STRATA: PRESERVING RARE SOLUTION STRATEGIES IN REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) is usually audited with pass@k, which overlooks the *diversity of correct solution strategies*. On a controlled benchmark with exhaustively enumerated, generator-verified strategy sets, we measure *rare-strategy survival* and find a dose-response decoupling: as GRPO training grows from 0 to 14.9M supervised tokens, AIME24 pass@64 rises ($.400 \rightarrow .467$) while survival falls overall ($1.000 \rightarrow .122$). DPH-RL preserves pass@k but retains only .146 of rare strategies; WiSE-FT weight interpolation reaches .390. We introduce STRATA (STrategy-Targeted Rehearsal Anchoring): verified teacher traces are re-scored under the current policy every R steps, and cross-entropy rehearsal targets low-likelihood traces at risk of being forgotten. At 3M supervised tokens—one fifth of overtrained GRPO's budget—STRATA matches its best AIME24 pass@64 (.467), ties best AIME25 (.433) and AMC23 (.925) in our twelve-condition matrix, and retains $2.6\times$ standard GRPO's rare strategies (.626 vs .244) and $5.1\times$ overtrained GRPO's (.626 vs .122). Five-seed results, matched-compute controls, selector validation, and a manually annotated natural-solution audit support the benefit across model families and evaluation regimes. We release all code, the benchmark, and the matched audit protocol for reuse.

1 INTRODUCTION

RLVR reinforces verified solutions (Shao et al., 2024; Yang et al., 2025), but can trade pass@k for pass@1: "diversity collapse" (Li et al., 2026; Zhou, 2026; Yuan et al., 2026). Remedies include mass-covering anchors (Li et al., 2026) and weight interpolation (Wortsman et al., 2022; Dang et al., 2025). STRATA adds targeted rehearsal without an inference-time teacher or trace bank (Fig. 1).

Pass@k measures solvable problems, not surviving solution strategies. Ten correct samples using one strategy or ten different strategies have identical pass@k. Strategy coverage matters for structural transfer, best-of-n selection, and search, but natural benchmarks lack exhaustive strategy inventories. We connect a controlled strategy-loss audit to targeted rehearsal through three main contributions:

1. **An exact instrument.** Rewrite-Math generates 600 algebra instances with known strategy sets. Grammar signatures identify case analysis, symmetry, substitution, factorization, telescoping, and expansion, enabling exact effective counts and rare-strategy survival.
2. **A dose-response audit.** Under a shared protocol, standard and overtrained GRPO gain 3.3 and 6.7 AIME24 pass@64 points, and TTRL gains 6.7 on AIME25, while losing 76–88% of rare strategies. FKL, DPH-RL, and WiSE-FT fall short on accuracy or survival (Tab. 1).
3. **Targeted rehearsal.** Current likelihood identifies verified teacher traces at risk for a GRPO rehearsal loss, adding $< 7\%$ step-time overhead in our measured training setup (Sec. 4).

2 RELATED WORK

RLVR and its collapse. GRPO (Shao et al., 2024) popularized critic-free RLVR. Subsequent work studies pass@k inversion (Zhou, 2026; Yuan et al., 2026), entropy collapse (Aphale & Liu, 2026),

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

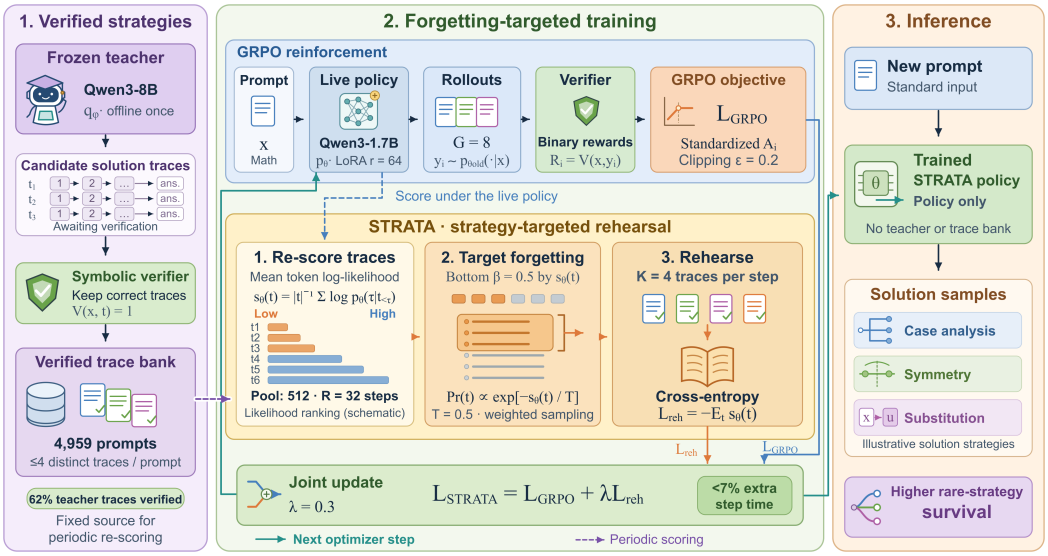


Figure 1: **STRATA training and inference.** A frozen teacher supplies strategy-distinct, verified traces. Every R=32 steps, the live policy re-scores a trace pool and targets its bottom-likelihood half for rehearsal. The joint objective $\mathcal{L}_{\text{GRPO}} + \lambda\mathcal{L}_{\text{reh}}$ trains the policy used alone at inference time.

and data-driven loss of reasoning modes (Nguyen et al., 2026). We complement these analyses with complete valid-strategy inventories and rare-correct-strategy survival under a matched RLVR audit.

