WHEN RL SUPPRESSES ITS OWN VOCABULARY: RECOVERING REASONING DIVERSITY IN PUZZLE-TO-MATH TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning using verifiable rewards (RLVR) improves LLM reasoning, but the conditions under which it transfers across domains, and why it does so, remain under-explored. We study cross-domain transfer in two model families whose SFT and RL post-training stages use only constraint-satisfaction puzzles, with no mathematics problems in the post-training data, using a reasoning primitive-level analysis framework: a 9-class span classifier that segments chain-of-thought reasoning into primitives and tracks how they are sequenced across recipe stages and domains. We find that puzzle SFT induces a reasoning-primitive vocabulary, and that vanilla GSPO then commits the model to longer uninterrupted chains of reasoning, yielding `pass@32` gains on OlymMATH-Hard in both families. But the same training also suppresses backtracking across families and in both domains. To address this, we propose a novelty bonus that rewards diverse correct rollouts using perplexity under the reference model as a signal, designed to preserve the backtracking present in the SFT checkpoint. It brings backtracking back toward its SFT level, and in a controlled A/B test against vanilla GSPO (matched data, algorithm, training budget) it adds a further `pass@32` gain on hard mathematics. The end-to-end recipe raises hard-math `pass@32` by 20 points on OlymMATH-Hard for OLMo3-7B and 7 points on HMMT for Qwen3-4B without adding any mathematics problems in the post-training stages studied here.

1 INTRODUCTION

Reinforcement learning has driven much of the recent progress in large language model (LLM) reasoning on math and code, but how much of this improvement reflects general problem-solving capability versus skills specific to the training distribution is unclear. We probe this in a setting designed to dissociate the two: models from two families are trained with supervised fine-tuning (SFT) and reinforcement learning (RL) on constraint-satisfaction puzzles only, with no mathematics problems anywhere in post-training, and are evaluated on hard mathematics benchmarks where current models remain far from saturation. Puzzles and mathematics share problem-solving operations such as search, planning, verification, and backtracking (Pólya, 1945; Simon & Newell, 1971; Schoenfeld, 2014), while differing substantially in problem framing and notation. Whatever RL contributes to math performance in this setup cannot come from post-training exposure to mathematics. If transfer occurs, it must reflect changes to the abstract operations the two domains share. This setup lets us ask: does RL on puzzles improve hard-mathematics performance, and if so, what reasoning-level changes accompany the gain?

Transfer to math occurs in both families: after SFT and RL on puzzles, `pass@32` on OlymMATH-Hard rises by 20 percentage points for OLMo3-7B (from 16.0% to 36.0%), and by 7 points on the HMMT rounds released after its base model for Qwen3-4B (from 66.7% to 73.5%), without any mathematics in post-training. To see what reasoning-level changes accompany this gain, we develop a primitive-level analysis framework. A span classifier segments chain-of-thought reasoning into 9 primitives (HYPOTHESIZE, COMPUTE, CHECK, BACKTRACK, PLAN, ENUMERATE, SETUP, SUMMARIZE, OTHER), and we track how these primitives are sequenced in a trace, in particular how long the model pursues one line of reasoning before it backtracks. Because the analysis rests entirely on the primitive classifier, we anchor it against blind human annotation (Appendix N). The

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.

