouragement of incorrect traces.

G.3 Same-Support Real-Math Evaluation

The held-out MATH-500 study is the main real-task result in the paper. In this finite-bank setting, bank solvability leaves little room for further movement in pass@8, so the meaningful empirical movement is in semantic coverage: DCR redistributes probability mass across semantically distinct correct strategies while leaving answer accuracy fixed. This is the setting that the semantic coverage metric is designed to summarize.

Table 4: **Held-out MATH-500 all-prompt summary.** All methods attain the same pass@8 = 0.8912, so the main effect is not a change in answer correctness. The semantic benefit appears in the multi-solution slice and in the coverage-oriented figures below; safety margins for the DCR variants remain comfortably positive.

Method	pass@8	Incorrect Mass	Safety Margin
Utility	0.8912	0.1088	1.0000
Entropy	0.8912	0.1088	1.0000
DCR (Gated)	0.8912	0.1088	0.7553
DCR (Ungated)	0.8912	0.1088	0.7553
DCR (Shuffled)	0.8912	0.1088	0.7561

Table 4 summarizes the full 500-prompt held-out set, whereas Table 5 isolates the 433-prompt multi-solution slice.

Real-math interpretation. The paired bootstrap shows the same pattern. On the full 500-prompt held-out set, DCR (Gated) improves coverage@8 relative to both Utility and Entropy by approximately +0.0512, with percentile bootstrap interval $[0.0464, 0.0562]$, while the held-out pass@8 difference is exactly zero. These are paired full-set prompt-level differences, not values read directly from the displayed tables. Table 5 separately reports the raw 433-prompt multi-solution slice means, where coverage@8 increases from 2.4662 to 2.5253. The empirical claim is therefore that DCR increases semantic-strategy coverage at fixed accuracy. The calibration split shows the same pattern under the threshold rule, which reduces concern that the held-out result reflects a selection effect. These held-out results also do not support a meaningful advantage of DCR (Gated) over DCR (Ungated) on real math, since the two variants are effectively tied at reporting precision.

Table 5: **Held-out MATH-500 multi-solution slice.** The held-out run identifies 433/500 prompts as multi-solution under the frozen clustering rule. On this slice, DCR improves semantic coverage and the effective number of correct strategies, while the DCR KKT residuals remain near zero. This table is the main summary for whether DCR changes the distribution over correct reasoning strategies rather than raw correctness.

Method	cov@8	N_{eff}	DCR KKT Residual
UTILITY	2.4662	2.7883	0.9307
ENTROPY	2.4662	2.7883	0.2400
DCR (Gated)	2.5253	2.8784	0.0004
DCR (Ungated)	2.5253	2.8784	0.0027
DCR (Shuffled)	2.4616	2.7807	0.0027

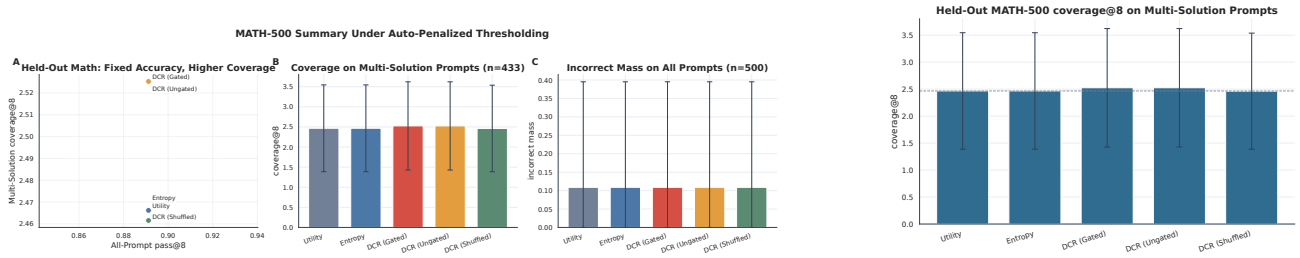

Figure 12: **Held-out MATH-500 overview and ablation view.** *Left composite:* the held-out summary, which separates accuracy from semantic coverage on the full 500-prompt set and shows the incorrect-mass baseline. *Right:* direct coverage comparison on the 433-prompt multi-solution slice. On that slice, DCR raises coverage@8 from 2.4662 (UTILITY and ENTROPY) to 2.5253 (DCR (Gated) and DCR (Ungated)) while all methods remain fixed at pass@8 = 0.8912 on the full set. The shuffled control is weaker, which supports the claim that the gain depends on semantic alignment rather than on a generic preference for flatter distributions.