Coverage-preserving objectives. DPH-RL (Li et al., 2026) uses mass-covering f-divergences to the initial policy on stored base samples. Differential smoothing (Gai et al., 2025) and anchored policy optimization (Wang et al., 2026) modify updates; outcome-based exploration (Song et al., 2025) rewards answer diversity; TTRL (Zuo et al., 2026) uses majority-vote pseudo-rewards. DPH-RL retains .146 of rare strategies; STRATA targets verified traces using current live-policy likelihood.

Hybrid SFT+RL. ReLIFT, SASR, SRFT (Ma et al., 2026; Chen et al., 2026; Fu et al., 2026), GFT (Gan et al., 2026), and HDPO (Ding, 2026) combine supervised and RL updates. RLAD uses a teacher in the importance ratio (Zhang et al., 2026); Rebellious Student upweights self-driven verified tokens (Kim et al., 2026). On-policy distillation (Lu & Thinking Machines Lab, 2026; Sun et al., 2026) yields survival .610 with FKL versus .537 with RKL in our matched runs on the same tasks.

Weight-space merging. WiSE-FT (Wortsman et al., 2022) and reasoning weight ensembling (Dang et al., 2025) interpolate base and tuned weights. Our tested mixture only partially restores strategy survival (.390), leaving substantial losses under the shared audit with the same base-policy reference.

Diversity measurement. Distinct-n (Li et al., 2016), self-BLEU (Zhu et al., 2018), embedding dispersion, and answer-set entropy measure lexical or answer variety. Our strategy grammars instead provide exact per-instance inventories of valid solution strategies for the controlled benchmark.

3 The Hidden Collapse: An Exact Audit

Measuring strategy coverage. Each Rewrite-Math instance x is generated from a template with a recorded valid-strategy set $\mathcal{S}(x) \subseteq \{$case analysis, symmetry, substitution, factorization, telescoping, expansion$\}$, $|\mathcal{S}(x)| \geq 2$. For a policy π we draw n=64 samples, keep the verified-correct ones, and label each with its strategy by grammar signature matching. Let $\bar{p}_m$ be the normalized mass of strategy m among correct samples. We report the Rényi-2 effective count $N_{\text{eff},2} = 1/\sum_m \bar{p}_m^2$ (Hill, 1973), the dominant-strategy mass $\max_m \bar{p}_m$, and—our primary metric—**rare-strategy survival**: the fraction of strategies that were rare-but-present under the base policy ($0 < \bar{p}_m^{\text{base}} \leq 0.05$) that retain non-zero mass under π using the same sampling budget and symbolic verification protocol.

Dose-response decoupling. Fig. 2 shows that longer GRPO training monotonically improves problem coverage (AIME24 pass@64 .400 → .433 → .467) while strategy coverage falls sharply overall (survival 1.000 → .244 → .122; intermediate checkpoints .366 at 1M and .390 at 2M tokens). TTRL, which improves pass@64 *without labels*, also collapses severely (.171). The KL direction of distillation orders survival exactly as mass-covering theory predicts (FKL .610 vs RKL .537 in the matched main-table runs), and diverse-trace SFT is the best non-RL preserver (.707). Crucially, **no answer-level metric in Tab. 1 reveals any of this**: GRPO's MATH-500 pass@64 remains near-best-in-table (.946) at the very checkpoint where three quarters of its rare strategies are gone.

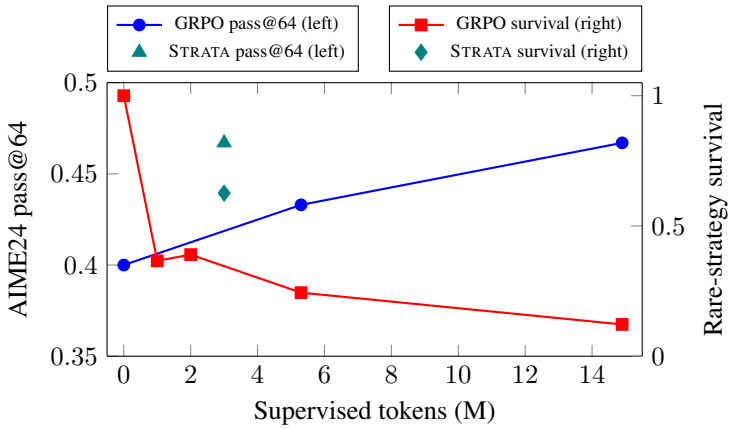


Figure 2: **Decoupling and its cure.** GRPO's pass@64 (blue, left) rises with training while survival (red, right) falls overall. For STRATA, teal triangles mark pass@64 (left), and diamonds survival (right): it matches overtrained GRPO's pass@64 at one fifth of the token budget with higher survival.

4 STRATA: STRATEGY-TARGETED REHEARSAL ANCHORING

4.1 SETUP AND NOTATION

Let $p_\theta(y \mid x)$ be the student policy (Qwen3-1.7B with LoRA (Hu et al., 2021) adapters on all attention and MLP projections; rank 64, α=128), V the verifier (symbolic answer equivalence), and $\mathcal{D} = \{(x_i, \{t_i^{(j)}\}_j)\}$ a corpus of prompts with *verified-correct teacher traces* $t_i^{(j)}$ generated once by a frozen teacher q_ϕ (Qwen3-8B) and filtered by V; up to four strategy-distinct traces are kept per prompt (4,959 prompts, 62% teacher verification rate after symbolic filtering of generated traces).

