BEYOND MAX-ENT TARGETS: DISENTANGLING ENTROPY AND DIVERSITY IN LLM POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Entropy-regularized trajectory-balance objectives have been proposed for LLM post-training as an alternative to GRPO-like objectives. By matching a distribution over trajectories rather than concentrating only on the highest-reward ones, they are expected to preserve diverse solutions, expand problem reachability, and improve accuracy on difficult tasks. We test these expectations under binary verifiable rewards. To expose the available target-design space, we expand the classical Shannon entropy with q-exponential targets in a common Tsallis family, used here as an analytical and experimental degree of freedom. We show that group-normalized binary rewards reduce every reward tilt to a single incorrect-to-correct reweighting profile $\rho(p)$. This reduction reveals a coupling between support and scale, compresses apparently different target constructions into a narrow region, and provides a direct criterion for choosing trajectory-balance parameters. In the experiments, we separate nominal target geometry from learned-policy behavior and evaluate the latter after quantifying procedural variation. Although our sweep changes the ρ-profile substantially, its behavioral effects reveal a clear separation between nominal target geometry and learned-policy properties: reachability and accuracy mostly remain within measured variation, while the strongest intervention reduces both. A matched GRPO comparison shows that Max-Ent objectives need not retain higher token entropy, even as an LLM judge reproduces the previously reported semantic-diversity effect. We thus identify nominal Max-Ent structure, token-level entropy, and semantic solution diversity as empirically distinct axes, showing that theoretical target changes cannot substitute for behavioral validation.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) has become the standard final stage of training large reasoning models (Guo et al., 2025; OpenAI et al., 2026). Widely used RLVR objectives (Schulman et al., 2017; Shao et al., 2024) optimize primarily for reward and may sharpen the policy onto a narrower set of trajectories, reducing the generation diversity that difficult problems may require (Cui et al., 2025; Chen et al., 2026). A natural response is Maximum-entropy RL (Max-Ent RL) (Ziebart et al., 2008; Haarnoja et al., 2018), where the entropy regularizer is added to the objective, so that the optimum is a distribution over trajectories rather than a single mode. Generative flow networks (Bengio et al., 2021; 2023) provide a convenient architecture for learning such a distribution: they sample terminal states in proportion to reward and are closely connected to entropy-regularized RL (Tiapkin et al., 2024; Mohammadpour et al., 2024), while the trajectory-balance criterion (Malkin et al., 2022) makes the target an explicit argument of the loss rather than an implicit consequence of a penalty. This is a route by which max-ent objectives have reached LLM fine-tuning (Hu et al., 2024; Zhu et al., 2026b). The Shannon entropy term can be extended with replacing it by the q-indexed Tsallis family(Tsallis, 1988; Lee et al., 2019), promising the more expressive target and diverse generations. These properties motivate growing interest in this paradigm as an alternative to conventional LLM post-training methods.

In contrast to purely reward-maximizing objectives, entropy-regularized objectives with a parametrized entropy regularizer make it possible to set different target distributions by varying their shape and a parameter that controls which low-relevance trajectories are filtered out of the support. Such targets are hypothesized to encourage more diverse generations (Zhu et al., 2026b) and,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in turn, improve accuracy and the ability to solve previously unreachable problems; suppressing poor trajectories may further concentrate sampling on useful regions.

In practice, we adopt the standard RLVR setting and inherit its constraints: a binary verifiable reward and group-normalized advantages computed over G trajectories per prompt (Shao et al., 2024). To realize this setting, we build on GFlowNets and train with the trajectory-balance objective (Bengio et al., 2021; 2023). We then ask whether these benefits materialize in RLVR and evaluate the resulting reasoning models.

The main contributions of this work can be summarized as follows:

1. **We characterize the targets** available to Max-Ent trajectory-balance objectives in binary-reward RLVR. Under group-normalized binary rewards, any multiplicative reward tilt takes only two values within a group. Its effect on the empirical target therefore reduces to a single reweighting ratio ρ between incorrect and correct responses, directly characterizing the nominal reward transform supplied to the optimizer.
2. **Target geometry is turned into a controlled experimental axis.** We express all compared objectives through their induced ρ-profiles and sweep this quantity across different targets. The Tsallis family is used here as a convenient extension of the objective that supplies an additional degree of freedom.
3. **We introduce a validation protocol matched to the resolution of RLVR post-training.** Before comparing objectives, we measure variation due to evaluation, initialization, seed, training regime, and chat-template convention. The protocol combines paired accuracy estimates, per-problem reachability at a fixed sampling budget, and explicit detection floors, separating changes attributable to an objective from changes below the resolution of the procedure.
4. **We show that the expected Max-Ent benefits do not follow from target variation alone.** Although the sweep substantially changes the nominal reweighting, behavioral outcomes do not track this change: reachability does not expand and accuracy does not improve consistently, and the strongest intervention degrades both. The Max-Ent checkpoints also show no token-entropy advantage over matched GRPO, while an LLM judge assigns their solutions greater semantic variety. Thus, semantic diversity, answer diversity, and policy entropy are distinct in practice, and no single criterion establishes a general claim of "richer diversity".

2 Related work

GFlowNets and Trajectory Balance. The framework of Bengio et al. (2021; 2023) was developed for sampling compositional objects, with trajectory balance introduced by Malkin et al. (2022), and adapted to LLM fine-tuning by Hu et al. (2024). Its theoretical connection to maximum-entropy RL (Tiapkin et al., 2024; Mohammadpour et al., 2024) motivates treating the target as an object of design. FlowRL (Zhu et al., 2026b) casts LLM post-training as distribution matching with an exponential target and provides the precedent for the claims about semantic diversity. We treat this result as a hypothesis to be tested, independently repeat the semantic evaluation with a stronger LLM judge, and complement it with matched answer-, text-, and token-level measurements.

Regularizers with generalized entropies. The Tsallis entropy family (Tsallis, 1988) gives a q-parameterized class of regularizers whose optima are not exponential; at $q = 2$ the induced policy is the sparsemax transformation (Martins & Astudillo, 2016), giving sparse MDPs with a constant performance-error bound (Lee et al., 2018; Chow et al., 2018), and Lee et al. (2019) unify these cases using a real-valued index. The same family recurs as a Bregman divergence in mirror-descent policy optimization (Tomar et al., 2020), as a q-deformed KL term (Zhu et al., 2023), and as q-exponential family policies (Zhu et al., 2025); convex regularizers and the prediction maps they induce are organized by the Fenchel–Young framework (Blondel et al., 2020) and by the theory of regularized MDPs (Geist et al., 2019). All of this is stated at the level of a per-step MDP with a rich action space. Our setting is the opposite extreme, and Section 4 shows it collapses the family to one scalar per group.

Group-relative objectives under binary rewards. GRPO (Shao et al., 2024) removes the value network by normalizing rewards within a sampled group, and underlies DeepSeek-R1 (Guo et al., 2025); DAPO (Yu et al., 2026) refines its clipping and sampling rules. The closed form induced by this normalization under binary rewards is known: Mroueh (2025) shows that GRPO is an adaptive weighted contrastive objective in which the two correctness classes receive weights determined by the group pass rate, and subsequent analyses adopt that form and examine its biases (Liu et al., 2025; Yang et al., 2026). Parallel lines treat the relative weight of positive and negative samples as a free hyperparameter (Zhu et al., 2026a; Arnal et al., 2026) or redesign the baseline (Wu et al., 2026; Xiao et al., 2025). Whereas that line treats the weight as a tunable quantity, we ask what freedom remains for target design.

Generalizing the loss versus generalizing the target. The nominal target, its theoretical matching loss, and the stabilized implementation are separate design layers. Trajectory balance admits different residual divergences (da Silva et al., 2024; Fawkes & Hartford, 2026), while the log-variance or VarGrad formulation removes the learned normalizer (Whitammer et al., 2026). FlowRL instead uses length-normalized sequence log-probabilities in a trajectory-balance residual (Zhu et al., 2026b). Related normalization choices are also known to change the effective optimization problem in policy-gradient and sequence-level importance-sampling objectives (Liu et al., 2025; Shen et al., 2026). Our theory varies the nominal target at a fixed practical loss; it does not assume that a finite, clipped training run reaches the nominal fixed point. Appendix A states the exact relation between the two.

What post-training changes, and at what resolution. That reward optimization improves sampling efficiency while narrowing the set of problems solved at least once is established (Chen et al., 2026), and several studies report the solved set turning over during prolonged training (Yuan et al., 2026; Zhou, 2026; Qin et al., 2026). Because these are claims about set membership estimated from finitely many samples, they depend on the budget and on the threshold at which a change is counted: $\mathrm{Pass@}k$ is unstable as k approaches n (Hariri et al., 2026), while an unbiased estimator with per-problem counts and an explicit floor exists (Chen et al., 2021). Entropy collapse under reward optimization is widely reported (Cui et al., 2025), while entropy regularization is reported to give little benefit (Shen, 2026); Section 4 offers an explanation for the latter that is compatible with the former.

3 PRELIMINARIES

Notation. An autoregressive policy $\pi_\theta(y \mid x)$ answers a prompt x and receives a binary verifiable reward $r \in \{-1, +1\}$. For each prompt, G responses are drawn and rewards are group-normalized, $\hat{r} = (r - \mu_r)/(\sigma_r + \varepsilon)$, with μ_r and σ_r the group mean and sampled standard deviation. We write c for the number of correct responses in a group and $p = c/G$ for the realized group correctness rate. A group with $c \in \{0, G\}$ is *degenerate*: all responses share the same reward, $\hat{r} = 0$ throughout, and the group carries no differential signal. Groups with $1 \le c \le G - 1$ are *informative*.

Reward maximization vs. Max-ent. Standard RL objectives maximize expected reward $\mathbb{E}_\pi r(x, y)$. A widely used RLVR objective, GRPO (Shao et al., 2024), replaces the value model with $\hat{r}$ in a PPO-style objective (Schulman et al., 2017). Max-Ent RL (Ziebart et al., 2008; Haarnoja et al., 2018) instead balances expected reward against policy entropy with $\mathbb{E}_\pi[r(x, y) + \lambda \sum_{t=1}^{|y|} \mathcal{H}(\pi(\cdot \mid x, y_{<t}))]$ as an objective for maximization, yielding an optimal policy proportional to exponentiated reward: $\pi^*(y \mid x) \propto \exp\big(\alpha\, r(x, y)\big)$.

Trajectory balance with a multiplicative reference. We study a Tsallis-regularized Max-Ent formulation, using the entropy (Tsallis, 1988) $\mathcal{H}_q\big(\pi(\cdot \mid x)\big) = \frac{1}{q-1}\Big(1 - \sum_y \pi(y \mid x)^q\Big)$ as a regularizer. It recovers the Shannon case as $q \to 1$. The exponential tilt is then replaced by a q-exponential one to reweight the reference distribution:

$$\pi_q^*(y \mid x) \;\propto\; \big[\,\hat{r}(x, y) + \delta\,\big]_+^{1/(q-1)}\, \pi_{\mathrm{ref}}(y \mid x), \qquad \delta = \frac{1}{(q-1)\,\alpha_q}, \tag{1}$$

where α_q is the inverse temperature and δ the reparametrization we adopt in its place: responses with $\hat{r} < -\delta$ get no weight. We also compare against a reward-maximizing GRPO baseline whose KL reference is its initialization.