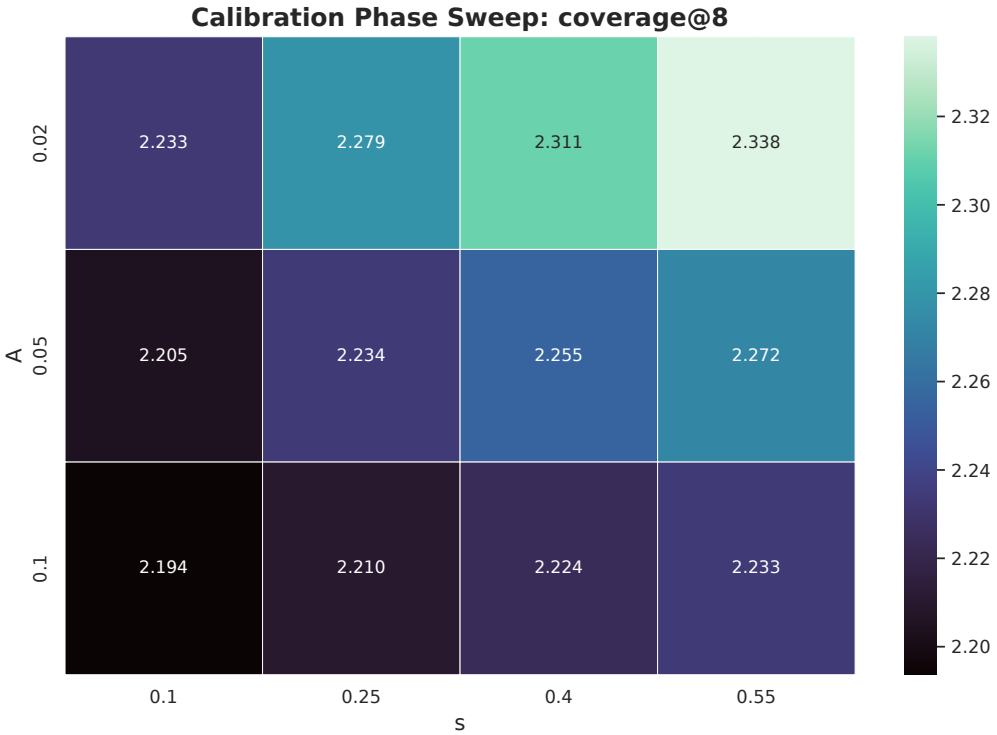

Figure 13: **Calibration phase sweep: coverage@8.** This phase plot shows that the meaningful movement across the (A, s) grid is in semantic coverage rather than in pass probability. On the calibration split, with the selected penalized-silhouette threshold, pass@8 stays at 0.89 while coverage@8 rises from 2.1825 for UTILITY/ ENTROPY to 2.2335 for DCR (Gated).