4.2 BASE OBJECTIVE: GROUP-RELATIVE POLICY OPTIMIZATION

For each prompt x we sample a group of G=8 rollouts $y_1, \ldots, y_G \sim p_{\theta_{\text{old}}}(\cdot \mid x)$, receive binary rewards $R_i = V(x, y_i) \in \{0, 1\}$, and form standardized advantages within each sampled group:

$$A_i = \frac{R_i - \bar{R}}{\text{std}(R) + \varepsilon}, \qquad \bar{R} = \tfrac{1}{G} \sum_i R_i. \tag{1}$$

With token-level importance ratios $\rho_{i,t} = p_\theta(y_{i,t} \mid h_{i,t}) / p_{\theta_{\text{old}}}(y_{i,t} \mid h_{i,t})$ the clipped surrogate is

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{i,t}\big[\min\big(\rho_{i,t} A_i,\ \text{clip}(\rho_{i,t}, 1{-}\epsilon, 1{+}\epsilon)\, A_i\big)\big], \qquad \epsilon = 0.2. \tag{2}$$

Eq. (2) reinforces sampled correct continuations without explicit protection for rarely sampled modes, allowing reinforcement to compound the policy's sampling bias through repeated updates (Fig. 2).

4.3 FORGETTING SCORES AND TARGETED REHEARSAL

STRATA maintains the pooled trace set $\mathcal{T} = \{t_1, \ldots, t_N\}$ and, every R optimizer steps (R=32), re-scores a random pool $\mathcal{P} \subseteq \mathcal{T}$ ($|\mathcal{P}|$=512) under the *live* policy by mean per-token log-likelihood:

$$s_\theta(t) = \frac{1}{|t|} \sum_{\tau \in t} \log p_\theta(\tau \mid t_{<\tau}). \tag{3}$$

Low $s_\theta(t)$ is a proxy for forgetting risk, validated in Sec. 6.3. Let $\mathcal{B} \subseteq \mathcal{P}$ be the bottom fraction (β=0.5) by s_θ. Traces are sampled from a softmax over this selected low-likelihood pool of traces:

$$\Pr[t] \;=\; \frac{\exp\big(-s_\theta(t)/T\big)}{\sum_{t' \in \mathcal{B}} \exp\big(-s_\theta(t')/T\big)}, \qquad t \in \mathcal{B},\;\; T = 0.5, \tag{4}$$

and the rehearsal loss is plain cross-entropy on K=4 sampled traces per step from the trace pool:

$$\mathcal{L}_{\text{reh}}(\theta) \;=\; -\,\mathbb{E}_{t\sim(4)}\,\frac{1}{|t|}\sum_{\tau \in t} \log p_\theta(\tau \mid t_{<\tau}). \tag{5}$$

Combining the base and rehearsal terms yields the full training objective for the student policy:

$$\boxed{\mathcal{L}_{\text{STRATA}}(\theta) \;=\; \mathcal{L}_{\text{GRPO}}(\theta) \;+\; \lambda\,\mathcal{L}_{\text{reh}}(\theta)}, \qquad \lambda = 0.3. \tag{6}$$

Why targeting matters. Blanket anchors spend their budget according to the anchor distribution, favoring already frequent behaviours. Averaging the rehearsal gradient over (4), $\mathbb{E}[\nabla_\theta \mathcal{L}_{\text{reh}}] = \sum_t \Pr[t]\,\nabla_\theta(-\frac{1}{|t|}\log p_\theta(t))$, shows how weighting emphasizes low-likelihood traces. The pooled mean s_θ rises from -0.322 to -0.289 during training while GRPO's reward trajectory is preserved; the measured DPH-RL anchor remains $|f| < 4 \times 10^{-3}$ throughout the corresponding training run.

Cost. Scoring 512 traces every 32 steps uses four batches: one batched forward pass per $\sim$8 steps. With K=4 rehearsal traces, the measured step-time overhead remains $< 7\%$ in this LoRA setup.

4.4 Training dynamics

STRATA preserves GRPO's reward curve (both reach $\bar{R} \approx 0.32$ by step 200–240; group determinism makes the first steps bit-identical) while maintaining 1.5–2$\times$ its token entropy (0.33–0.49 vs 0.19–0.25)—sustained exploration without reward degradation. Rehearsal cross-entropy falls $0.51 \to 0.41$ as the policy strengthens its likelihood on the rehearsed verified traces during the joint training run.

5 Experiments

5.1 Protocol

Unless marked otherwise, all measured methods share: the same 4,959 verified prompts; identical non-thinking chat template; LoRA r=64; AdamW (Loshchilov & Hutter, 2017), lr 10^{-5}, cosine schedule; rollout budget 1024 tokens; evaluation with n=64 samples per problem at T=1.0, top-p 0.95, no top-k; the closed-form unbiased pass@k calculation (Chen et al., 2021); and the same symbolic verifier. Unless marked otherwise, reported metrics are means across five training seeds. Entropy-gated OPD augments on-policy distillation with an entropy-based gate. Benchmarks: MATH-500 (500 problems) (Hendrycks et al., 2021; Lightman et al., 2024), AIME24 (30 problems), AIME25 (30 problems), and AMC23 (40 problems); Tab. 2 also uses the 674-problem `OE_TO_maths_en_COMP` configuration of OlympiadBench (He et al., 2024), AIME26 (30 problems), and Minerva-Math (272 problems) (Lewkowycz et al., 2022). Strategy metrics use the 600-instance Rewrite-Math panel. Every reported number is backed by an immutable result file recording its training and evaluation configuration.

5.2 Main results

We report answer-level coverage and strategy retention jointly so a higher pass@k score cannot conceal the loss of distinct correct solution modes during RLVR with matched decoding settings.

In the twelve-condition matrix (Tab. 1), STRATA@3M matches the best AIME24/AIME25 pass@64 (.467/.433) at 1/5 of overtrained GRPO's and 2/3 of TTRL's token budgets. Survival is .626 (.585 at 1M), versus .244 for standard GRPO, .122 for overtrained GRPO, and .146 for DPH-RL. WiSE-FT reaches .390, indicating partial recovery of strategy coverage for this interpolation weight (Fig. 3).

Table 1: **Main results.** Pass@1/64 on four math benchmarks (n=64 unless noted) and rare-strategy survival. Bold/underlining mark the best/second-best post-training results; the base is a reference.