We optimize these targets with the practical trajectory-balance surrogate (Malkin et al., 2022; Zhu et al., 2026b). Let $\tilde{r} = \frac{1}{q-1}\log[\hat{r}+\delta]_+$, so that the nominal target is $\pi_q^* \propto \exp(\tilde{r})\pi_{\mathrm{ref}}$, and replace its partition function by a learned $Z_\phi(x)$:

$$\mathcal{L}(\theta,\phi) \;=\; \frac{1}{|\mathcal{B}|}\sum_{(x,y)\in\mathcal{B}} w_{\mathrm{imp}}\cdot\Big(\log Z_\phi(x) + \tfrac{1}{|y|}\log\pi_\theta(y\mid x) - \tilde{r}(x,y) - \tfrac{1}{|y|}\log\pi_{\mathrm{ref}}(y\mid x)\Big)^2, \quad (2)$$

where $w_{\mathrm{imp}} = \mathrm{clip}\big(\pi_\theta(y\mid x)/\pi_{\mathrm{behav}}(y\mid x), 1-\epsilon, 1+\epsilon\big)$ is a stop-gradient trust-region weight centered on the behavior policy. Equation 1 specifies the nominal target, whereas Equation 2 is its stabilized training surrogate. In particular, normalizing only the policy–reference log-ratio changes the exact zero-residual fixed point. This follows the practical construction used in prior LLM trajectory-balance training, but means that the target geometry analyzed below and the distribution attained by a finite training run must be evaluated separately; Appendix A makes this distinction explicit. In our experiments the multiplicative reference is the checkpoint from which optimization starts, as detailed in Appendix B.

4 The design space of Max-Ent targets

Entropy-regularized trajectory balance is motivated by the possibility of retaining response diversity that pure reward maximization may reduce (Chen et al., 2026; Zhu et al., 2026b). If this premise holds, a distributional target should preserve multiple high-quality modes, maintain probability on rare correct trajectories, and improve both sampling coverage and difficult-problem accuracy.

To study the premise beyond a single construction, we use the Tsallis family as an experimental extension of the Shannon objective. For a scalar score z, its reward tilt has the form

$$w_{q,\delta}(z) = [z+\delta]_+^{t}, \qquad t = \frac{1}{q-1}. \quad (3)$$

Our use of this family is instrumental: q supplies an additional handle with which to vary the target, rather than constituting a new objective in itself.

The two parameters control different features of the tilt. The shift δ places the support boundary at $z = -\delta$, while the exponent t controls how rapidly the weight grows inside the support. At fixed q, however, changing δ also changes relative weights away from the boundary, coupling support and scale. Changing q can separate these effects, although smoothness restricts the available range: the truncated power is continuously differentiable at the boundary only for $t > 1$, or $1 < q < 2$. The boundary case $q = 2$ is continuous but has the sparsemax kink; among integer exponents yielding differentiability, the smallest is $t = 2$, corresponding to $q = 3/2$.

The Shannon objective is the limiting reference point. As $q \to 1$, retaining a finite affine slope requires scaling δ jointly with q, and the truncated power converges, up to normalization, to an exponential tilt

$$w_{(\mathrm{Shannon})}(z) = \exp(\alpha z).$$

It has no finite-support boundary and only its temperature varies the target. The generalized family therefore appears to provide control over target shape and support that the Shannon objective cannot express. We next ask how much of this apparent flexibility remains in the setting in which these objectives are trained.

4.1 What binary RLVR leaves of the design space

Standard RLVR supplies a particularly restrictive score: rewards are binary and normalized within a group of G sampled responses. For a group with correctness rate $p \in (0,1)$, write $s = \sqrt{(1-p)/p}$. Under population normalization, the advantage takes only two values, $\hat{r} = s$ on correct responses and $\hat{r} = -1/s$ on incorrect ones (Mroueh, 2025). Sample normalization multiplies both values by $\sqrt{(G-1)/G}$, without changing the two-valued reduction.

One-scalar reduction. For a sampled group $y_{1:G}$, define the reweighted empirical distribution

$$\widehat{\pi}_G^{(w)}(y_i \mid x) = \frac{w_i}{\sum_{j=1}^{G} w_j}, \qquad w_i = f\big(\hat{r}(x, y_i)\big).$$

Proposition 1 (Two-valued group reweighting)**.** *For any informative sampled group, the tilt takes two values, w_1 on correct responses and w_0 on incorrect responses. The reweighted empirical group distribution is therefore determined by*

$$\rho(p) = \frac{w_0}{w_1} = \frac{f(-1/s)}{f(s)}, \qquad p'_G(p) = \frac{p}{p + \rho(p)(1-p)}, \tag{4}$$

where p'_G is the correct-class mass of the reweighted empirical group. Conversely, every profile $\rho : (0,1) \to [0,1]$ is realized by some tilt f.

Proof sketch. The two classes have unnormalized masses $w_1 p$ and $w_0(1-p)$, so normalization leaves only their ratio. Since $p \mapsto s$ is bijective, the converse follows by defining f to realize any admissible $\rho(p)$. Appendix C gives the full derivation, the unbiased-variance correction, and the extension to non-binary rewards. □

The reduction concerns the reward tilt rather than the full target $\pi^*(y \mid x) \propto w(y)\pi_{\mathrm{ref}}(y \mid x)$: the fixed reference can still redistribute mass within either correctness class. It is held fixed in our sweep, while the tilt is varied.

For the q-exponential family, the induced profile is

$$\rho(p) = \left(\frac{[\delta - 1/s]_+}{s + \delta} \right)^{1/(q-1)}. \tag{5}$$

Its endpoints have a direct interpretation: $\rho = 0$ suppresses the incorrect class, whereas $\rho = 1$ makes the reweighting blind to correctness. Hence, all the freedom introduced above reaches the group-level target through the single profile $\rho(p)$.

Where the available targets lie. Our reference configuration combines two independently motivated choices. We set $q = 3/2$, the largest differentiable q with an integer exponent, and $\delta = 1$, which places the idealized support boundary at $p^* = 1/2$. Although each choice is natural in isolation, neither controls how strongly their combination reweights correct versus incorrect responses. We therefore locate the resulting ρ-profile within the attainable target space.

Table 1: Analytic support boundaries and realized group-level target statistics. p^* and $k^* = \lceil Gp^* \rceil$ use the population-normalized idealization; the remaining quantities are computed with the implementation's unbiased group standard deviation.

Configuration	p^*	k^*	$\max_k \rho$	$p'_G(k{=}2)$	$p'_G(k{=}4)$	$\rho < 10^{-2}$
$q{=}3/2,\ \delta{=}1$ (canonical)	0.500	4	0.0354	0.909	0.992	$k \geq 4$
$q{=}3/2,\ \delta{=}2$	0.800	7	0.1627	0.672	0.884	$k = 7$
$q{=}3/2,\ \delta{=}4$	0.941	8	0.3956	0.468	0.722	none
exponential, $\beta{=}15$	–	–	$6.5{\times}10^{-13}$	≈ 1	≈ 1	all

At the canonical pair, the correct class already receives 0.91 of the reweighted mass at $k = 2$, while ρ never exceeds 0.035. Nearly half of the informative groups have $\rho < 10^{-2}$. The published exponential objective is still more suppressive: for $G = 8$ and $\beta = 15$, $\rho \leq 6.5 \times 10^{-13}$. Thus their nominally different shapes are strongly compressed in the realized region.

The remaining degree of freedom. The general support–scale trade-off now becomes discrete. At fixed q, increasing δ makes the target less concentrated but moves the boundary towards less frequent group compositions: at $\delta = 2$, only groups with seven of eight correct responses are suppressed; at $\delta = 4$, no informative group is suppressed. Consequently, a sweep over δ cannot isolate support from scale.

The index retains one degree of freedom that the shift does not. At the fixed effective shift $\delta_{\mathrm{eff}} = 1$, the idealized boundary stays fixed while $\max_k \rho$ grows from 0.002 at $q = 1.25$ to 0.155 at $q = 1.9$,

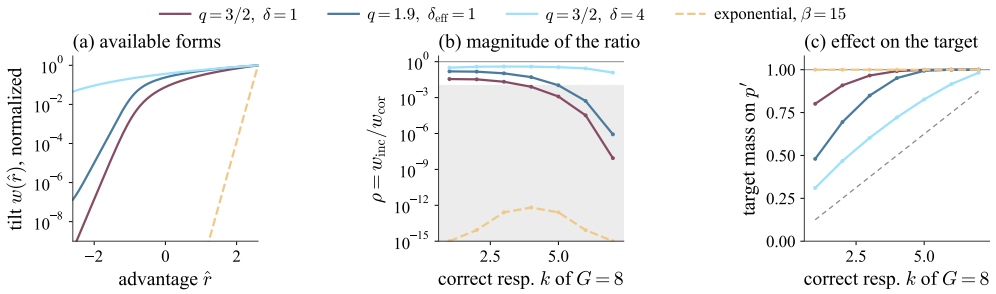

Figure 1: **(a)** Tilt against the group-normalized advantage. **(b)** Reweighting ratio by group composition; shading marks $\rho < 10^{-2}$. **(c)** Correct-class mass of the reweighted group distribution, against the untilted $p'_G = p$ and the hard threshold $p'_G = 1$.

approaching 0.19 as $q \to 2$. Thus the generalized family can change the scale at a fixed boundary, but only within the smoothness-constrained range.

The contrast with the Shannon's is stronger than the absence of a finite support boundary. Its entire reweighting profile obeys a symmetry that no choice of temperature can remove:

Proposition 2 (Symmetry of the exponential target)**.** *For the affine transform $\tilde{r} = \alpha\hat{r}$,*

$$\rho(p) = \exp(-2\alpha/\sigma_r).$$

Since σ_r depends on p only through $p(1-p)$, $\rho(p) = \rho(1-p)$ for every α. The profile is therefore symmetric in group competence, has its maximum at $p = 1/2$, and cannot be shifted by temperature.

For $q > 1$, the maximum occurs at $k = 1, 2, 3$ for $\delta = 1, 2, 4$. The generalized family retains one experimentally useful degree of freedom after the binary reduction: it can control the competence level at which the target discriminates most strongly, something no Shannon temperature can reproduce.

5 Experiments

The analysis so far concerns the targets specified by the objectives; here we ask how the trained checkpoints behave. A sweep can move ρ substantially while its behavioral consequences remain hidden by training variation or sampling noise. We organize the experiments around three expectations from the Max-Ent motivation: a more diverse target should expand reachability, improve accuracy on difficult problems, and produce a more entropic policy. In each case, the observed result is weaker than, or opposite to, that expectation.

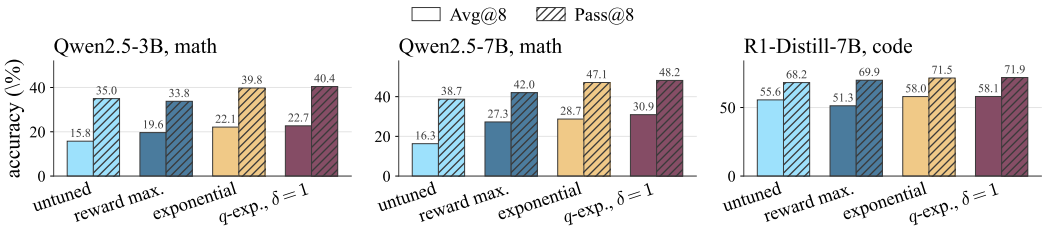

Figure 2: A sweep of the objective family at 400 steps. Bars average five mathematics benchmarks for two backbones and the code benchmarks for a distilled code model.

5.1 Basic fine-tuning sweep

We first compared the two objectives at 400 steps on two mathematics backbones and a distilled code model, using GRPO as the reward-maximizing baseline. Both TB objectives have higher point

estimates in almost every cell, and the tuned q-exponential outperforms the exponential on most of them (Figure 2; per-benchmark results are in Appendix I).

Section 4 shows that these targets are numerically close to each other and to a hard threshold, so the objectives alone do not predict a stable ordering between them. Neither does Figure 2: differences among the fine-tuned bars are 0.6–2.2 points, whereas re-evaluating identical weights moves the same aggregate by 3.3. We therefore evaluate the family using a single protocol whose resolution is established first.