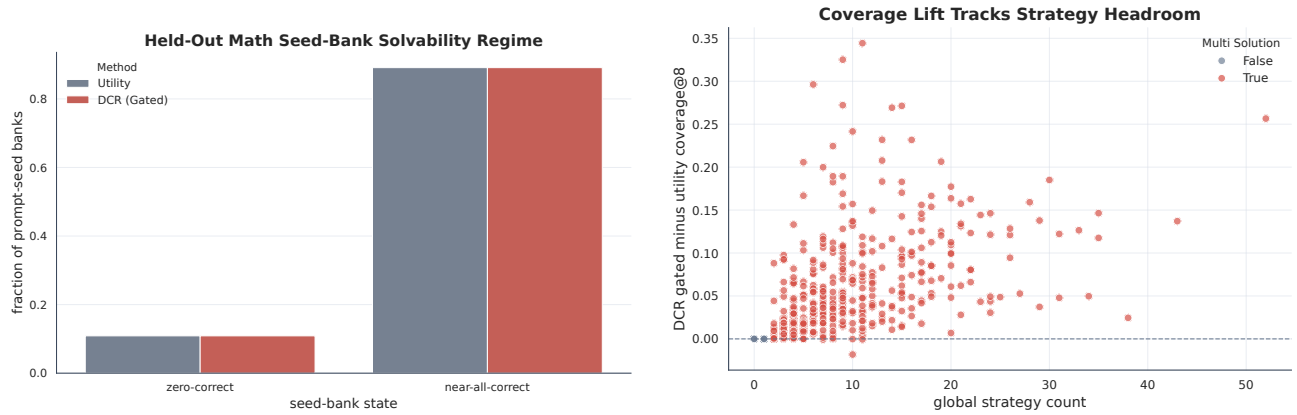

Figure 14: **Why coverage moves while pass remains fixed.** *Left:* a bank-solvability diagnostic showing that most prompt-seed banks are effectively either solved or unsolved, leaving little room for further movement in pass@8. *Right:* coverage lift versus discovered strategy count. The DCR gain is concentrated on prompts with semantic headroom, which is consistent with the intended mechanism.

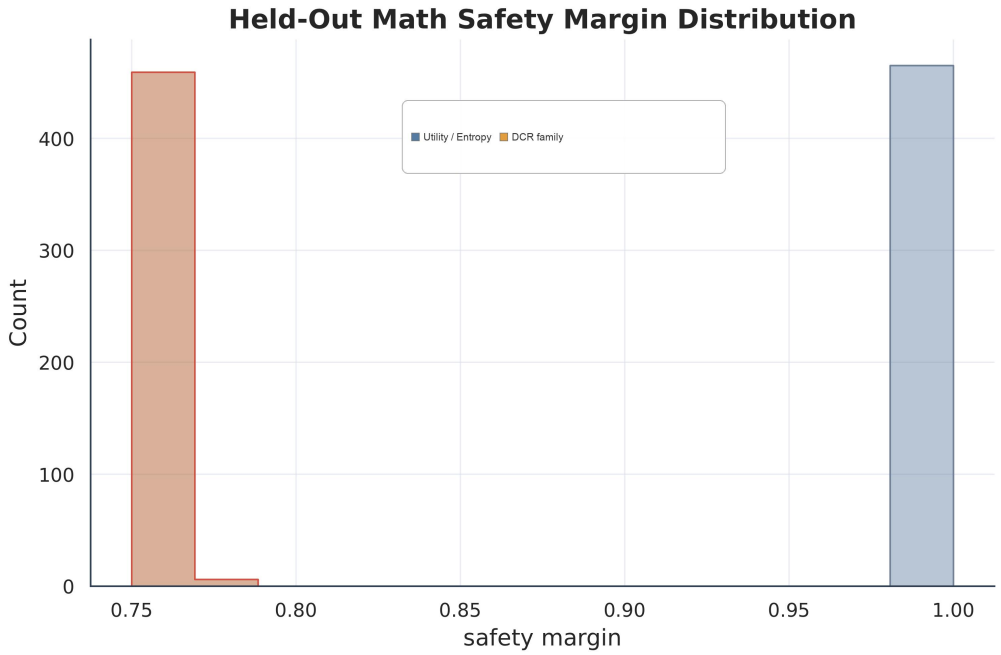

Figure 15: **Real-math safety-margin distribution.** The held-out DCR solutions remain comfortably inside the no-incorrect-reward regime. The orange spike near 0.755 is the DCR family, whereas the blue spike at 1.0 is the UTILITY/ENTROPY baseline. The semantic-coverage gain is therefore not bought by drifting into a regime where incorrect traces receive net encouragement.

G.4 Controlled Symbolic Strategy Study

The symbolic study is an exact-label comparison in a non-toy setting because both correctness and strategy labels are exact. Its candidate banks are drawn from the public Game of 24 and Countdown corpora reused from the external `llm-reasoning-uft` release of Ni et al. (2025), so the pipeline evaluates a public source rather than a privately regenerated bank. Unlike the real-math section, there is no need to approximate semantic strategies with embeddings before the oracle comparison can be stated. The resulting gain is not new-trace discovery; it is an exact-label test of how the optimized distribution reallocates mass across strategies already present in the bank.