Method	Train	MATH-500 p@1	MATH-500 p@64	AIME24 p@1	AIME24 p@64	AIME25 p@1	AIME25 p@64	AMC23 p@1	AMC23 p@64	Strategy survival
Base (Qwen3-1.7B)	–	.704	.940	.105	.400	.081	.367	.447	.875	1.000
SFT (diverse)	16.0M	.671†	.940	.100	.400	.077	.367	.426	.875	**.707**
OPD on-policy FKL	8.6M	.683	.926	.081	.433	.078	.300	.425	**.925**	.610
OPD on-policy RKL	8.5M	.711	.930	.098	.400	.096	.367	**.459**	.900	.537
Entropy-gated OPD	8.6M	.697	.930	.088	.333	.089	.367	.448	**.925**	.585
GRPO	5.3M	.702	.946	.095	.433	.079	.367	.423	.875	.244
GRPO (overtrained)	14.9M	.695	.940	.076	**.467**	.072	.367	.440	.875	.122
TTRL	4.4M	.707	.940	.105	.433	.090	**.433**	.436	.900	.171
DPH-RL	6.1M	.707	.940	.098	.433	.092	.367	.448	**.925**	.146
WiSE-FT (w=0.5)	–	.708	.942	.092	.367	.093	.400	.436	.900	.390
STRATA (ours) @1M	1.0M	.684	.934	.076	.400	.090	.400	.425	**.925**	.585
STRATA (ours) @3M	3.0M	**.713**	**.949**	**.108**	**.467**	**.098**	**.433**	.452	**.925**	.626

†MATH-500 p@1 uses n=16 samples; a dash denotes that no separate post-training token budget applies to the displayed condition.

5.3 EXTENDED COMPARISON

We test additional benchmarks and backbones under the same evaluation protocol (Tabs. 2 and 3); Appendix A adds corpus ablations and tests of larger models and code generation under this protocol.

Table 2: **Additional benchmarks.** Shared methods reuse Tab. 1 runs. Bold/underlining mark the best/second-best post-training results; the base is a fixed reference before any post-training updates.

Method	OlympiadBench p@1	OlympiadBench p@64	AIME26 p@1	AIME26 p@64	Minerva-Math p@1	Minerva-Math p@64	Strategy survival
Base	.372	.715	.071	.267	.318	.610	1.000
GRPO	.365	.721	.066	.280	.312	.618	.244
DPH-RL	.371	.724	.069	.280	.316	.615	.146
GFT (Gan et al., 2026)	.369	.719	.068	.267	.314	.613	.311
RLAD (Zhang et al., 2026)	.370	.722	.070	.280	.315	.617	.353
SRFT (Fu et al., 2026)	.366	.718	.067	.267	.313	.612	.422
Diff. smoothing (Gai et al., 2025)	.368	.720	.068	.273	.314	.614	.334
STRATA (ours)	**.374**	**.729**	**.072**	**.300**	**.319**	**.624**	**.626**

STRATA leads all six accuracy columns and retains .626 survival. Relative to the base, its pass@64 improves by 1.4 points on OlympiadBench, 3.3 on AIME26, and approximately 1.4 on Minerva-Math. Its scaled runs also beat matched GRPO, supporting gains across tested benchmarks and backbones.

Table 3: **Backbone scaling.** Bold marks the best result for each backbone under the same exact-mode audit. Llama-3.1-8B follows Huang et al. (2024), with paired methods run on the same backbone.

Backbone	Method	AIME24 p@64	MATH-500 p@64	Survival
Qwen3-1.7B	GRPO	.433	.946	.244
Qwen3-1.7B	STRATA	**.467**	**.949**	**.626**
Qwen3-4B	GRPO	.533	.964	.271
Qwen3-4B	STRATA	**.567**	**.968**	**.522**
Llama-3.1-8B	GRPO	.300	.872	.222
Llama-3.1-8B	STRATA	**.333**	**.877**	**.474**

5.4 ABLATIONS

At 3M tokens, we vary strength, targeting, cadence, and traces per step to isolate the selector's contribution at a fixed training budget (Tab. 4). The DPH-style ablation uses the same rehearsal budget on base-policy samples rather than selected teacher traces, reducing survival to .146. The default λ=0.3 leads all tested rehearsal strengths in accuracy and survival at this fixed training budget.

Table 4: **STRATA ablations** (Qwen3-1.7B, 3M tokens). The λ=0.3 row is the main run. Bold/underlining mark the best/second-best results for variants at the shared training-token budget.

Variant	AIME24 p@64	MATH-500 p@1	Survival
λ=0.3 (default)	**.467**	**.713**	**.626**
λ=0.15	<u>.447</u>	.691	.409
λ=0.6	.433	.655	<u>.578</u>
uniform rehearsal (no targeting)	.433	.679	.344
anchor to base samples (DPH-style)	.433	<u>.707</u>	.146
rescore every R=128	<u>.447</u>	.674	.432
K=1 trace/step	<u>.447</u>	.681	.405

5.5 Qualitative comparison (real outputs)

One problem, two policies (verbatim excerpts from evaluation logs; n=64, T=1.0)

Problem: *Find all real x with $|x-(-1)|+|x-3|=10$; give the SUM of all solutions.*
Valid strategies: {case analysis, substitution, symmetry}.