5.2 Setup and resolution

We fine-tune **Qwen2.5-7B-Instruct** on DAPO-Math-17k with a reward $r(x, y) \in \{-1, +1\}$, and evaluate 128 samples for each of 90 AIME 2024–2026 problems. Eleven trajectory-balance checkpoints share the instruction-tuned initialization; the $q = 1.9$ run is compared against its documented step-200 warm start. GRPO resumes from a valid step-200 GRPO checkpoint for another 200 steps, giving 400 steps in total. Run configurations, estimator definitions, tests, chat templates, and sampling parameters are deferred to Appendix E.

Scoring identical initial weights through two evaluation pipelines produces 6 gained and 3 lost problems. Reseeding one configuration moves accuracy by 2.71 points $[0.58, 4.96]$, and changing only the chat-template convention moves it by 6.96 points $[5.04, 9.00]$. Within the present pool, the warm–cold spread at one δ is 1.06 percentage points, larger than the 0.61-point change across the entire δ-sweep. Full controls are in Appendix F.

5.3 The sweep moves the target

Across $\delta \in \{1, 2, 4\}$, the realized target spans an order of magnitude in $\max_k \rho$, the batch-weighted mass on the correct class falls from 0.93 to 0.63, and the fraction of informative groups suppressed below $\rho = 10^{-2}$ falls from 0.47 to zero. Every diagnostic agrees with the closed form of Section 4 to within 5×10^{-4}. The $q = 1.9$ extension retains the interpretable boundary $p^* = 1/2$ while raising $\max_k \rho$ from 0.0354 to 0.1547. The sweep therefore changes the intended target substantially; the following results ask whether the learned policies change with it.

5.4 Reachability does not expand

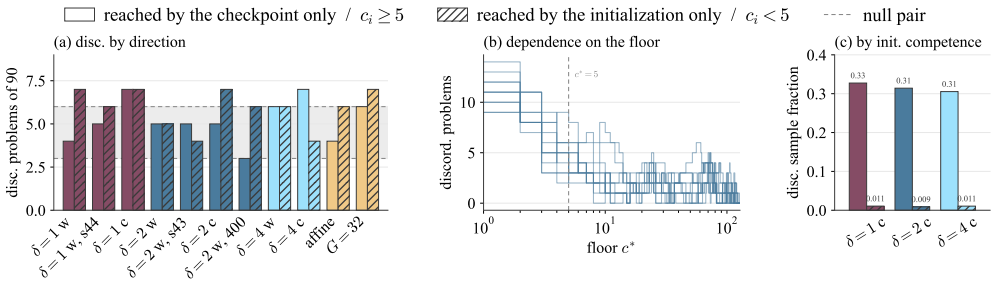

Figure 3: Reachability against each checkpoint's own initialization on the 90-problem pool at $n = 128$. **(a)** Discordant problems by direction for the eleven common-initialization configurations, with the re-evaluation control shown as a band. **(b)** Discordance as a function of the solved-count floor c^*. **(c)** Sample-level discordance on problems solved at least 5 of 128 times by the initialization and on the remaining problems.

We expected the more diverse targets to make previously unreachable problems reachable. For the eleven common-initialization configurations, they do not: each gains 3–7 problems and loses 4–7, the same scale as the $6/3$ re-evaluation control. No problem-level McNemar test is significant, with the smallest $p = 0.51$, and Pass@128 remains between 0.422 and 0.478 (see Figure 3).

GRPO behaves similarly at the reachability level despite improved accuracy. Against its step-200 checkpoint, the final gains seven problems and loses six, with Pass@128 $= 0.478$ and a problem-

level McNemar $p = 1.0$. At the robust floor $c^* = 5$, it gains five problems and loses one. Its accuracy gain is not accompanied by a detectable expansion of the reachable set.

The $q = 1.9$ run moves in the opposite direction. Against its own warm start, it gains 5 problems and loses 24, reducing Pass@128 from 0.467 to 0.256 ($p = 5.5 \times 10^{-4}$). At the prespecified robustness floor $c^* = 5$, it gains none and loses nine. Appendix H reports the complete floor curve.

Nor does the sweep shift probability toward difficult problems. For the original cold runs, sample-level discordance is 30.5–32.7% on problems already solved at least 5 times by the initialization, but only 0.9–1.1% on the rest. For $q = 1.9$, the low-competence stratum contains 5 gained and 17 lost problems. The observed changes therefore remain within the existing competence region or contract it rather than expanding its boundary.

5.5 Accuracy does not improve on difficult problems

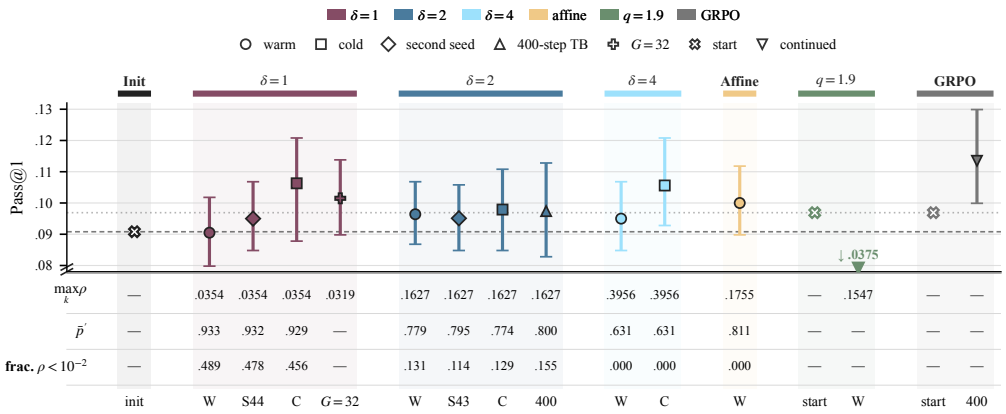

Figure 4: Checkpoint accuracy and target diagnostics. Points show Pass@1, with paired 95% intervals shifted to the corresponding starting checkpoint. Colors denote target families and markers training regimes; aligned rows report target statistics. The $q = 1.9$ result is marked below the displayed range.

We expected the less suppressive targets to improve per-sample accuracy, particularly where the initialization is weak. Instead, the original eleven checkpoints show no monotone relation between target diversity and accuracy: ten of the eleven paired bootstrap intervals cross zero, all marginal Pass@1 intervals overlap, and the warm–cold spread exceeds the effect of δ across its full range (see Figure 4). The small positive result for cold $\delta = 4$ is not accompanied by an expansion in reachability.

The $q = 1.9$ run produces the significant negative paired effect: Pass@1 falls from 0.0969 to 0.0375, with $\Delta = -0.0594$ $[-0.100, -0.023]$. Moving the target does not yield the expected gain on complex problems: the strongest intervention instead reduces quality. Full results in Appendix G.

GRPO shows a different pattern: continuing it from step 200 to step 400 raises Pass@1 from 0.0969 to 0.1135, with a paired difference of 0.0166 $[0.0026, 0.0334]$. Its Pass@8 and Pass@128 are 0.2491 and 0.4778, respectively. The per-sample gain is therefore detectable, although the reachability comparison above remains balanced.

5.6 Policy entropy falls while judged semantic diversity rises

Max-ent targets are expected to leave more diverse policies (Zhu et al., 2026b). Against the initialization, the opposite occurs. The final answer Gini–Simpson index is 0.8333, while all eleven trained checkpoints lie between 0.7836 and 0.8141 (Figure 5 (b)). The ordering survives the correctness controls, and the checkpoint matching the initialization's pool accuracy (0.0905 versus 0.0908) remains 0.037 $[0.017, 0.057]$ less diverse. The $q = 1.9$ run also decreases pointwise from 0.8085 to 0.7957 relative to its own warm start. Per-class estimates and their effective sample sizes are in Appendix J.

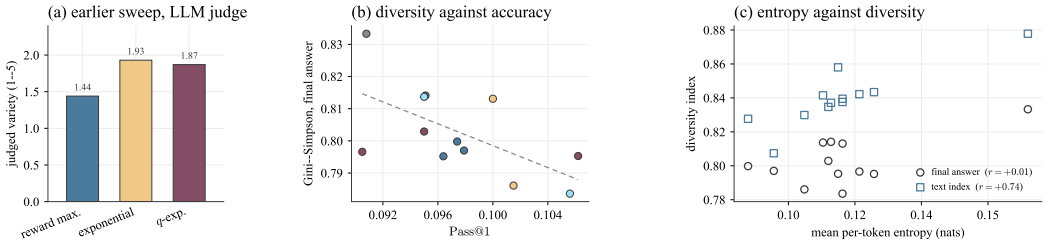

Figure 5: **(a)** Methodological variety judged by Claude Sonnet 4.6 in the earlier sweep on AIME pool. **(b)** Gini–Simpson diversity over final answers against pool accuracy for the common-initialization sweep. **(c)** Mean per-token entropy against text- and answer-level diversity indices.

On the available teacher-forcing artifacts, per-token entropy falls from 0.159 at initialization to 0.0959–0.1243 across the eleven checkpoints. However, the policies do not move far: argmax divergence from the initialization is 1.77–2.88%, and at most 0.0221% of generated tokens fall outside the reference support. Fine-tuning sharpens mostly the same token choices rather than replacing them. Entropy correlates with the text-level diversity index ($r = 0.74$) but not with final-answer diversity ($r = 0.01$) (see Figure 5c). Appendix J isolates this discrepancy.

This initialization comparison alone does not identify an objective effect, because fine-tuning generally sharpens policies. The relevant comparator is a GRPO checkpoint trained under the same setting. GRPO reaches a mean per-token entropy of 0.127 $[0.104, 0.156]$, above the point estimates of all available TB checkpoints (0.096–0.124), although their marginal bootstrap intervals overlap. Thus, the Max-Ent objectives provide no measurable token-entropy advantage over reward maximization in this setup. At the same time, GRPO improves $\mathrm{Pass@1}$ by 0.0166 $[0.0026, 0.0334]$ against its initialization, but gains and loses a similar number of reachable problems ($7/6$, $p = 1.0$).

The independent semantic evaluation gives slightly different picture. Using Claude Sonnet 4.6, we reproduce the direction of the FlowRL result (Appendix J.1): the exponential and q-exponential objectives receive methodological-variety scores of 1.93 and 1.87, respectively, compared with 1.44 for GRPO (Figure 5 (a)). The relative separation is significantly smaller than in the original FlowRL evaluation (1.85x vs 1.34x), making it highly dependent on the judge-model choice.

The disagreement is substantive rather than contradictory. The judge evaluates differences between complete solution strategies, final-answer diversity measures variation in extracted outcomes, and teacher-forced entropy measures local predictive uncertainty. The horizon-matched paired comparison finds higher token entropy under GRPO, whereas the semantic judge assigns higher variety scores to the Max-Ent solutions. Consequently, these quantities should not be used interchangeably as measurements of "diversity."

6 Conclusion

We examined whether the distributional motivation of Max-Ent trajectory-balance training survives the constraints of RLVR with binary rewards. The generalized formulation exposes the remaining target-design freedom, which reduces under binary rewards to a single reweighting profile; support and scale remain coupled, and standard parameter choices can place apparently different objectives in the same narrow, strongly suppressive region. Substantial changes in this profile do not translate into the expected behavioral ordering.

Reachability does not expand, accuracy changes are mostly below the resolution of the procedure, and the strongest intervention contracts both accuracy and coverage. Max-Ent training provides no token-entropy advantage over GRPO, despite producing solutions judged as more semantically varied. These results argue for validating reachability, quality, semantic diversity, and policy entropy separately. They also delimit the scope of our analysis: the theory characterizes the nominal target transform, whereas training uses a practical surrogate under substantial procedural variation. Whole-residual normalization and direct residual are therefore important controls for future comparisons.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing, for the retrieval and discovery of related work, and for drafting sections of the paper (in particular, structuring the appendix). We also used AI agents to assist with writing configuration files and implementing parts of the codebase. All AI-assisted content, including text, references, related-work summaries, and all realizations in code, was reviewed and edited by the authors. We take full responsibility for the content of this work, including all text and claims.