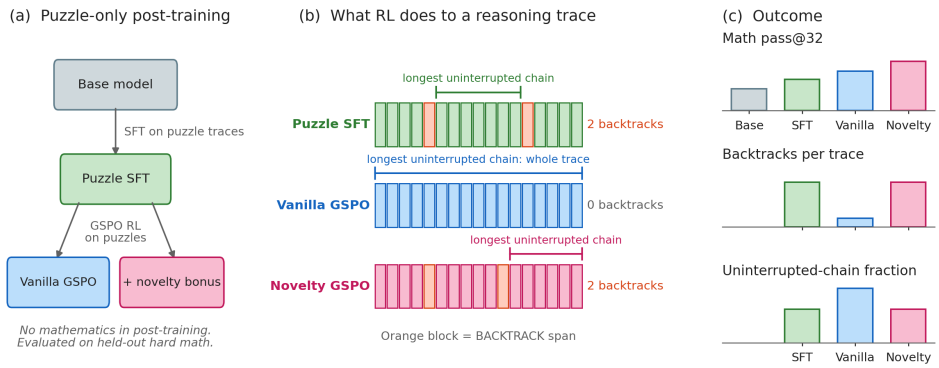


Figure 1: Overview. **(a)** The puzzle-only post-training pipeline. **(b)** Illustrative reasoning trace split into spans, with BACKTRACK spans in orange and the longest uninterrupted chain bracketed. **(c)** Illustrative outcome across the recipe. Vanilla GSPO lengthens the chain and suppresses backtracking, and the novelty bonus restores backtracking while raising `pass@32`.

framework reveals a chain–recovery tradeoff (Figure 1b). RL lengthens the *uninterrupted chains*: the longest stretch of a trace in which the model never backtracks grows, on hard mathematics to nearly the whole trace. It also suppresses *recovery*: BACKTRACK falls in both families and both domains, by 80% on OlymMATH-Hard for OLMo3.

This chain–recovery tradeoff motivates a modification to the RL reward: a perplexity-based novelty bonus that rewards rollouts unlikely under a frozen reference model, designed to preserve the backtracking present in the SFT checkpoint. The bonus brings backtracking back toward its SFT level, and in a controlled A/B test against vanilla GSPO with matched data, algorithm, and training budget, it adds a further 7pp `pass@32` on OlymMATH-Hard for OLMo3 and 5pp on HMMT for Qwen3. On the held-out mathematics evaluations of both families, the restored backtracking and the accuracy gain appear together.

Our conclusions, that RL on puzzles transfers, that it does so via primitive-level changes, and that a novelty bonus amplifies the effect, rely on a setup that cleanly isolates RL's contribution. Recent work on puzzle-to-math transfer has not had this property. **Enigmata** (Chen et al., 2025a) trains on a 36-puzzle suite at scale, but includes math problems in both SFT and RL, making it difficult to isolate whether transfer performance originates from in-domain math exposure. **Logic-RL** (Xie et al., 2025) and **LogicPuzzleRL** (Wong et al., 2025) train on puzzles only but evaluate primarily at `pass@1` on benchmarks no harder than AIME24. At `pass@1`, it is impossible to distinguish whether RL expands the set of problems the model can solve at all (capability ceiling expansion) or merely sharpens a solution path it could already reach under sufficient sampling. Whether puzzle-only post-training can raise the hard-math capability ceiling therefore remains undemonstrated in the prior literature.

Contributions. (1) We show that puzzle-only post-training, SFT followed by RL on constraint-satisfaction puzzles, transfers to hard mathematics in two model families, with no mathematics in post-training, (2) We introduce a human-validated primitive-level analysis of chain-of-thought reasoning and use it to identify a chain–recovery tradeoff under vanilla RL: longer uninterrupted chains, suppressed backtracking, (3) We propose a perplexity-based novelty bonus that brings backtracking capability back and adds a further `pass@32` gain over a matched vanilla baseline in both families.

2 Related Work

RL for LLM reasoning and synthetic puzzle tasks. Reinforcement learning from verifiable rewards (RLVR), popularised by DeepSeek-R1 (Guo et al., 2025) and developed through objectives such as GRPO (Shao et al., 2024), GSPO (Zheng et al., 2025), and DAPO (Yu et al., 2025), has become a standard recipe for post-training reasoning models. Synthetic puzzles with deterministic verifiers are a natural testbed for this recipe, and recent works such as Enigmata (Chen et al.,