Table 6: **Symbolic strategy study on Game of 24 and Countdown.** Every method attains pass@8 = 1.0, so the entire comparison is about strategy coverage rather than correctness. The oracle kernel produces the largest gains, as expected when semantic geometry is captured without approximation.

Task	Method	Case Count	cov@8	N_{eff}
Game of 24	DCR (Embedding)	250	2.5105	2.6134
Game of 24	DCR (Oracle)	250	2.9189	3.2431
Game of 24	DCR (Shuffled)	250	2.4654	2.5532
Game of 24	DCR (Ungated)	250	2.5105	2.6134
Game of 24	Entropy	250	2.4681	2.5568
Game of 24	Utility	250	2.4681	2.5568
Countdown	DCR (Embedding)	250	2.0901	2.0651
Countdown	DCR (Oracle)	250	2.1314	2.1535
Countdown	DCR (Shuffled)	250	2.0835	2.0535
Countdown	DCR (Ungated)	250	2.0901	2.0651
Countdown	Entropy	250	2.0840	2.0543
Countdown	Utility	250	2.0840	2.0543

Symbolic interpretation. The symbolic bootstrap comparisons reinforce the same story, but the comparison is informative because oracle labels remove the main approximation layer from real math. On Game of 24, the oracle kernel improves coverage@8 over Utility by about +0.451; on Countdown the corresponding gain is +0.0474. The oracle–embedding gap therefore measures how much is lost when semantic geometry must be inferred rather than specified exactly. The symbolic section therefore provides an exact-label validation: exact semantic labels reveal large strategy-coverage gains at saturated pass.

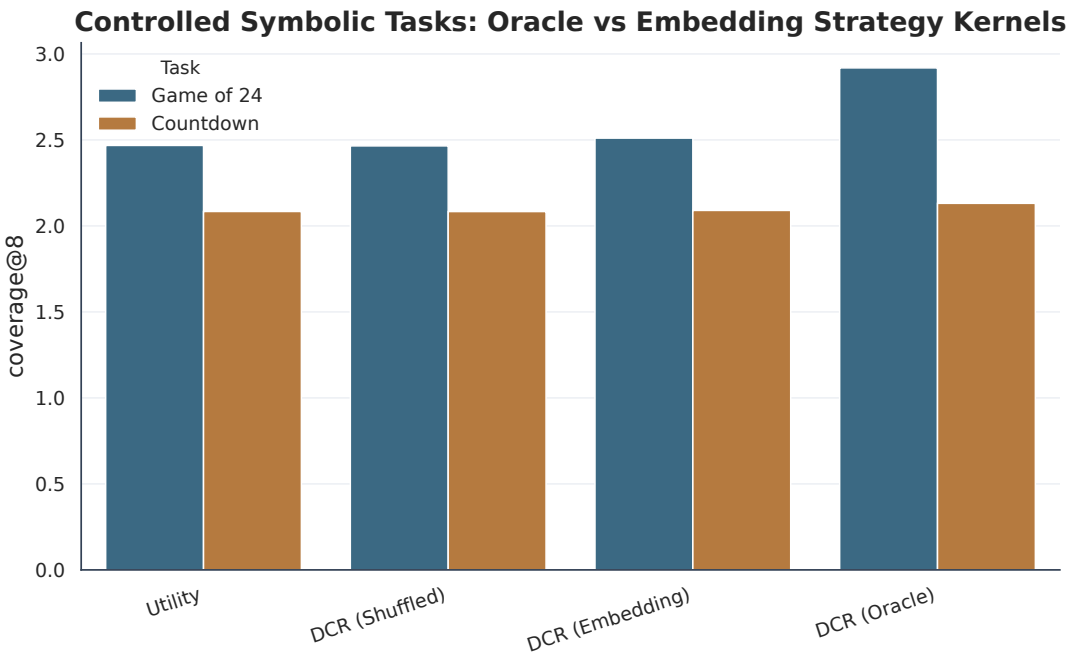

Figure 16: **Oracle-kernel versus embedding-kernel DCR in the symbolic study.** This figure compares the oracle and embedding kernels in a non-toy setting. Because symbolic strategy labels are exact, the oracle kernel can directly target semantic diversity rather than an embedding approximation. The resulting coverage gains are largest on Game of 24, where DCR (Oracle) increases coverage@8 from approximately 2.468 to 2.919 at unchanged pass@8 = 1.0.