ETHICS STATEMENT

This work studies objective design for post-training language models with verifiable mathematical rewards. It involves no human participants, no collection of personal data, and no deployment study. We use publicly available mathematical training and evaluation data subject to their respective licenses and evaluate answers with an automatic verifier.

Improving mathematical reasoning may increase the capabilities of general language models, including in settings where stronger problem-solving ability could be misused. Our contribution is methodological rather than a release of a new deployed system, and we do not claim that the proposed objectives are universally beneficial. In particular, we report negative and inconclusive results, distinguish nominal targets from their practical optimization surrogates, and disclose limitations arising from finite training, evaluation variance, and incomplete controls.

Our experiments require substantial accelerator compute. We limit this cost by reusing fixed evaluation pools and teacher-forced diagnostic trajectories where appropriate, and by reporting the provenance of checkpoints and evaluation protocols. The semantic diversity analysis uses an automated LLM judge and is reported as a limited auxiliary measurement rather than as a substitute for human evaluation.

REPRODUCIBILITY STATEMENT

The experimental appendix specifies the model and data, optimization settings, checkpoint horizons, hardware, evaluation protocol, estimators, statistical procedures, and known integrity caveats. The supplementary archive contains the training and evaluation code, experiment configurations, figure scripts, and machine-readable CSV and JSON artifacts underlying the reported results. It also includes dependency information, provenance manifests, and validation instructions for the released components. Model checkpoints and large raw rollout files are excluded because of their size, while unavailable artifacts and remaining release caveats are documented explicitly in the archive.

REFERENCES

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Remi Munos. Asymmetric reinforce for off-policy reinforcement learning: Balancing positive and negative rewards. *Advances in Neural Information Processing Systems*, 38:9640–9664, 2026.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394, 2021.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. Evaluating large language models trained on code, 2021. URL `https://arxiv.org/abs/2107.03374`.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026.

Yinlam Chow, Ofir Nachum, and Mohammad Ghavamzadeh. Path consistency learning in tsallis entropy regularized mdps. In *International conference on machine learning*, pp. 979–988. PMLR, 2018.

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.

Tiago da Silva, Eliezer d da Silva, and Diego Mesquita. On divergence measures for training gflownets. *Advances in Neural Information Processing Systems*, 37:75883–75913, 2024.

Jake Fawkes and Jason Hartford. f-trajectory balance: A loss family for tuning gflownets, generative models, and llms with off-and on-policy data. *arXiv preprint arXiv:2605.15417*, 2026.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't pass@k: A bayesian framework for large language model evaluation, 2026. URL https://arxiv.org/abs/2510.04265.

Edward Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 43053–43083, 2024.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018.

Kyungjae Lee, Sungyub Kim, Sungbin Lim, Sungjoon Choi, and Songhwai Oh. Tsallis reinforcement learning: A unified framework for maximum entropy reinforcement learning. *arXiv preprint arXiv:1902.00137*, 2019.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

MAA. American invitational mathematics examination (AIME). https://maa.org/, 2025. Accessed: 2026-07-10.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Sobhan Mohammadpour, Emmanuel Bengio, Emma Frejinger, and Pierre-Luc Bacon. Maximum entropy gflownets with soft q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2593–2601. PMLR, 2024.

Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, et al. Openai o1 system card, 2026.

Chuanyu Qin, Chenxu Yang, Qingyi Si, Naibin Gu, Peng Fu, and Zheng Lin. Learning to solve, forgetting to retain: Correct-set turnover in rlvr, 2026. URL https://arxiv.org/abs/2606.03087.

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

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

Guobin Shen, Chenxiao Zhao, Xiang Cheng, Lei Huang, and Xing Yu. Vespo: Variational sequence-level soft policy optimization for stable off-policy llm training. *arXiv preprint arXiv:2602.10693*, 2026.

Han Shen. On entropy control in llm-rl algorithms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 136535–136556, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/dd0f29b61307b08e69cf93a50e3298fb-Paper-Conference.pdf.

Daniil Tiapkin, Nikita Morozov, Alexey Naumov, and Dmitry P Vetrov. Generative flow networks as entropy-regularized rl. In *International Conference on Artificial Intelligence and Statistics*, pp. 4213–4221. PMLR, 2024.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.

Esmeralda S. Whitammer, Salem Lahlou, Tristan Deleu, Xu Ji, Edward Hu, Katie Everett, Dinghuai Zhang, and Yoshua Bengio. Gflownets and variational inference, 2026. URL https://arxiv.org/abs/2210.00580.

Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Quantile advantage estimation: Stabilizing rlvr for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 132218–132236, 2026.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. Bnpo: Beta normalization policy optimization. *arXiv preprint arXiv:2506.02864*, 2025.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, et al. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Suqin Yuan, Jinkun Chen, Jiyang Zheng, Muyang Li, Lei Feng, Dadong Wang, Tao Xiang, Tongliang Liu, and Bo An. Understanding diversity collapse in rlvr via the lens of overtraining. *arXiv preprint arXiv:2606.15455*, 2026.

Todd Zhou. When rlvr shrinks the reasoning boundary: Diagnosing pass@ k inversion. *arXiv preprint arXiv:2607.20543*, 2026.

Lingwei Zhu, Zheng Chen, Matthew Schlegel, and Martha White. Generalized munchausen reinforcement learning using tsallis kl divergence. *arXiv preprint arXiv:2301.11476*, 2023.

Lingwei Zhu, Haseeb Shah, Han Wang, Yukie Nagai, and Martha White. q-exponential family for policy optimization. In *International Conference on Learning Representations*, volume 2025, pp. 40717–40744, 2025.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. *Advances in Neural Information Processing Systems*, 38:126546–126573, 2026a.

Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Henry Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, et al. Flowrl: Matching reward distributions for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 151758–151780, 2026b.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 1433–1438. AAAI Press, 2008.

A LENGTH NORMALIZATION AND THE FIXED POINT

Equation equation 1 defines a nominal distribution, while Equation equation 2 follows the usual practical sequence: target specification, a trajectory-level matching loss, and stabilization for LLM optimization. In particular, it divides the policy and reference log-probabilities by $|y|$, since sequence log-probabilities grow with length whereas verifier rewards remain bounded. The same separation between a theoretical objective and a stabilized estimator appears in FlowRL-style trajectory balance, log-variance GFlowNet training, policy-gradient RLVR, and sequence-level importance sampling (Zhu et al., 2026b; Whitammer et al., 2026; Shao et al., 2024; Yu et al., 2026; Liu et al., 2025; Shen et al., 2026). This convention improves optimization scaling, but it is not neutral with respect to the distribution matched at zero residual.

We therefore distinguish the nominal target in equation 1 from the length-normalized implementation. Setting the latter residual to zero gives

$$\pi_\theta(y \mid x) = \pi_{\text{ref}}(y \mid x) \exp\big(|y|(\tilde{r}(x,y) - \log Z_\phi(x))\big).$$

For an incorrect and a correct response of the same length L, Proposition 1 gives the effective class-tilt ratio $\rho_{\text{eff}} = \rho^L$. Thus, when $\rho = 0.4$ and $L = 100$, $\rho_{\text{eff}} \approx 1.6 \times 10^{-40}$. At unequal lengths, however, the exact ratio also contains

$$\exp((|y_{\text{i}}| - |y_{\text{c}}|) \log Z_\phi(x)),$$

together with the corresponding length-weighted rewards. Hence, ρ^L is an equal-length diagnostic, not the exact incorrect-to-correct mass ratio of a variable-length group.

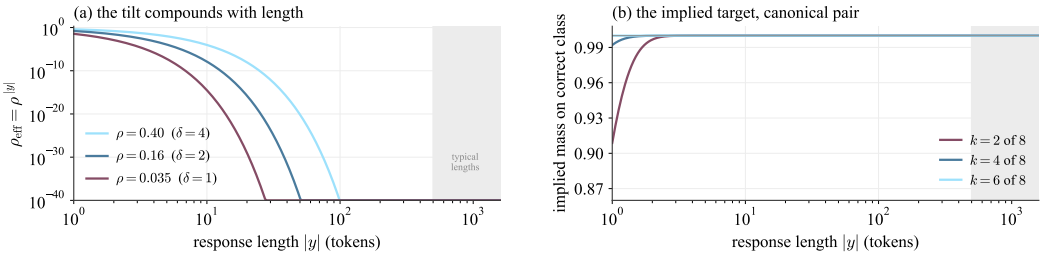

Figure 6: **(a)** The equal-length effective reweighting ratio ρ^L for the three shifts of Section 5; shading marks the lengths this setting produces. **(b)** The corresponding equal-length fixed-point proxy for correctness mass at the canonical pair.

This qualification fixes the scope of the theory. The analysis of $\rho(p)$ characterizes the nominal reward transform specified for the optimizer, not the attained class reweighting of a finite, clipped training run. We therefore use ρ to verify that the sweep changes its intended input, and behavioral measurements to determine whether this change survives optimization; we do not infer the latter from the former. Accordingly, the diagnostics in the main text measure the realized transform on observed groups; they are not evidence that the checkpoint has converged to equation 1. In practice, $\tilde{r}$ is an energy scale derived from normalized binary rewards and $\log Z_\phi$ is learned jointly as a prompt-dependent offset. Together with a bounded training horizon, trust-region weighting, and

gradient clipping, this makes the implemented objective a practical surrogate rather than a literal solver for its nominal fixed point.

Normalizing the whole residual would restore the nominal zero-residual equation. It would nevertheless change per-example gradient weights, clipping behavior, and the optimization of Z_ϕ, so it is a substantive robustness ablation rather than a drop-in replacement with the same optimization properties. We retain the FlowRL-style form to keep the implementation fixed while sweeping the target transform. The effect of whole-residual normalization is an important separate control.

B WHY THE MULTIPLICATIVE REFERENCE APPEARS

The reference enters equation 1 multiplicatively because a KL-regularized optimum has the form $\pi^* \propto \exp(\alpha\hat{r})\pi_{\text{ref}}$, and the q-deformation gives equation 1. In our experiments, the reference is the checkpoint from which optimization starts, so reference and initialization are not varied separately.

Accordingly, Proposition 1 concerns the binary reward tilt conditional on a fixed reference; it does not claim that the full learned policy is two-valued. The reference continues to determine relative probabilities within each correctness class, while finite optimization and length normalization determine how closely the resulting checkpoint approaches the nominal product.

The reward tilt and reference are not symmetric. The former acts on the two correctness classes induced by binary RLVR rewards, whereas the reference acts on complete sequences. A tilt can formally be absorbed into a redefined reference, but the converse requires an arbitrary sequence-dependent reward and no longer preserves the binary-reward setting. Length normalization does not change this distinction: it scales the policy–reference log-ratio, while the reward and learned normalizer define the length-dependent tilt above. Because Z_ϕ depends only on the prompt, it cannot generally cancel response-level length variation.

Removing the reference term at the same step count and learning rate leaves the objective poorly anchored: the learned normalizer and mean log-policy each drift by roughly ten nats while their sum remains within 0.85. Off-support generation rises from under 8% to 96.8%, and token entropy reaches $3.5\times$ that of the reference. This is a diagnostic of the reference anchor; by comparison, the off-support fractions in Section 5.6 are at most 0.022%.