2025a), Logic-RL (Xie et al., 2025), and LogicPuzzleRL (Wong et al., 2025) report transfer from puzzle training to mathematical reasoning. However, existing evidence either mixes math data into the training recipe or evaluates puzzle-only training primarily at `pass@1` on less difficult math benchmarks, leaving open whether non-math RL can expand the mathematical reasoning ceiling, measured by `pass@k` at large k. This question connects to a broader debate about RL post-training transfer: some work reports broad cross-domain gains from math RL (Cheng et al., 2025b), while other studies find that single-domain RL mainly sharpens in-domain behaviour (Huan et al., 2025; Hu et al., 2026a; Yue et al., 2025). We study this in a clean isolation: our pipeline uses no math data in SFT or RL and evaluates `pass@k` on hard out-of-domain math. Closest in spirit, Yuan et al. (2025) show that RL can compose previously learned atomic skills into combinations that transfer across tasks. We also find that RL works on primitives the SFT stage introduced, but our question is the opposite one: which of those primitives vanilla RL suppresses, and whether a novelty bonus restores them.

Exploration and diversity in RL fine-tuning. Recent work has argued that RLVR can improve `pass@1` while reducing `pass@k` diversity, motivating method families that preserve exploration during RL fine-tuning (Yue et al., 2025; Cui et al., 2025; Yu, 2025; Chen et al., 2025b). One family modifies token-level signals via covariance-based clipping, entropy-term additions, or low-probability regularisation (Cui et al., 2025; Cheng et al., 2025a; Huang et al., 2025). A second family reweights whole trajectories by rarity, using current-policy ranks, strategy clusters, or outcome-level novelty (He et al., 2025; Hu et al., 2026b; Song et al., 2025). Our novelty bonus belongs to the trajectory-level family but differs from prior methods in its anchor: novelty is scored under a frozen reference model rather than the current policy, so the diversity target does not drift as training progresses. The bonus applies only to verifier-correct rollouts and uses top-k token NLL rather than LLM-judge clustering, rewarding alternative successful reasoning paths without per-trace judge calls at training time. Unlike work that optimises `pass@k` directly (Walder & Karkhanis, 2025), we retain the standard verifiable-reward objective and use novelty as an auxiliary exploration-preserving bonus.

Mechanism analyses of reasoning. A growing literature analyzes reasoning mechanisms by labeling behaviors in chain-of-thought traces, including verification, backtracking, subgoal setting, and related cognitive operations (Gandhi et al., 2025; Kim et al., 2025). These studies show that RL reshapes the distribution of reasoning primitives, amplifying some and suppressing others (Gandhi et al., 2025; Bai et al., 2025). Closest to our taxonomy, Li et al. (2025) segment reasoning into functional episodes in the style of Schoenfeld's problem-solving protocols (Schoenfeld, 2014). Our analysis differs in scale and method: we distill judge annotations into a lightweight span classifier that can label 25k-token traces across tens of thousands of rollouts and ten checkpoints, avoiding both per-trace judge calls and injected behavior markers. Unlike prior analyses, we use the resulting primitive sequences to study out-of-domain transfer rather than tracking only per-trace counts.

3 Post-Training Pipeline and Results

Setup and evaluation. We use OLMo3-7B-Instruct-SFT (Team Olmo et al., 2025) as our primary base model and repeat the full recipe on Qwen3-4B-Instruct-2507 (Qwen Team, 2025) (Appendix O). OLMo3 is fully open, has undergone no prior RL fine-tuning, and retains upstream mathematical knowledge from pretraining and instruction tuning. This last property is critical as any post-training gain reflects improved search and verification over existing knowledge, not the injection of new mathematical content. Our post-training pipeline uses only constraint-satisfaction puzzles from the Simon Tatham collection (Tatham, 2024; Estermann et al., 2024) (Bridges, Pattern, Undead, and Galaxies), with no mathematics problems in either the SFT or RL stages. We evaluate on three hard out-of-domain mathematics benchmarks: OlymMATH-Hard (Sun et al., 2025a), HMMT (Balunović et al., 2025) (four rounds, February 2024 through February 2026, 123 problems; for Qwen3-4B the 63 problems from the two rounds released after its base model, which we call HMMT post-cutoff), and OMEGA (Sun et al., 2025b), all of which remain far from saturation for the models studied here and therefore probe the capability ceiling rather than benchmark familiarity. Our primary metric is `pass@32`, since a `pass@1` improvement can reflect either capability ceiling expansion, where new problems become solvable, or sharpening of an already-reachable path, and the two are indistinguishable at the single-sample level. Large-k pass rates isolate the former, and all evaluations use 32 rollouts per problem at temperature $T{=}0.6$.