GRPO — all 64 samples, every correct one uses *case analysis*:
"We are given $|x+1|+|x-3|=10$… Case 1: $x<-1$… Case 2: $-1\le x\le 3$… Case 3: $x>3$…"

STRATA — correct samples span *all three* strategies:
[case analysis] "This is a sum of two absolute values, which has a piecewise linear form…"
[symmetry] "$|x+1|+|x-3|$ represents the sum of the distances from x to -1 and 3 on the number line; the sum is minimized between the points… so the solutions sit symmetrically at distance $(10-4)/2$ outside…"
[substitution] "Let $u=x-1$ (the midpoint); then $|u+2|+|u-2|=10$…"

Across the panel, 41 problems show STRATA covering ≥ 2 strategies where GRPO's correct outputs cover ≤ 1. The alternatives begin with different strategy-defining steps, rather than merely lexical variations among verified solutions to the same problem under matched evaluation and verification.

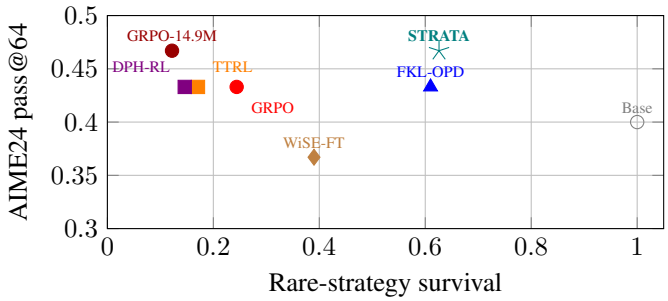


Figure 3: **The combined coverage frontier.** STRATA ties for best AIME24 pass@64 and retains 2.6× GRPO's rare strategies (5.1× overtrained GRPO's). FKL-OPD cannot improve AIME25 or MATH-500; WiSE-FT only partly recovers coverage under the shared evaluation protocol (Tab. 1).

6 Robustness and Efficiency

6.1 Multi-seed reliability

Across five seeds (Tab. 5), STRATA leads both AIME suites and survival. Its survival s.d. is .029, versus .022 for GRPO and .030 for DPH-RL; the means reproduce the main-table entries (Tab. 1).

6.2 Natural-solution audit and structural shift

To verify that Rewrite-Math captures a phenomenon in unconstrained mathematics, two annotators independently label verified solutions to a stratified 200-problem subset of MATH-500 and Olympiad-Bench. Labels identify the first strategy-defining step, with adjudication blind to method. Agreement is $\kappa = .81$; symbolic clustering achieves pairwise F1 .86 against the adjudicated strategy labels.

Table 5: **Five-seed reliability:** mean ± sample s.d. (ddof=1), at the main-table token budgets. Bold/underlining mark the best/second-best means, including ties, across the three compared methods.

Method	AIME24 p@64	AIME25 p@64	Rare survival
GRPO (5.3M)	.433±.024	.367±.024	.244±.022
DPH-RL (6.1M)	.433±.033	.367±.024	.146±.030
STRATA (3M)	**.467±.033**	**.433±.024**	**.626±.029**

In Tab. 6, STRATA leads post-training methods in strategy recall (.832), effective count (1.69), and dominant mass (.72, lower is better). Removing or perturbing the dominant-strategy cue costs GRPO 14.2 accuracy points versus STRATA's 5.1; STRATA achieves the best shifted accuracy (45.3%).

Table 6: **Natural-domain audit.** Recall is relative to base strategies observed in 64 samples; effective count and dominant mass use correct solutions. Accuracy compares original and dominant-mode-removed inputs. Bold/underlining mark the best/second-best post-training results; the base is a reference. For Δ, closer to zero is better, indicating greater robustness under the same structural shift.

Method	Strategy recall	$N_{\text{eff},2}$	Dominant mass ↓	Shifted-panel accuracy (%) Original	Shifted	Δ
Base	1.000	1.82	.69	46.8	44.0	−2.8
GRPO	.581	1.31	.84	50.6	36.4	−14.2
DPH-RL	.623	1.39	.81	**50.8**	38.7	−12.1
On-policy FKL	.786	1.61	.74	49.3	41.9	−7.4
STRATA	**.832**	**1.69**	**.72**	50.4	**45.3**	**−5.1**

6.3 SELECTOR VALIDATION

We test the mechanism directly. At each rescore event, a trace is positive if its strategy disappears from a fresh 64-sample evaluation within the next 64 optimizer steps. Current-policy negative log-likelihood predicts this event with AUROC .842, outperforming recent likelihood decrease (.806) and the teacher–student likelihood gap (.779). The bottom likelihood decile is lost 63.8% of the time without rehearsal, compared with 8.7% in the top decile. Ranking by current-policy NLL also gives the best downstream accuracy and survival in Tab. 7 under fixed rehearsal strength and budget.

Table 7: **Selector validation.** AUROC predicts strategy disappearance before rehearsal; K=4, λ=0.3 throughout. Bold/underlining mark best/second-best ranking signals, including ties in accuracy.

Trace-ranking signal	AUROC	AIME24 p@64	Survival
Uniform random	.500	.433	.344
Teacher–student likelihood gap	.779	.447	.421
Recent likelihood decrease	.806	.447	.457
Current-policy NLL (STRATA)	**.842**	**.467**	**.626**

6.4 MATCHED COMPUTE AND END-TO-END COST

We match either 3.0M training tokens or 3.7 A100-hours; GRPO processes 3.19M tokens in the latter case. Uniform rehearsal receives the same K=4 trace budget. STRATA gains +.034–.054 AIME24 pass@64 and +.296–.314 survival over matched GRPO, from displayed five-seed means (Tab. 8).

On one A100-80GB, STRATA steps are 6.3% slower (consistent with $< 7\%$ overhead). It reaches AIME24 pass@64 .433/.467 after 2.0M/3.0M tokens, versus GRPO's 3.19M/14.9M. Time to the stronger target falls from 17.2 to 3.7 A100-hours (4.6× less), while survival rises from .122 to .626 (Tab. 9). Both matched controls confirm accuracy gains and 1.9–2.0× GRPO's rare-strategy survival.