C PROOF OF PROPOSITION 1 AND EXTENSIONS

C.1 THE TWO-VALUED ADVANTAGE

Fix a prompt and a group of G responses of which c are correct, and write $p = c/G$ with $0 < p < 1$. With $r \in \{-1, +1\}$ the group mean is $\mu_r = 2p - 1$. The biased (population) group standard deviation is

$$\sigma_r = \sqrt{\mathbb{E}[r^2] - \mu_r^2} = \sqrt{1 - (2p-1)^2} = 2\sqrt{p(1-p)}. \tag{6}$$

Writing $a = \sqrt{1-p}$ and $b = \sqrt{p}$, so that $a^2 + b^2 = 1$ and $\sigma_r = 2ab$, the group-normalized advantage on each correctness class is

$$\hat{r}_1 = \frac{1-\mu_r}{\sigma_r} = \frac{2(1-p)}{2ab} = \frac{a}{b}, \qquad \hat{r}_0 = \frac{-1-\mu_r}{\sigma_r} = \frac{-2p}{2ab} = -\frac{b}{a}, \tag{7}$$

that is, $\hat{r}_1 = s$ and $\hat{r}_0 = -1/s$ with $s = a/b = \sqrt{(1-p)/p}$. The advantage depends on a response only through its correctness, and the smoothing constant ε in $\hat{r} = (r - \mu_r)/(\sigma_r + \varepsilon)$ is negligible for informative groups; on degenerate groups ($c \in \{0, G\}$) the numerator vanishes as well and $\hat{r} = 0$ throughout, so such groups carry no differential signal and are excluded from what follows. This closed form is established in the analysis of group-relative objectives (Mroueh, 2025); we restate it because every statement below is a consequence of it.

The unbiased estimator. The implementation computes σ_r with Bessel's correction, so the realized normalizer is $\sigma_r = 2\sqrt{\frac{G}{G-1}\,p(1-p)}$ and the advantages are $\hat{r}_1 = s\sqrt{\frac{G-1}{G}}$, $\hat{r}_0 = -\frac{1}{s}\sqrt{\frac{G-1}{G}}$. At $G = 8$ this rescales both by $\sqrt{7/8} \approx 0.935$, which shifts ρ by roughly one fifth at the canonical setting and moves the truncation boundary correspondingly. All numerical values reported in the pa-

per use the unbiased form, matching the implementation; the algebraic statements below are written with the population form for readability and hold verbatim after the substitution $s \mapsto s\sqrt{(G-1)/G}$.

C.2 Collapse to a single ratio

Let $w = f(\hat{r})$ be any multiplicative reward tilt. Within the sampled group, consider the reweighted empirical distribution defined before Proposition 1. Since w is constant on each correctness class, write $w_1 = f(s)$ and $w_0 = f(-1/s)$. Its correct-class mass is

$$p' \;=\; \frac{w_1 p}{w_1 p + w_0(1-p)} \;=\; \frac{p}{p + \rho\,(1-p)}, \qquad \rho \;=\; \frac{w_0}{w_1}, \tag{8}$$

the common factor w_1 having cancelled. Two tilts with the same ρ therefore induce the same target within the group, whatever their functional form, and the entropic index and the additive shift can affect the target only through ρ. This establishes the first part of the proposition.

C.3 The q-exponential tilt

For $f(\hat{r}) = [\hat{r} + \delta]_+^{1/(q-1)}$ with $q > 1$,

$$\rho(p) \;=\; \left(\frac{[\,-1/s + \delta\,]_+}{s+\delta}\right)^{1/(q-1)} \;=\; \left(\frac{[\,\delta - 1/s\,]_+}{s+\delta}\right)^{1/(q-1)}, \tag{9}$$

the denominator being positive since $s > 0$ and $\delta > 0$. The numerator turns to zero when $\delta \le 1/s$, i.e. when $s \le 1/\delta$, i.e. when

$$\sqrt{\frac{1-p}{p}} \le \frac{1}{\delta} \quad\Longleftrightarrow\quad p \;\ge\; p^* \;=\; \frac{\delta^2}{1+\delta^2}, \tag{10}$$

which is the truncation boundary quoted in the main text; at $G = 8$ the corresponding count is $k^* = \lceil G p^* \rceil$. Note that p^* depends on δ alone and the exponent on q alone, which is the decoupling exploited in Paragraph 4.1.

At $q = 3/2$ the exponent is 2, and with $a = \sqrt{1-p}$, $b = \sqrt{p}$ the two class weights are

$$w_1 p = \left(\frac{a}{b} + \delta\right)^2 b^2 = (a + \delta b)^2, \qquad w_0(1-p) = \left(\delta - \frac{b}{a}\right)^2 a^2 = (\delta a - b)^2, \tag{11}$$

so the normalizer collapses by the identity $(a+\delta b)^2 + (\delta a - b)^2 = (1+\delta^2)(a^2+b^2) = 1+\delta^2$, the cross terms $\pm 2\delta ab$ cancelling, and

$$p' \;=\; \frac{\left(\sqrt{1-p} + \delta\sqrt{p}\,\right)^2}{1+\delta^2} \quad (p \le p^*), \qquad p' = 1 \quad (p \ge p^*). \tag{12}$$

At the boundary $\delta a = b$ the first branch evaluates to 1 and the two agree.

C.4 Surjectivity

We show every profile is realized, so that ρ exhausts the design space. Let $\rho : (0,1) \to [0,1]$ be arbitrary. The map $p \mapsto s = \sqrt{(1-p)/p}$ is a bijection of $(0,1)$ onto $(0,\infty)$, so prescribing ρ as a function of p is the same as prescribing the ratio $f(-1/s)/f(s)$ at every $s > 0$. Choose f arbitrarily and strictly positive on $(0,\infty)$ — for instance $f \equiv 1$ — and define, for $u < 0$,

$$f(u) \;=\; \rho\!\left(\frac{1}{1+u^2}\right) \cdot f\!\left(-\frac{1}{u}\right), \tag{13}$$

which is well defined because $u \mapsto -1/u$ maps $(-\infty, 0)$ bijectively onto $(0,\infty)$, and because $s = -1/u$ corresponds to $p = 1/(1+u^2)$. By construction $f(-1/s)/f(s) = \rho(p)$ for every p. Hence the correspondence between tilts and two-class empirical reweighting profiles is onto. This statement concerns the reward tilt; it does not imply that the full sequence-level target is determined by ρ alone. The q-exponential and exponential tilts, by contrast, trace only a two-parameter and a one-parameter subset respectively (Figure 7b).

For numerical stability, the implementation replaces the hard truncation $[u]_+$ with $\mathrm{softplus}_\beta(u) = \beta^{-1}\log(1 + \exp(\beta u))$, using $\beta = 5$, and adds $\eta = 10^{-6}$ before taking the power: $w =$

$[\mathrm{softplus}_5(\delta_{\mathrm{cfg}} + c_q\hat{r}) + \eta]^{1/(q-1)}$, where $c_q = (q-1)(3/q)$. This avoids zero weights, undefined logarithms, and unstable boundary gradients. All reported ρ-statistics and figures use the actual softplus transform.

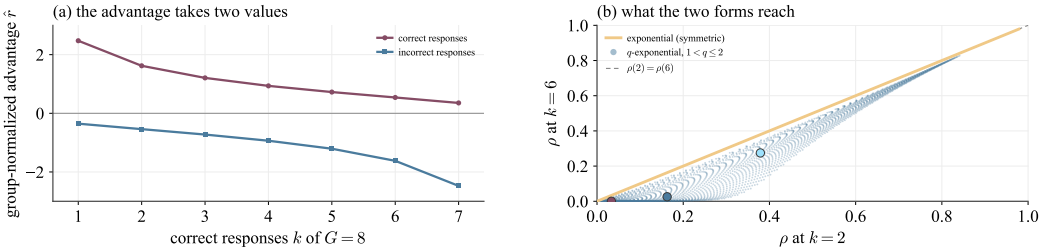

Figure 7: **(a)** The group-normalized advantage against group composition: two values on every informative group, diverging as p approaches a degenerate endpoint. At the endpoints themselves, the implementation sets the normalized advantage to zero. **(b)** Profile space projected on $(\rho(k{=}2), \rho(k{=}6))$. Every point of the square is an admissible profile by Proposition 1. The exponential family is confined to the diagonal, since σ_r is symmetric about $p = 1/2$; the q-exponential family fills a wedge below it, swept over $1 < q \leq 2$ and $\delta_{\mathrm{eff}} \in [0.2, 12]$. Markers are the three shifts at $q = 3/2$.

C.5 How much of the design space the available forms reach

Proposition 1 says the design space is the space of profiles $\rho : (0,1) \to (0,1]$, and the construction above shows every profile is realized by some tilt. At a fixed group size G only the values $\rho(k/G)$, $k = 1, \ldots, G-1$, enter any computation, so the design space relevant to training is $(G{-}1)$-dimensional; at $G = 8$ it has seven degrees of freedom.

The forms used here cover far less. The q-exponential tilt is parameterized by (q, δ) and the exponential by β alone, so they trace at most two- and one-dimensional subsets, respectively, of this $G{-}1$-coordinate space. Which subset is reached is also constrained in shape rather than only in dimension. Writing $t = 1/(q-1)$ for the exponent, the q-exponential profile is

$$\rho(p) \;=\; g(p)^t, \qquad g(p) \;=\; \frac{[\,\delta - 1/s\,]_+}{s + \delta}, \tag{14}$$

so $\log \rho = t \log g$: varying the index rescales $\log \rho$ uniformly in p without changing where it vanishes, and varying the shift moves the zero and deforms g, but g remains monotone in s and continuous. Three consequences follow, each visible in Figure 1.

First, each q-exponential profile is continuous: as $p \uparrow p^*$ the numerator $\delta - 1/s \downarrow 0$, so $\rho \to 0$ with a rate governed by t. A profile that holds ρ away from zero on the admitted set and then drops would require a discontinuity at p^*, which no member of either family has — contrary to what the phrase "truncation at p^*" suggests. What we call the shift is therefore a boundary at which the weight has already decayed over several group compositions: at the canonical pair, ρ has already fallen below 10^{-2} by $k = 4$, where the idealized boundary lies.

Second, the exponential family reaches no profile with a zero at all, since $\exp(\cdot) > 0$ everywhere, and is in addition symmetric (Proposition 2); a one-dimensional family constrained to be symmetric about $p = 1/2$ is a thin subset even of the profiles the q-exponential reaches.

Third, the two design goals in Paragraph 4.1 concern different features of the same curve: where it vanishes and how large it remains away from zero. The two parameters address these features separately, with δ controlling p^* and t controlling scale. Smoothness, however, restricts t: the truncated power is continuously differentiable at the boundary only for $t > 1$, equivalently $1 < q < 2$. At $q = 2$ it is continuous but has a kink. The smooth design corridor is therefore the image of $\{(q, \delta) : 1 < q < 2\}$, restricted to shifts whose p^* lies within the range of realized group compositions.

We do not characterize which profiles outside this image would be preferable; Proposition 1 says only that they exist and are reachable by some tilt.

Figure 8 places the evaluated configurations within the larger smooth target family. The sweep covers substantially different reweighting strengths, although only a small set of parameter pairs is trained.

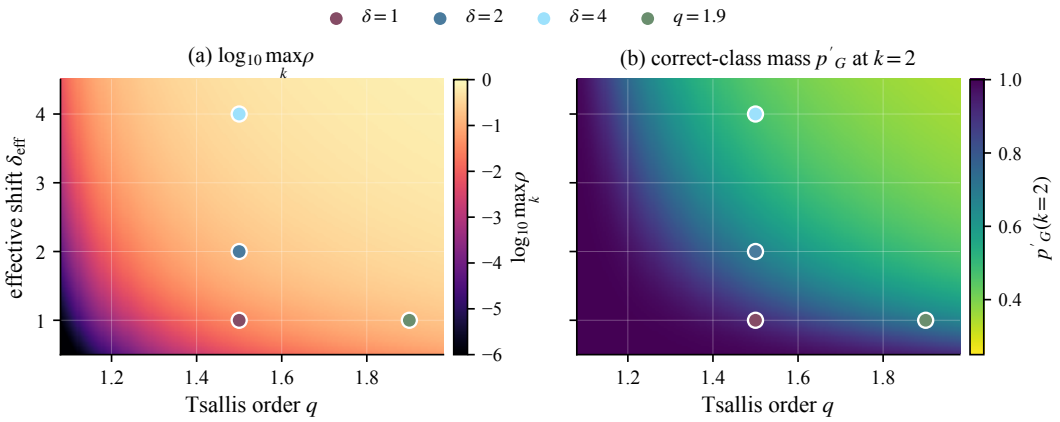

Figure 8: Target-design space under the implemented softplus transform. **(a)** Maximum incorrect-to-correct ratio across informative group compositions. **(b)** Correct-class mass at $k = 2$. Markers identify the evaluated configurations.