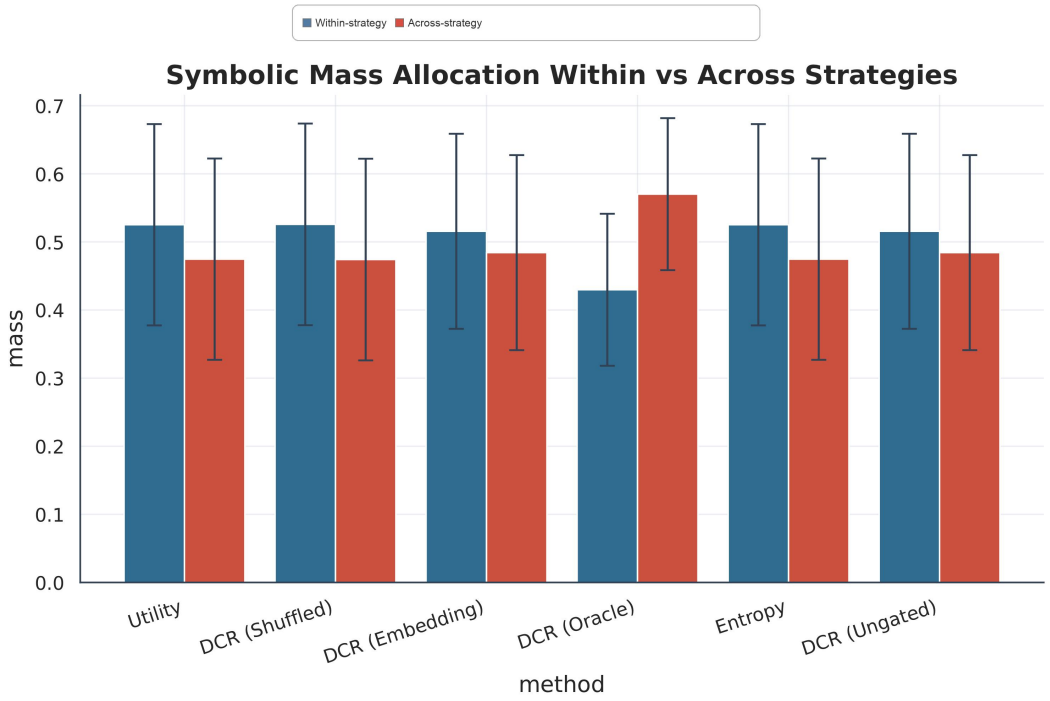

Figure 17: **Symbolic coverage geometry.** Because every method already attains pass@8 = 1, the meaningful difference is how mass is distributed across correct strategies. The oracle variant reduces redundant within-strategy concentration and increases coverage across distinct strategies. Blue bars denote within-strategy mass and orange bars denote across-strategy mass.