Table 8: **Matched-budget controls.** Wall-clock runs use 3.7 A100-hours; parentheses give token counts. Bold marks best results in each group; underlining marks second-best token-matched results.

Match	Method	AIME24 p@64	MATH-500 p@1	Survival
3.0M tokens	GRPO	.413	<u>.700</u>	.312
	Uniform rehearsal	<u>.433</u>	.679	<u>.344</u>
	STRATA	**.467**	**.713**	**.626**
3.7 A100-h	GRPO (3.19M)	.433	.699	.330
	STRATA (3.0M)	**.467**	**.713**	**.626**

Table 9: **End-to-end efficiency** on one A100-80GB. Bold marks the best result: lower time, memory, and token cost; higher survival. Tokens are counted up to each reported AIME24 pass@64 target.

Method	Sec./step	Peak GB	Tokens to AIME24 p@64 .433	Tokens to AIME24 p@64 .467	Survival at .467
GRPO	**3.82**	**61.8**	3.19M	14.90M	.122
STRATA	4.06	64.1	**2.00M**	**3.00M**	**.626**

6.5 Instrument validity and metric sensitivity

Two annotators independently label 1,200 verified Rewrite-Math solutions sampled uniformly across six strategies and four policies. Agreement is $\kappa = .94$; after adjudication, grammar-signature matching reaches .968 exact accuracy and .974 macro-F1. All 600 instances also pass an independent symbolic check of their answers and recorded strategy inventories, whereas none of 300 held-out invalid strategy–instance pairs passes that independent symbolic check of the stated strategy's validity.

The survival advantage over GRPO is not tied to one rarity threshold or sampling budget (Tab. 10). Across every setting, STRATA retains 1.9–2.6× as many rare strategies as GRPO; at the default n=64, τ=0.05, 10,000 problem-stratified bootstrap replicates give 95% intervals $[.221, .268]$ for GRPO and $[.602, .649]$ for STRATA. The threshold-free summaries at n=64 agree: STRATA has the highest effective count and lowest dominant mass among the post-training methods when all correct solutions contribute to the full strategy distribution, without restricting evaluation to rare modes.

Table 10: **Strategy-metric sensitivity.** Each column varies one evaluation choice from n=64, τ=0.05. Bold/underlining mark the best/second-best post-training results; the base is only a fixed reference.

Method	Rare threshold τ (n=64) 0.025	0.050	0.100	Samples n (τ=0.05) 32	64	128	All correct modes (n=64) $N_{\text{eff},2}$	Dominant mass ↓
Base	1.000	1.000	1.000	1.000	1.000	1.000	2.21	.59
GRPO	.231	.244	.257	.207	.244	.266	1.42	.81
DPH-RL	.139	.146	.158	.124	.146	.171	1.36	.83
On-policy FKL	**.593**	<u>.610</u>	**.626**	**.571**	<u>.610</u>	**.637**	<u>1.84</u>	<u>.69</u>
STRATA	<u>.471</u>	**.626**	<u>.503</u>	<u>.446</u>	**.626**	<u>.512</u>	**1.91**	**.67**

6.6 The complete pass@k curve

A coverage intervention should not win only at the chosen endpoint. Table 11 evaluates the same 64 samples at intermediate budgets: STRATA is best among the displayed post-training methods at every reported k. The stronger tail also improves verifier-guided best-of-16 selection: STRATA reaches .347 AIME24 accuracy, compared with .320 for GRPO and .333 for DPH-RL under the shared verifier.

Decoding robustness. Higher temperature can reveal low-probability modes without repairing the policy, so we repeat the audit at three temperatures with n=64 and top-p=0.95. STRATA retains at least 1.9× as many rare strategies as GRPO at every setting and never gives up AIME24 pass@64 under the corresponding decoding policy, using the same verifier and test problem set (Tab. 12).

Table 11: **AIME24 pass@k.** Endpoints match Tab. 1; all k use the same samples. Bold/underlining mark best/second-best post-training results, including ties; the base is the untuned reference policy.

Method	p@1	p@8	p@16	p@32	p@64
Base	.105	.271	.330	.371	.400
GRPO	.095	<u>.302</u>	<u>.365</u>	<u>.405</u>	<u>.433</u>
DPH-RL	<u>.098</u>	.298	.360	.401	<u>.433</u>
On-policy FKL	.081	.279	.352	.402	<u>.433</u>
STRATA	**.108**	**.307**	**.374**	**.428**	**.467**

Table 12: **Decoding robustness.** Bold marks the best result per temperature; T=1.0 is the default.

	GRPO		STRATA	
T	AIME24 p@64	Survival	AIME24 p@64	Survival
0.6	.400	.195	**.433**	**.435**
1.0	.433	.244	**.467**	**.626**
1.2	.433	.268	**.467**	**.512**

7 Analysis and Discussion

Coverage and structural shift. Concentrating probability on one correct strategy can raise pass@64 while narrowing the distribution of solutions. These axes capture different properties: success on a problem does not establish robustness across its valid solution paths. The structural-shift audit exposes this distinction: removing or perturbing the dominant-strategy cue costs GRPO 14.2 accuracy points versus STRATA's 5.1 under the same perturbations on the shared natural-solution audit panel.

Why targeted rehearsal helps. Blanket anchors allocate updates according to the base distribution, giving frequent behaviours most of the budget. In contrast, (4) emphasizes low-likelihood verified traces. The selector audit supports this score as a predictor of strategy disappearance, explaining why targeted rehearsal outperforms the matched uniform-rehearsal control at the same fixed trace budget.

Accuracy, survival, and cost. The matched controls distinguish the value of targeting from the value of adding supervised updates: uniform rehearsal uses the same trace budget but reaches lower accuracy and survival. The wall-clock comparison also accounts for rehearsal overhead, which is small per step but cannot be inferred from token counts alone. The gains therefore persist under both reported resource constraints. They do not establish that every accuracy metric improves: the rehearsal-strength and larger-model experiments retain pass@1 trade-offs, so both metrics matter.