C.6 Rewards with more than two values

If the reward takes m distinct values on a group, the group-normalized advantage takes m values and any multiplicative tilt takes m values, of which one is absorbed by normalization. The target within a group is therefore determined by $m-1$ ratios, and the design space is $(m-1)$-dimensional. The collapse studied in this paper is the case $m = 2$; it is a property of the binary reward, not of the regularizer.

D The affine limit and matched ρ

The transform $\tilde{r} = \frac{1}{q-1}\log[\hat{r}+\delta]_+$ is singular at $q = 1$: the exponent diverges, and setting q numerically close to one gives values of $\tilde{r}$ in the thousands rather than the Shannon target. Instead, the limit is taken analytically. Inside the support, where the smoothing constant is negligible,

$$\tilde{r} = \frac{1}{q-1}\log\delta + \frac{1}{q-1}\log(1+\hat{r}/\delta) \longrightarrow \frac{1}{q-1}\log\delta + \alpha\,\hat{r}, \qquad \alpha = \frac{1}{(q-1)\delta}, \tag{15}$$

and the constant is absorbed by Z_ϕ, leaving the affine transform $\tilde{r} = \alpha\hat{r}$ used in the Shannon runs. The reweighting ratio is then $\rho(p) = \exp(-2\alpha/\sigma_r)$, the two classes differing by exactly 2 in raw reward.

Symmetry makes matching approximate. Since σ_r depends on p only through $p(1-p)$, the affine profile satisfies $\rho(p) = \rho(1-p)$ at every α (Proposition 2), while the q-exponential profile is skewed towards low-competence groups. No α therefore reproduces a q-exponential profile, and the choice of α is interpreted as a choice of where to agree. Matching the $\delta = 2$ profile at the operating point $k = 2$ gives $\alpha = 0.841$; matching its maximum gives $\alpha = 0.971$. We use $\alpha = 0.93$ for approximate minimization of the mean absolute log discrepancy over the informative range.

The residual in Table 2 is not a defect of the matching. The reason is the content of Proposition 2: the affine profile peaks at $k = G/2$ and is symmetric about it, the q-exponential peaks at $k = 2$ and decays to zero at k^*. The tail ratio diverges because the exponential tilt has no zero to reach. Comparisons involving the affine run are therefore read as comparisons at similar scale and different shape, not at a matched profile.

Table 2: Residual mismatch between the q-exponential profile at $\delta = 2$ and the affine profile at $\alpha = 0.93$, $G = 8$. These two agree to within a third over $k \leq 4$ and diverge without bound at the truncation boundary, where one profile reaches zero and the other cannot.

k	1	2	3	4	5	6	7
ρ, q=3/2, δ=2	0.1354	0.1627	0.1582	0.1318	0.0854	0.0258	0.00006
ρ, affine, α=0.93	0.0721	0.1341	0.1658	0.1755	0.1658	0.1341	0.0721
ratio	0.5	0.8	1.0	1.3	1.9	5.2	1.3×10^3

The published temperature. With σ_r maximized at $p = 1/2$ the affine ratio is bounded uniformly, $\rho \leq \exp(-2\beta/\sigma_{\max})$; at $G = 8$ and the published $\beta = 15$ this is 6.5×10^{-13}, the figure quoted in Paragraph 4.1. We reserve $\beta = 15$ for the published FlowRL temperature and use α for the affine coefficient in our matched-profile control.

E Training and evaluation protocol

E.1 Training

All runs except $q = 1.9$ and GRPO fine-tune with veRL and vLLM rollouts on 2× NVIDIA H100 80GB GPUs. The $q = 1.9$ and GRPO runs use 2× PPU-ZW810E-96GB accelerators. We use batches of 256 trajectories, comprising 32 distinct prompts with $G = 8$ generations per prompt, learning rate 10^{-6} with a linear warmup of 10 iterations, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$), framework-default gradient clipping, and maximum prompt and response lengths of 2048 and 8192 tokens, respectively. Training rollouts are sampled at temperature 1.0 without nucleus truncation. The training set is DAPO-Math-17k (Yu et al., 2026), with a binary verifiable reward $r \in \{-1, +1\}$. The learned normalizer Z_ϕ is a 3-layer MLP over the masked mean of prompt-token hidden states, so $\log Z_\phi$ depends on the prompt alone.

Cold runs train for 400 steps from the initialization. Warm continuations resume from a common step-200 checkpoint ($\delta = 1$) and run for 100 further steps, except the long-horizon run, which continues to step 400. The q-exponential runs use $q = 3/2$ and the shifts stated in the main text; the affine run uses $\tilde{r} = \alpha\hat{r}$ with $\alpha = 0.93$, matched to the $\delta = 2$ target at the operating point (Appendix D). Configuration entries are passed as δ_{cfg} and rescaled internally by $c_q = (q-1)(3/q)$, so the effective shift is $\delta_{\text{eff}} = \delta_{\text{cfg}}/c_q$; at $q = 3/2$ the factor is exactly one, and the distinction matters only for runs at other indices.

The $q = 1.9$ experiment is a 100-step continuation from its documented step-200 warm start. GRPO is resumed from a valid step-200 GRPO checkpoint and trained for 200 additional steps; its reported checkpoint therefore has a total training horizon of 400 steps. The `step200` string retained in the artifact name describes the restart checkpoint and not the final horizon.

We use one seed for each configuration, with two configurations repeated under a second independent seed. This supplies the variance estimate of Section 5.2 and also prevents us from ordering configurations that differ by less than it.

E.2 Evaluation

Problem pool. We use 90 problems, comprising 30 problems each from AIME24, AIME25, and AIME26 (MAA, 2025). We draw $n = 128$ samples per problem at temperature 0.7 and top-$p = 0.95$. We compute unbiased $\text{Pass@}k$ as $|Q|^{-1} \sum_i \left[1 - \binom{n_i - c_i}{k} / \binom{n_i}{k}\right]$ (Chen et al., 2021), and retain per-problem counts c_i for the reachability analysis of Appendix H.

Pairing. We compute differences against a comparator on matched (problem, sample index) pairs. We verified the matching directly: for a sample of 2000 shared keys between two evaluation files the logged sampling seed agrees in every case, which is what makes the paired bootstrap and the McNemar test valid here. Intervals are 95% cluster bootstrap over problems, 1000 resamples, seed 42; each resampled problem's 128 samples are kept together. McNemar is the exact two-sided

binomial test on the discordant-pair counts. The two answer different questions: the magnitude of the mean paired difference under problem-level resampling, and whether the discordant counts alone are inconsistent with an even split. They are not expected to agree, so we report both rather than selecting one.

Integrity. A single verifier commit scores all evaluation files, and we record its hash alongside each. We apply the same chat-template convention to every checkpoint; the cost of that is quantified in Section 5.2.

E.3 The fixed trajectory set

Token-level diagnostics, including argmax divergence, off-support fraction and entropy, are computed under teacher forcing on a fixed set of 187 generations from 153 problems, comprising 189,972 evaluated tokens and produced by the initialization, reused unchanged for every checkpoint. This prevents checkpoint differences from being confounded with differences in generated trajectories. Diagnostic files for each checkpoint contain an identical set of (problem, sample index) keys, which we verified pairwise. The set comprises the initialization trajectories in an unweighted random draw from pooled held-out generations of three models, taken without stratification and without filtering by length, parseability, or truncation. Its correctness rate differs from the initialization's held-out accuracy by less than the standard error of a draw this size, so we do not treat it as class-imbalanced by selection.

E.4 Compute

Each 100 training steps take approximately 6 wall-clock hours on the 2×H100 setup and 8 hours on the 2×PPU-ZW810E setup. The controlled H100 sweep comprises 2,300 training steps: one shared 200-step warm-start prefix, three 400-step cold runs, seven 100-step warm continuations, and one 200-step long continuation. This corresponds to approximately

$$\frac{2300}{100} \times 6 \times 2 = 276 \quad \text{H100-hours.}$$

The PPU experiments comprise 500 steps: the shared 200-step GRPO prefix, its 200-step continuation, and the 100-step $q = 1.9$ continuation. Counting the shared prefix once gives

$$\frac{500}{100} \times 8 \times 2 = 80 \quad \text{PPU-ZW810E-hours.}$$

Generating one $n = 128$ AIME pool takes approximately 1.5 hours on two accelerators. Counting the 16 pool-generation jobs used in the controlled analysis gives approximately 48 additional accelerator-hours. Teacher-forced diagnostics take under one minute per checkpoint and contribute less than 0.5 accelerator-hours in total. Thus the accounted controlled study uses approximately 276 H100-hours, 80 PPU-ZW810E-hours, and 48 accelerator-hours for evaluation-pool generation. CPU-only scoring, bootstrap analysis, and figure generation are excluded. The earlier exploratory sweep is also excluded because complete job-level timing records for it were not retained.

F Variance of the procedure

The controls in this appendix measure different sources of variation and are therefore not pooled into a single error bar. The top block of Table 3 contains historical controls under the earlier five-benchmark protocol. The bottom block uses the matched-pair $\mathrm{Pass@1}$ comparison on the present 90-problem AIME pool. Its purpose is narrower: to establish the scale against which differences within the δ-sweep should be read.

Figure 9 compares the target effect with the measured variation of the training and evaluation procedure. The change across the full δ-sweep is smaller than several nuisance effects.

The $\delta = 2$ seed comparison should be read conservatively: the seed-43 checkpoint available for evaluation is at step 250, whereas its matched run is evaluated at step 300. We exclude the $q = 1.9$ continuation from the variance estimate because it changes both the Tsallis order and the optimization trajectory.

Table 3: Variance controls measured before interpreting the sweep. The top block reports accuracy points under the earlier protocol; the bottom block reports pooled paired Pass@1 differences on the 90-problem pool at $n = 128$. All magnitudes are reported in accuracy percentage points.

Source	Comparison	Magnitude
Re-evaluation, mean accuracy	same weights, alternate evaluation pipeline	3.33
Re-evaluation, Pass@8	same weights and protocol family	7.50
Reseeding one configuration	configuration and data fixed	2.71 [0.58, 4.96]
Analytic 2σ threshold	300 prompts at $n = 8$	2.70
Chat-template convention	same base checkpoint	6.96 [5.04, 9.00]
Seed, $\delta = 2$	warm runs	0.13
Seed, $\delta = 1$	warm runs	0.45
Warm versus cold, $\delta = 4$	shift fixed	1.06
Effect of δ from 1 to 4	mean over q-exponential rows	0.61

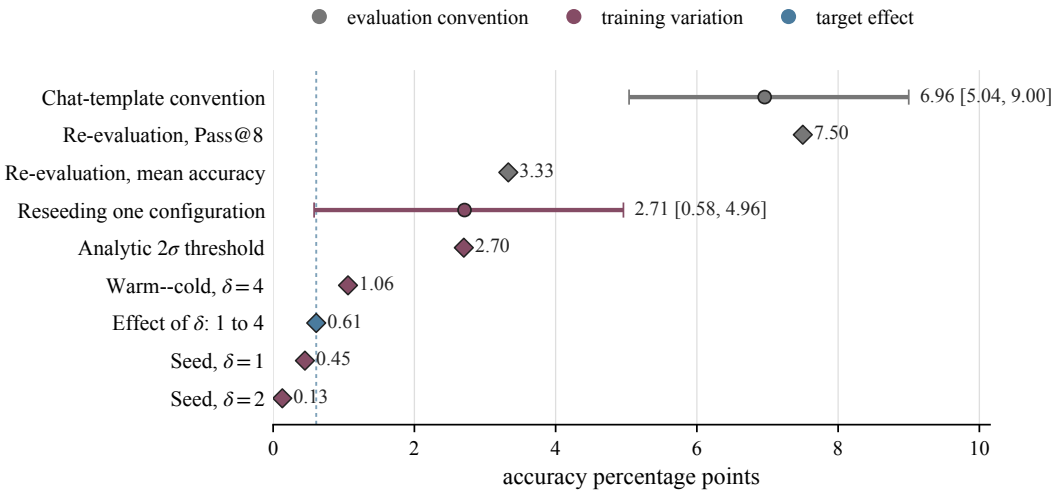

Figure 9: Empirical resolution of the procedure. Intervals are shown where repeated estimates are available; the other points report measured magnitudes. The dashed line marks the accuracy change across the complete $\delta = 1$ to $\delta = 4$ sweep.