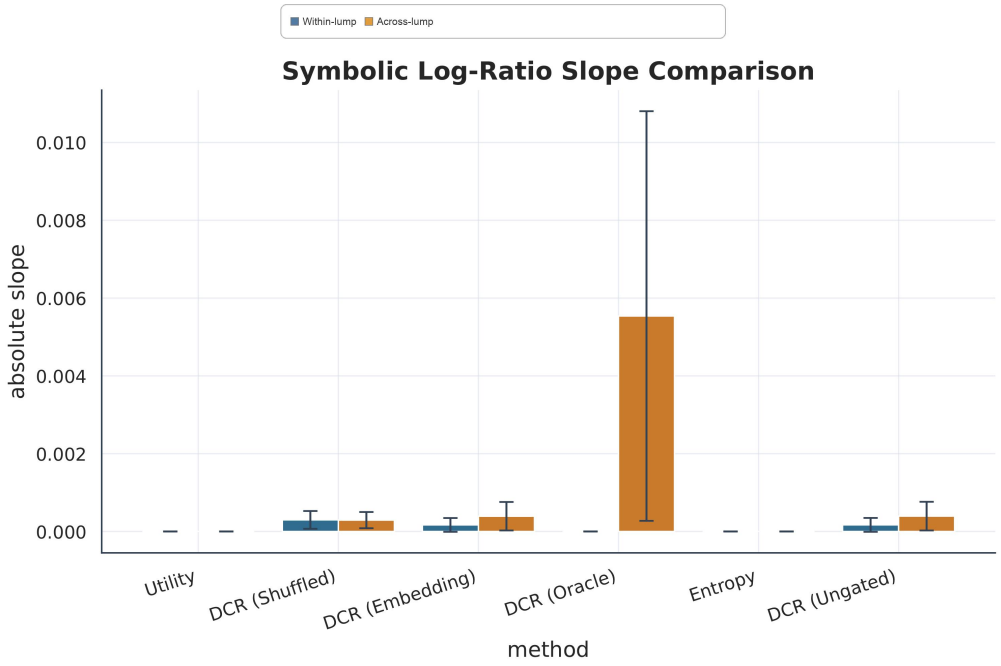

Figure 18: **Within-lump versus across-lump slope comparison in the symbolic study.** This panel connects the symbolic empirical results back to the DCR log-ratio theory. In the oracle setting, within-lump slopes are near zero while across-lump slopes remain positive, which is the expected pattern from Theorem C.8. Blue bars denote within-lump slopes and orange bars denote across-lump slopes.

G.5 ReasoningTrap Robustness Diagnostic

ReasoningTrap (Jang et al., 2025) serves as a secondary robustness diagnostic rather than the main empirical focus. The pipeline evaluates the public ReasoningTrap release together with the Hugging Face datasets `ReasoningTrap/MATH500` and `ReasoningTrap/AIME`. In this setup, the modified variants are easier than the originals and exhibit richer strategy headroom, which weakens the benchmark as a pure rigidity validator.

Table 7: **ReasoningTrap rigidity-gap summary.** All methods have the same negative rigidity_gap_pass@8, which means the modified public variants are easier than the originals. The semantic coverage gap is also negative for every method, so this release does not cleanly isolate the intended rigidity effect.

Method	Rigidity gap in pass@8	Coverage Gap
DCR (Gated)	-0.1905	-0.9249
DCR (Shuffled)	-0.1905	-0.9073
DCR (Ungated)	-0.1905	-0.9249
Entropy	-0.1905	-0.9136
Utility	-0.1905	-0.9136

ReasoningTrap interpretation. In this appendix, ReasoningTrap functions as a secondary robustness diagnostic rather than as the core empirical test. The evaluation uses all 84 public-release pairs: 50 MATH500-derived items and 34 AIME-derived items. The bootstrap comparison of DCR (Gated) against Utility on the coverage gap is only about -0.0113 and crosses zero, so these results do not isolate a DCR-specific failure or make ReasoningTrap the main empirical evidence for rigidity.