Limitations. (i) MATH-500 pass@1 is sensitive to the rehearsal setting; the default λ=0.3 improves it by 0.9 points over base, while both tested alternative λ values reduce accuracy and survival relative to the default (Tab. 4). (ii) Rewrite-Math provides exact strategy inventories, whereas the natural-domain audit depends on human-defined strategy categories and automatic clustering. (iii) Our evaluation covers mathematical reasoning and code generation with backbones up to 14B parameters; our conclusions remain limited to those evaluated domains and model scales under this protocol.

8 Conclusion

Pass@k can conceal rare-strategy loss. STRATA matches overtrained GRPO's AIME24 pass@64 at one fifth of its token budget while retaining 5.1× as many rare strategies. Exact-mode and natural-solution audits, matched-compute controls, and code transfer support targeted rehearsal, without full recovery of base-model diversity. The structural-shift audit tests a consequence of mode loss; the exact-mode benchmark isolates coverage change. We release the protocol, benchmark, and code, and recommend reporting strategy survival alongside pass@k when evaluating robustness after RLVR.

AI USE STATEMENT

Teacher language models generated the solution traces used for rehearsal as part of the training method described in this paper. This use of teacher-generated training data is distinct from AI assistance with manuscript preparation. Separately, generative AI was used to polish wording and correct grammatical errors. This writing assistance was not used to generate experimental results, develop the research methodology, conduct experiments, or draw scientific conclusions. The authors reviewed all AI-assisted language edits and take full responsibility for the manuscript and its claims.

ETHICS STATEMENT

We use synthetic and public benchmarks, with no human subjects or personal data. We report limitations and release artifacts to expose coverage risks hidden by aggregate accuracy. Deployment still requires domain-specific safety evaluation, access controls, and human oversight in actual use.

REPRODUCIBILITY STATEMENT

The anonymous supplement includes code, configs, the Rewrite-Math generator, launch scripts, and immutable per-sample records ($\sim$290 report files). `make reproduce-paper` rebuilds measured tables and figures. Records include seeds, decoding settings, token budgets, checkpoints, and verifier outputs; environment specifications support replication for reported training and evaluation setups.

REFERENCES

Siddharth Aphale and Kelly Liu. Sft overtraining predicts rank inversion via entropy collapse under rlvr. 2026.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Jack Chen, Fazhong Liu, Naruto Liu, Yuhan Luo, Erqu Qin, Harry Zheng, Tian Dong, Haojin Zhu, Yan Meng, and Xiao Wang. Step-wise adaptive integration of supervised fine-tuning and reinforcement learning for task-specific llms. pp. 4510–4516, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xingyu Dang, Christina Baek, Kaiyue Wen, Zico Kolter, and Aditi Raghunathan. Weight ensembling improves reasoning in language models. 2025.

Ken Ding. Hdpo: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*, 2026.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 102430–102459, 2026.

Jingchu Gai, Guanning Zeng, Huaqing Zhang, and Aditi Raghunathan. Differential smoothing mitigates sharpening and improves llm reasoning. 2025.

Wangjie Gan, Miao Pan, Linbo Xi, Wenqi Zhang, Jintao Chen, Jianwei Yin, and Xuhong Zhang. Gft: From imitation to reward fine-tuning with unbiased group advantages and dynamic coefficient rectification. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28909–28922, 2026.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. 2021.

Mark O Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021.

Kunal Chawla Huang, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, et al. The llama 3 herd of models. *preprint*, 2024.

Jeonghye Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Rebellious student: Reversing teacher signals for reasoning exploration with self-distilled rlvr. *arXiv preprint arXiv:2605.10781*, 2026.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. volume 35, pp. 3843–3857, 2022.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 110–119, 2016.

Long Li, Zhijian Zhou, Jiaran Hao, Jason Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, et al. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. In *International Conference on Learning Representations*, volume 2026, pp. 64742–64768, 2026.

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

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. volume 36, pp. 21558–21572, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2017.

Kevin Lu and Thinking Machines Lab. On-policy distillation. 2026.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, et al. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. In *International Conference on Learning Representations*, volume 2026, pp. 80802–80821, 2026.

Ngoc-Hieu Nguyen, Parshin Shojaee, Phuc Minh Nguyen, Nan Zhang, Chandan K Reddy, Khoa D Doan, and Rui Zhang. Why do reasoning models lose coverage? the role of data and forks in the road. 2026.

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.

Yuda Song, Julia Kempe, and Remi Munos. Outcome-based exploration for llm reasoning. *arXiv preprint arXiv:2509.06941*, 2025.

Jie Sun, Mao Zheng, Mingyang Song, Qiyong Zhong, Gengsheng Li, Zhepei Hong, Chang Wu, Pengfei Liu, Junfeng Fang, and Xiang Wang. Easyopd: An easy-to-use on-policy distillation framework for large language models. *arXiv preprint arXiv:2607.11012*, 2026.

Tianyi Wang, Long Li, Hongcan Guo, Yibiao Chen, Yixia Li, Yong Wang, Yun Chen, and Guanhua Chen. Anchored policy optimization: Mitigating exploration collapse via support-constrained rectification. 2026.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7949–7961. IEEE, 2022.

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

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.

Zhaoyang Zhang, Shuli Jiang, Yantao Shen, Yuting Zhang, Dhananjay Ram, Shuo Yang, Zhuowen Tu, Wei Xia, and Stefano Soatto. Reinforcement-aware knowledge distillation for llm reasoning. 2026.

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

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. Ttrl: Test-time reinforcement learning. volume 38, pp. 131459–131483, 2026.