Weight displacement. For each run, we compute the relative parameter displacement

$$d_{\text{rel}} = \frac{\|\theta_{\text{final}} - \theta_{\text{init}}\|_2}{\|\theta_{\text{init}}\|_2},$$

after flattening the corresponding policy checkpoint tensors. Cold runs have $d_{\text{rel}} = 0.00077$–0.00088, while warm continuations have $d_{\text{rel}} = 0.00054$–0.00062. Thus, under this diagnostic, the start regime explains more variation than the shift within either regime. Weight displacement is only a checkpoint-space diagnostic and is not used as a proxy for behavioral distance.

G PER-RUN TARGET AND CHECKPOINT DIAGNOSTICS

Figure 4 summarizes the checkpoint comparison visually. Table 4 gives its exact numerical values, while Table 5 reports the training-log quantities from which the target diagnostics are computed.

Table 4: Numerical values underlying Figure 4. Target statistics are averaged over the last 40 logged steps. Differences and 95% confidence intervals are paired against the corresponding starting checkpoint.

		Target			Measured	
Configuration	Regime	$\max_k \rho$	$\bar{p}'$	$\rho < 10^{-2}$	P@1	Δ [95% CI]
initialization	–	–	–	–	.0908	–
$\delta = 1$	warm	.0354	.933	.489	.0905	$-.0003$ [$-.011$, .011]
$\delta = 1$, seed 44	warm	.0354	.932	.478	.0950	$+.0042$ [$-.006$, .016]
$\delta = 1$	cold	.0354	.929	.456	.1063	$+.0155$ [$-.003$, .030]
$\delta = 2$	warm	.1627	.779	.131	.0964	$+.0056$ [$-.004$, .016]
$\delta = 2$, seed 43	warm	.1627	.795	.114	.0951	$+.0043$ [$-.006$, .015]
$\delta = 2$	cold	.1627	.774	.129	.0979	$+.0071$ [$-.006$, .020]
$\delta = 2$, 400 steps	warm	.1627	.800	.155	.0974	$+.0066$ [$-.008$, .022]
$\delta = 4$	warm	.3956	.631	.000	.0950	$+.0042$ [$-.006$, .016]
$\delta = 4$	cold	.3956	.631	.000	.1056	$+.0148$ [.002, .030]
affine, $\alpha = .93$	warm	.1755	.811	.000	.1000	$+.0092$ [$-.001$, .021]
$\delta = 1$, $G = 32$	warm	.0319	–	–	.1015	$+.0107$ [$-.001$, .023]
$q = 1.9$ warm start	–	–	–	–	.0969	–
$q = 1.9$, $\delta_{\text{eff}} = 1$	warm	.1547	–	–	.0375	$-.0594$ [$-.100$, -.023]
GRPO, step-200 start	–	–	–	–	.0969	–
GRPO, 400 steps	cont.	–	–	–	.1135	$+.0166$ [.003, .033]

We extract the target diagnostics from training logs and average them over the last 40 distinct steps of each run. The logged group histogram counts groups by the number c of correct responses among the G samples. Its mass at $c = 0$ and $c = G$ gives the degenerate fractions; all remaining groups are informative. We compute $\bar{p}'$ by weighting each informative composition by its realized frequency, and report the fraction for which $\rho < 10^{-2}$. The *Informative* column is the total number of informative group instances in the averaging window.

Two rows require qualification. The $G = 32$ run records its group histogram only over $c = 0, \ldots, 8$, so its group-level fractions cannot be recovered; the reported ρ comes from the correctly computed aggregate in the training code. For the second $\delta = 2$ seed, the latest complete checkpoint is at step 250 rather than 300, so its comparison also includes a 50-step horizon difference.

Loss scale. The configurations change residual magnitude as well as target shape. The transformed reward spans 20.4, 11.1, and 2.9 nats across classes and group compositions for $\delta = 1, 2, 4$, respectively. Because the residual enters equation 2 squared, these configurations expose the optimizer to different loss scales. Gradient clipping limits, but does not eliminate, this difference.

McNemar tests. For the eleven common-initialization TB runs, the exact sample-level McNemar p-values are: $\delta = 1$, warm 0.95, seed 44 0.13, cold 0.022; $\delta = 2$, warm 0.038, seed 43 0.12, cold 0.010, 400-step warm 0.017; $\delta = 4$, warm 0.13, cold 7×10^{-8}; affine 0.0009; and $G = 32$, 0.0001.

Table 5: Training-log target diagnostics. Degenerate fractions are computed over all groups; the suppressed fraction and $\bar{p}'$ use informative groups only.

Configuration	Window	$c = 0$	$c = G$	Informative	$\rho < 10^{-2}$	$\bar{p}'$	$\bar{p}$
$\delta = 1$ w	261–300	.429	.105	597	.489	.933	.317
$\delta = 1$ w, s44	261–300	.390	.102	651	.478	.932	.327
$\delta = 1$ c	361–400	.382	.130	625	.456	.929	.345
$\delta = 2$ w	261–300	.427	.115	586	.131	.779	.315
$\delta = 2$ w, s43	260–299	.401	.120	613	.114	.795	.339
$\delta = 2$ c	361–400	.387	.129	620	.129	.774	.339
$\delta = 2$ w, 400	361–400	.423	.113	594	.155	.800	.330
$\delta = 4$ w	261–300	.425	.109	597	.000	.631	.314
$\delta = 4$ c	361–400	.384	.118	638	.000	.631	.338
affine, $\alpha = .93$	261–300	.414	.105	615	.000	.811	.319
$\delta = 1$, $G = 32$	261–300	–	–	–	–	–	.335

Table 6: Per-run reachability against the corresponding initialization. G_1/L_1 denote gained/lost problems at ordinary reachability, and G_5/L_5 those at the robust floor ($c^* = 5$). The McNemar test is exact and two-sided at $c^* = 1$. All rows except ($q = 1.9$) use the common instruction-tuned initialization; ($q = 1.9$) uses its documented warm start.

Configuration	Regime	G_1	L_1	G_5	L_5	p_{McN}
same initialization, two pipelines	control	6	3	3	2	.508
($\delta = 1$)	warm	4	7	1	1	.549
($\delta = 1$), seed 44	warm	5	6	2	1	1.000
($\delta = 1$)	cold	7	7	3	0	1.000
($\delta = 2$)	warm	5	5	5	3	1.000
($\delta = 2$), seed 43	warm	5	4	3	2	1.000
($\delta = 2$)	cold	5	7	3	3	.774
($\delta = 2$), 400 st.	warm	3	6	2	3	.508
($\delta = 4$)	warm	6	6	1	2	1.000
($\delta = 4$)	cold	7	4	5	1	.549
affine, ($\alpha = .93$)	warm	4	6	4	0	.754
($\delta = 1$), ($G = 32$)	warm	6	7	2	1	1.000
GRPO, 400 steps	cont.	7	6	5	1	1.000
($q = 1.9$), ($\delta_{\mathrm{eff}} = 1$)	warm	5	24	0	9	5.46×10^{-4}

Seven are below 0.05, whereas only the cold $\delta = 4$ paired bootstrap interval in Table 4 excludes zero. This difference is expected: McNemar's test uses only discordant sample counts, whereas the bootstrap estimates the mean paired effect while preserving problem-level clustering.

H REACHABILITY TABLES

For problem i, let $c_i(\pi)$ be the number of correct samples among $n = 128$ generations from policy π. At a floor c^*, the reached set is

$$\mathcal{R}_\pi(c^*) = \{i : c_i(\pi) \geq c^*\}.$$

We call $\mathcal{R}_\theta(c^*) \setminus \mathcal{R}_{\mathrm{init}}(c^*)$ gained problems and $\mathcal{R}_{\mathrm{init}}(c^*) \setminus \mathcal{R}_\theta(c^*)$ lost problems. Thus $c^* = 1$ is ordinary reachability, while $c^* = 5$ requires at least five successful samples from the policy.

The comparison is recomputed after thresholding both policies at every c^*. Consequently, gained and lost counts need not be monotone in c^*: a problem reached by both policies at $c^* = 1$ can become discordant at a higher floor. We therefore report the entire curve in Figure 3 and summarize representative thresholds below.

Table 7: Sensitivity of total discordance ($N_{c^*} = G_{c^*} + L_{c^*}$) to the solved-count floor. Values summarize the full curves in Figure 3; they are descriptive and are not separate hypothesis tests.

Configuration	Regime	N_1	N_5	N_{10}	N_{20}	N_{50}	N_{100}
same initialization with two pipelines	control	9	5	2	0	1	0
($\delta = 1$)	warm	11	2	2	1	1	0
($\delta = 1$), seed 44	warm	11	3	3	1	2	3
($\delta = 1$)	cold	14	3	6	1	3	0
($\delta = 2$)	warm	10	8	1	2	0	2
($\delta = 2$), seed 43	warm	9	5	1	1	0	1
($\delta = 2$)	cold	12	6	2	4	3	3
($\delta = 2$), 400 steps	warm	9	5	3	3	2	1
($\delta = 4$)	warm	12	3	1	1	1	1
($\delta = 4$)	cold	11	6	7	3	4	2
affine, ($\alpha = .93$)	warm	10	4	3	2	0	1
($\delta = 1$), ($G = 32$)	warm	13	3	1	2	1	1
GRPO, 400 steps	cont.	13	6	5	1	3	3
($q = 1.9$), ($\delta_{\text{eff}} = 1$)	warm	29	9	10	9	7	2

Table 8: Where discordance occurs for the three cold runs. The first two columns are fractions of matched samples that disagree between the checkpoint and initialization, stratified by the initialization's solved count. The ratio is their quotient.

Configuration	$c_{\text{init}} \geq 5$	$c_{\text{init}} < 5$	Ratio
$\delta = 1$, cold	.3274	.0107	30.6
$\delta = 2$, cold	.3142	.0087	36.1
$\delta = 4$, cold	.3054	.0106	28.8

Table 6 shows that the common-initialization sweep produces gained and lost counts on the scale of the pipeline control. No configuration yields a significant problem-level McNemar test at $c^* = 1$. The $q = 1.9$ continuation differs qualitatively: it loses substantially more problems than it gains, and has no gained transition at the robust floor.

The original sweep produces discordance of the same order as the pipeline control at both ordinary and robust reachability. The $q = 1.9$ row remains exceptional over a wide range of floors: its 29 discordant problems at $c^* = 1$ reduce at higher floors, but remain predominantly losses.

Table 8 shows that the observed churn in the cold sweep is concentrated on problems already reached reliably by the initialization. For the $q = 1.9$ continuation, the analogous problem-level split is also unfavourable: among 22 problems with $c_{\text{init}} \geq 5$, it gains none and loses seven; among the remaining 68 problems, it gains five and loses seventeen. Thus, neither the common sweep nor the $q = 1.9$ run provides evidence that the target expands reachability on previously hard problems.