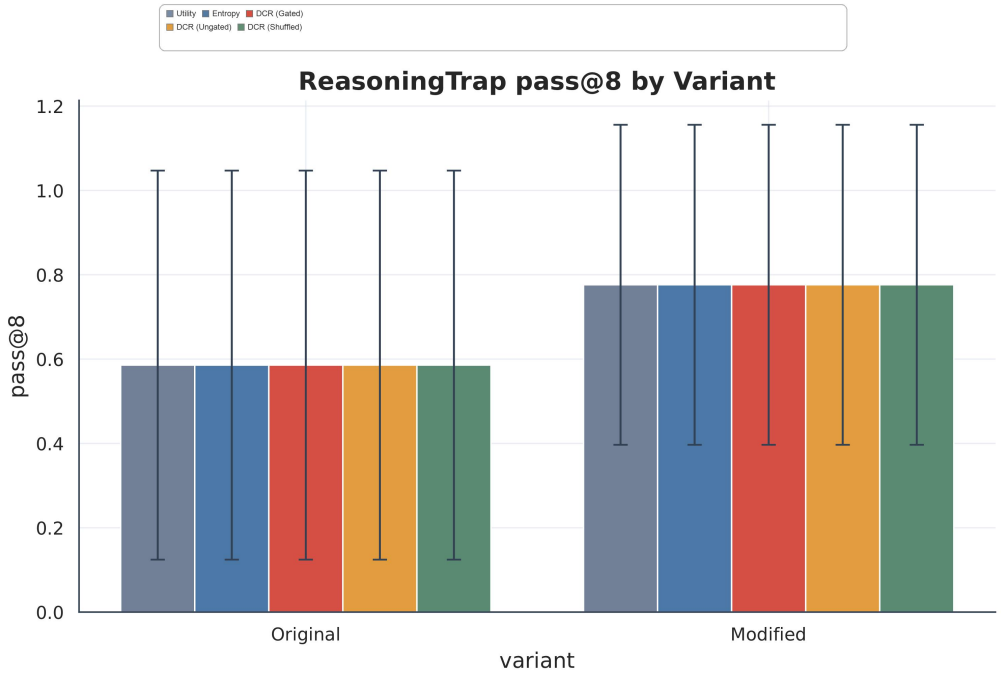

Figure 19: **ReasoningTrap pass shift by variant.** The modified variants are easier than the originals for every method. The negative rigidity gap is driven primarily by a benchmark shift, not by a DCR-specific drop in answer accuracy. Within each Original/Modified group, the five bars correspond to UTILITY, ENTROPY, DCR (Gated), DCR (Ungated), and DCR (Shuffled).

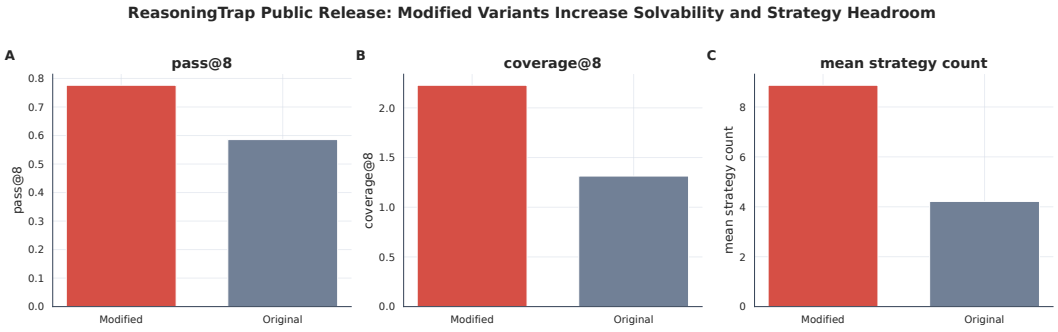

Figure 20: **ReasoningTrap headroom-shift diagnostic.** The modified variants have both higher pass and richer strategy structure than the originals. In this setup, the benchmark is informative but does not behave like a clean rigidity-only stress test.

G.6 Cross-Experiment Synthesis

Across the non-toy sections, the evidence is conditional on fixed candidate banks: it supports better redistribution over discovered correct strategies, not new-strategy discovery outside the sampled support.

1. The toy suite validates the core DCR mechanism directly: global convergence and repulsion across correct strategies both appear clearly.
2. The supporting synthetic diagnostics show that the same mechanism occupies a broad safe region rather than a single tuned point.
3. The held-out MATH-500 study is the main real-task support: DCR improves semantic-strategy coverage by roughly $+0.051$ at fixed pass@8 $= 0.8912$, and the real-math results do not support a meaningful advantage of DCR (Gated) over DCR (Ungated).
4. The symbolic study provides a non-toy mechanism validation: exact strategy labels remove embedding ambiguity and reveal the largest gains when correctness is already saturated.
5. The ReasoningTrap diagnostic is informative but mixed, so it is best viewed as a boundary case rather than as the main empirical support for rigidity.

Across fixed candidate banks, the empirical results support improved redistribution across semantically distinct correct strategies when raw correctness is already saturated.