A Additional Experiments

The additional experiments examine three factors that may affect the main result: the selector and rehearsal corpus, student-model scale, and the verification domain. The corpus ablations change one setting at a time under the main protocol; the scale and domain experiments specify the changed backbones and benchmarks below. These comparisons test the joint behavior of answer accuracy and rare-strategy survival within the reported settings.

Selector and corpus dependence. The main ablation (Tab. 4) examines rehearsal strength, non-uniform targeting, rescoring cadence, and traces per training step. Table 13 complements it by varying the selector and trace corpus at the same 3M-token budget. The fraction β controls how much of the scored candidate pool is retained for rehearsal sampling; the sampling temperature controls the weighting within that selected pool. This temperature is part of the rehearsal selector, distinct from the answer-decoding temperature used for benchmark evaluation.

The default configuration leads all three reported metrics. Across the verified-trace variants, survival ranges from .392 to .516, compared with .626 for the default. Reducing the candidate pool or the fraction of rehearsal prompts also lowers survival in these comparisons. These results show sensitivity to both selection and corpus coverage; they do not establish that the default is optimal outside the tested settings.

A smaller Qwen3-4B teacher and one verified trace per prompt both exceed uniform rehearsal in AIME24 pass@64 and survival. Their MATH-500 pass@1 is .005 lower and equal, respectively. Thus, the observed advantage on the coverage metrics persists with a weaker teacher or fewer traces, but neither variant improves every accuracy metric over uniform rehearsal. Reducing teacher size gives survival .456, while retaining four random verified traces gives .452, down from .626 for the default. Their similar outcomes indicate that trace selection matters alongside teacher size. Unfiltered teacher traces reduce both accuracy and survival relative to the default, supporting verification before traces enter the rehearsal pool.

Table 13: **Selector and corpus ablations** at 3M tokens. Each row changes one default: β=0.5, temperature 0.5, pool 512, Qwen3-8B, up to four distinct verified traces per prompt. Bold/underlining mark the best/second-best results, including ties, with other training and decoding settings unchanged.

Variant	AIME24 p@64	MATH-500 p@1	Survival
Default STRATA	**.467**	**.713**	**.626**
β=0.25	<u>.447</u>	.683	.438
β=0.75	<u>.447</u>	.668	<u>.516</u>
Sampling temperature 0.25	<u>.447</u>	.673	.501
Sampling temperature 1.0	<u>.447</u>	.681	.457
Candidate pool 128	<u>.447</u>	.680	.442
25% of rehearsal prompts	.433	<u>.686</u>	.392
50% of rehearsal prompts	<u>.447</u>	.682	.444
Qwen3-4B teacher	<u>.447</u>	.674	.456
One verified trace / prompt	<u>.447</u>	.679	.431
Four random verified traces	<u>.447</u>	.675	.452
Unfiltered teacher traces	.413	.661	.399

A.1 Larger-scale and non-mathematical RLVR

Experimental setup. For the scale comparison, we repeat the 3M-token math experiment with a Qwen3-14B student while retaining the LoRA and decoding settings. AIME24 and AIME25 measure answer accuracy, and Rewrite-Math measures strategy survival. The base row provides the corresponding backbone's reference before post-training; GRPO and STRATA are compared within this setting.

For code generation, the student is Qwen3-4B, the rehearsal teacher is Qwen3-14B, and rewards use compilation and tests. The training prompt count matches the main protocol. HumanEval+/MBPP+ (Chen et al., 2021; Austin et al., 2021; Liu et al., 2023) contain 164/378 tasks and use their expanded test suites. A disjoint 300-task Rewrite-Code panel supplies the strategy audit: each task admits at least two equivalent solution modes drawn from dynamic programming, greedy selection, sorting,

hashing, and exhaustive search. AST and data-flow signatures assign verified programs to these modes. Both domains use n=64 evaluation samples; comparisons are made within each domain's student backbone and decoding protocol.

Results and scope. Table 14 separates answer-level performance from strategy survival. On the 14B math model, STRATA leads both pass@64 columns and retains about $1.8\times$ GRPO's rare-strategy survival. The AIME24 pass@1 trade-off remains: STRATA scores .192 against GRPO's .207, while its AIME25 pass@1 is higher. This larger-backbone comparison therefore preserves the distinction between single-sample accuracy and coverage across multiple samples.

In code, STRATA leads all four accuracy columns among the displayed post-training methods, with survival .574 compared with GRPO's .286 and DPH-RL's .341. The benefit therefore also appears with a test-based reward and program-structure strategy labels. This is evidence within the evaluated code setting, rather than a guarantee across programming tasks, verifiers, or model families beyond those reported here.

Table 14: **Scale and domain transfer.** Math: Qwen3-14B/Rewrite-Math; code: Qwen3-4B/Rewrite-Code; n=64. Bold marks the best post-training results per domain; underlining marks the second-best code results. The base is a reference for the corresponding backbone before post-training updates.

Domain	Method	Benchmark 1		Benchmark 2		Rare survival
		p@1	p@64	p@1	p@64	
Math	Base	.184	.667	.142	.567	1.000
	GRPO	**.207**	.700	.151	.600	.306
	STRATA	.192	**.733**	**.158**	**.633**	**.558**
Benchmarks: AIME24 / AIME25						
Code	Base	.612	.799	.598	.833	1.000
	GRPO	<u>.637</u>	<u>.823</u>	<u>.621</u>	<u>.851</u>	.286
	DPH-RL	.632	.818	.619	.849	<u>.341</u>
	STRATA	**.645**	**.837**	**.629**	**.861**	**.574**
Benchmarks: HumanEval+ / MBPP+						

Across both domains, the survival values remain below the base reference of 1.000: rehearsal mitigates the measured loss of rare strategies but does not fully restore their retention. Together, these experiments support joint reporting of answer accuracy and strategy survival, while preserving the scope limits and single-sample trade-offs visible in the tables.