I THE EARLIER SWEEP IN FULL

These runs predate the instrument of Section 5.2 and were produced under a different protocol: 400 training steps, $\mathrm{Avg@8}$ and $\mathrm{Pass@8}$ read from a single draw of eight samples per problem. Emphasis in the tables marks the best and second-best fine-tuned entry as originally reported; Section 5.1 states why we do not read an ordering from it.

J DIVERSITY

We report three complementary quantities. D_{all} is the plug-in Gini–Simpson index over final answers across all 128 samples per problem. D_{inc} is the unbiased Gini–Simpson index restricted to incorrect answers. $D_{\text{cor}}^{\text{Rao}}$ is the unbiased Rao diversity index induced by 4-gram Jaccard distance be-

Table 9: Mathematical reasoning benchmarks (%), Avg@8 and Pass@8 from the same eight samples per problem. Best and second-best among the fine-tuned methods per model and metric are **bold** and *italic*; Base is the untuned reference and is not ranked.

Model	Method	MATH-500	AMC-23	AIME24	AIME25	AIME26	Mean
	Average accuracy over eight samples (Avg@8)						
Qwen2.5-3B	untuned	45.63	25.63	3.75	2.08	1.67	15.75
	reward maximization	55.63	36.25	1.67	*2.50*	2.08	19.63
	exponential	*59.75*	*40.00*	**4.58**	2.08	**4.17**	*22.12*
	q-exponential, δ=1	**60.13**	**42.50**	**4.58**	**2.92**	*3.33*	**22.69**
Qwen2.5-7B	untuned	41.00	27.19	6.25	4.17	2.92	16.31
	reward maximization	68.25	45.94	8.75	*6.67*	*6.67*	27.26
	exponential	**71.50**	*51.25*	*11.67*	5.00	4.17	*28.72*
	q-exponential, δ=1	*70.25*	**51.88**	**15.00**	**9.58**	**7.92**	**30.93**
	Problems solved by at least one of eight samples (Pass@8)						
Qwen2.5-3B	untuned	79.00	62.50	16.67	3.33	13.33	34.97
	reward maximization	73.00	62.50	10.00	10.00	**13.33**	33.77
	exponential	**83.00**	*72.50*	**20.00**	*13.33*	*10.00*	*39.77*
	q-exponential, δ=1	*78.00*	**77.50**	*16.67*	**16.67**	**13.33**	**40.43**
Qwen2.5-7B	untuned	82.00	65.00	23.33	16.67	6.67	38.73
	reward maximization	**85.00**	65.00	23.33	*23.33*	13.33	42.00
	exponential	*84.00*	**85.00**	*26.67*	*23.33*	**16.67**	*47.13*
	q-exponential, δ=1	**85.00**	*82.50*	**30.00**	**26.67**	**16.67**	**48.17**

Table 10: Code generation benchmarks on DeepSeek-R1-Distill-Qwen-7B. Best and second-best among the fine-tuned methods are **bold** and *italic*; Base is the untuned reference and is not ranked. The Codeforces rating and percentile are derived from Pass@8.

	LiveCodeBench		HumanEval+		Codeforces		
Method	Avg@8	Pass@8	Avg@8	Pass@8	Pass@8	Rating	Percentile
basemodel	30.47	43.73	80.67	92.64	17.90	723.7	5.70
reward-max	30.60	46.59	71.93	**93.25**	20.84	890.6	19.9
exponential	**34.54**	*49.82*	*81.44*	**93.25**	*29.17*	*1204.9*	*56.3*
q-exponential, δ=1	*33.60*	**50.54**	**82.52**	**93.25**	**29.66**	**1220.7**	**58.2**

Table 11: Answer-level and class-conditioned diversity for the common- initialization sweep. D_{all} is diversity of final answers; D_{inc} controls for changes in the number of correct answers; $D_{\text{cor}}^{\text{Rao}}$ measures diversity of correct response texts. Entries are point estimates with 95% cluster-bootstrap intervals.

Configuration	D_{all}	D_{inc}	$D_{\text{cor}}^{\text{Rao}}$
initialization	.833 [.797, .868]	.872 [.841, .899]	.878 [.865, .894]
$\delta = 1$, warm	.797 [.758, .832]	.833 [.797, .864]	.842 [.826, .859]
$\delta = 1$, warm, seed 44	.803 [.762, .842]	.847 [.811, .877]	.835 [.818, .850]
$\delta = 1$, cold	.795 [.757, .833]	.835 [.798, .868]	.858 [.843, .875]
$\delta = 2$, warm	.795 [.751, .835]	.837 [.803, .871]	.843 [.821, .863]
$\delta = 2$, warm, seed 43	.814 [.776, .849]	.851 [.818, .879]	.837 [.819, .856]
$\delta = 2$, cold	.797 [.755, .839]	.839 [.801, .872]	.807 [.785, .830]
$\delta = 2$, warm, 400 steps	.800 [.758, .838]	.849 [.818, .875]	.828 [.809, .847]
$\delta = 4$, warm	.814 [.773, .849]	.857 [.823, .884]	.842 [.823, .863]
$\delta = 4$, cold	.784 [.741, .827]	.830 [.796, .863]	.840 [.822, .859]
affine, $\alpha = .93$, warm	.813 [.774, .850]	.858 [.826, .882]	.838 [.818, .858]
$\delta = 1$, $G = 32$, warm	.786 [.745, .824]	.830 [.794, .860]	.830 [.815, .846]
GRPO, 400 steps	.801 [.761, .839]	.848 [.815, .877]	.850 [.831, .868]

tween correct reasoning traces; this is needed because correct final AIME answers are almost always identical. Class- conditioned quantities use the common eligible-problem set, since the number of correct samples differs across checkpoints. All intervals are 95% cluster-bootstrap intervals over problems.

Paired entropy comparison. All teacher-forced checkpoints are evaluated on the same trajectories. We therefore average token entropy within each trajectory and apply a paired cluster bootstrap over problems, retaining all trajectories from a resampled problem together. For the horizon-matched comparison between GRPO and the 400-step $\delta = 2$ continuation,

$$H_{\text{GRPO}} - H_{\delta=2} = 0.0312 \quad [0.0259, 0.0377].$$

The interval excludes zero. The confidence intervals reported separately for other rows in Table 12 are marginal intervals and are not used as pairwise significance tests.

The initialization has the largest point estimate of overall answer diversity. This ordering is not explained only by accuracy: every trained checkpoint also falls below the initialization in the incorrect-answer and correct-text controls. The intervals are reported to show estimator resolution; the intended comparison is the paired ordering across a common problem set, not independent significance tests for every table cell.

All available fine-tuned checkpoints have lower mean token entropy than the initialization. At the same time, argmax divergence remains below 2.9% and off-support generation below 0.023%. Thus, fine-tuning predominantly sharpens existing token preferences rather than moving probability mass to a new support. The $\delta = 1$ cold run is the sole checkpoint with a step mismatch: answer-level statistics use its step-300 evaluation, whereas the only surviving teacher-forcing artifact is from step 400.

The two diversity measures answer different questions. Across the eleven trajectory-balance checkpoints, entropy correlates with correct-text diversity ($r = 0.74$), but not with final-answer diversity ($r = 0.01$). The former tracks local predictive uncertainty; the latter is constrained by AIME's small answer space.

The $q = 1.9$ continuation lowers overall and incorrect-answer diversity relative to its own warm start. Its correct-answer estimate is based on only 20 problems with at least two correct samples, compared with 35 for the warm start, and is consequently not interpreted as evidence for a change in correct-trace diversity. Teacher-forced diagnostics for $q = 1.9$ were not available at the time of analysis.

Table 12: Teacher-forced token-level diagnostics on the fixed set of 187 initialization trajectories. H_{tok} is mean per-token entropy. ΔH_{len} is the incorrect-minus-correct entropy gap after controlling for response length. Argmax divergence is measured against the instruction-tuned initialization; off-support denotes generated tokens with reference probability below 10^{-4}.

Configuration	eval./TF step	H_{tok} [95% CI]	ΔH_{len}	argmax div. (%)	off-support (%)
initialization	0 / 0	.159 [.130, .197]	.085	–	.0005
$\delta = 1$, warm	300 / 300	.121 [.099, .149]	.067	2.081	.0032
$\delta = 1$, warm, seed 44	300 / 300	.114 [.094, .140]	.063	1.894	.0026
$\delta = 1$, cold	300 / 400	.116 [.094, .144]	.068	2.764	.0063
$\delta = 2$, warm	300 / 300	.124 [.102, .153]	.068	2.337	.0032
$\delta = 2$, warm, seed 43	250 / 250	.115 [.094, .141]	.065	1.774	.0042
$\delta = 2$, cold	400 / 400	.102 [.083, .126]	.061	2.630	.0068
$\delta = 2$, warm, 400 steps	400 / 400	.096 [.078, .118]	.059	2.849	.0153
$\delta = 4$, warm	300 / 300	.113 [.093, .139]	.064	2.114	.0068
$\delta = 4$, cold	400 / 400	.117 [.096, .145]	.071	2.879	.0221
affine, $\alpha = .93$, warm	300 / 300	.117 [.096, .144]	.065	2.452	.0053
$\delta = 1$, $G = 32$, warm	300 / 300	.109 [.089, .134]	.062	2.265	.0032
GRPO, continued, 400 steps	400 / 400	.127 [.104, .156]	.068	2.779	.0047

Table 13: Available $q = 1.9$ answer-diversity comparison against its own step-200 warm start. Correct-answer diversity here is computed over final answers and is therefore not comparable to the correct-text Rao–Jaccard index in Table 11.

Checkpoint	D_{all}	D_{inc}	D_{cor}
warm start	.809 [.769, .844]	.849 [.815, .878]	.011 [0, .032]
$q = 1.9$, step 100	.796 [.759, .832]	.801 [.764, .833]	.018 [0, .054]

J.1 Semantic-judge protocol

We independently repeat the semantic-diversity evaluation using Claude Sonnet 4.6. For each item, the judge groups eight complete solutions by their underlying method and assigns a variety score from 1 to 5 according to the following rubric.

We use one fixed prompt and item-selection rule throughout. Each item is judged once, so the reported scores are descriptive and do not quantify judge-repeat variability. The direction agrees with FlowRL, although the separation from reward maximization is smaller. The complete prompt and evaluation rubric are included in the supplementary archive.

The semantic judge and the local diagnostics do not estimate the same latent quantity. The judge evaluates methodological variation across complete solutions, whereas Gini–Simpson diversity measures variation in extracted answers and token entropy measures local predictive uncertainty. In the matched teacher-forcing comparison, the 400-step trajectory-balance checkpoint retains significantly less token entropy than GRPO; the remaining trajectory-balance point estimates are also no larger, despite their higher semantic-variety scores.

Table 14: Semantic methodological variety in the earlier sweep, judged by Claude Sonnet 4.6. This is a single judging pass under a different horizon and protocol from the checkpoint-level measurements above; no confidence interval is available.

Objective	judged variety (1–5)
Reward maximization	1.44
Exponential target	1.93
q-exponential target	1.87

Table 15: Rubric for LLM-judged methodological variety. A cluster contains solutions that share the same core setup and solution strategy.

Score	Criterion
1	**Minimal variety:** essentially one method cluster, containing 7–8 attempts. Differences are limited to arithmetic, notation, or phrasing.
2	**Low variety:** the largest cluster contains 6 attempts, with at most one genuine alternative. Variation remains anchored to one dominant method.
3	**Moderate variety:** the largest cluster contains 4–5 attempts, with approximately two distinct strategies and noticeable differences in setup or technique.
4	**High variety:** the largest cluster contains at most 3 attempts, with approximately three distinct strategies and several mathematical techniques or problem framings.
5	**Maximum variety:** no cluster contains more than 2 attempts and at least four distinct strategies are represented, spanning substantially different techniques or lines of attack.