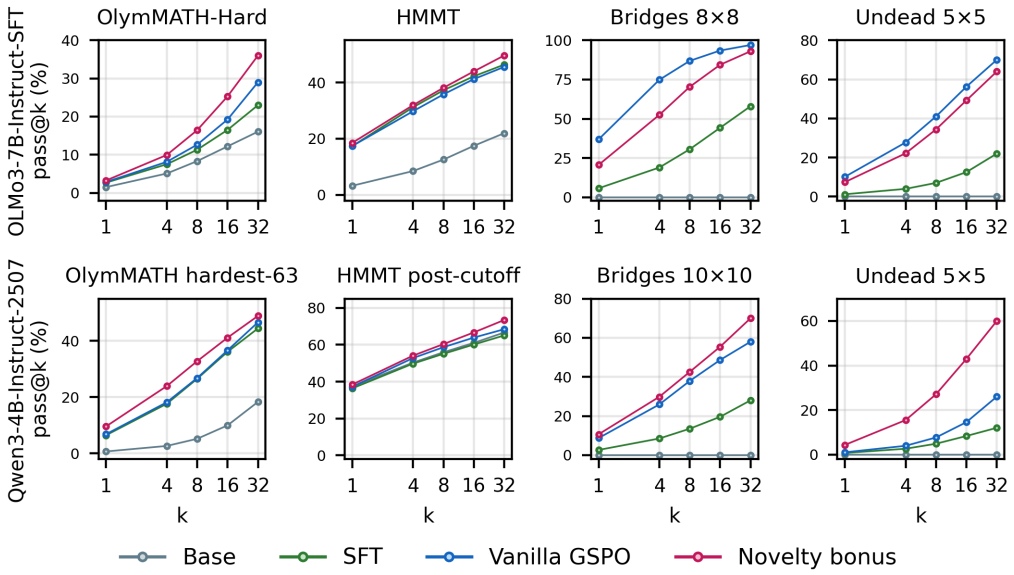


Figure 2: pass@k across the recipe for OLMo3-7B-Instruct-SFT (top) and Qwen3-4B-Instruct-2507 (bottom), on held-out mathematics (left) and held-out puzzle grids larger or harder than the training grids (right). Puzzle-only post-training raises pass@k on both domains in both families, and the novelty bonus leads on mathematics.

Training Recipe. The training pipeline has two stages. First, because the base model has near-zero puzzle competence, direct verifier RL would provide little reward variance, so we perform supervised fine-tuning on rejection-sampled reasoning traces distilled from Deep-Seek Reasoner (DSR), keeping only traces that verify as correct under the puzzle scorers. Second, from this SFT checkpoint we train two multi-puzzle variants of the RL algorithm Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025). Both variants use the same verifier reward, which combines an exact reward for fully correct boards with a partial completeness reward. The partial completeness reward is proportional to how much of the final board state matches the solution. The vanilla GSPO condition uses only this verifier reward, while the novelty condition additionally rewards correct rollouts whose reasoning traces are assigned low probability under the frozen SFT reference. We use GSPO rather than token-level Group Relative Policy Optimization (GRPO) because our target responses are 25–28k tokens and GSPO trained reliably in long-context pilots where GRPO was unstable. Full hyperparameters and reward details are in Appendix H.

Results. Puzzle SFT alone raises OlymMATH-Hard pass@32 from 16.0% at the base to 23.0%. Both RL variants, trained from this SFT checkpoint, improve further: vanilla GSPO reaches 29.0% (+6pp over SFT), while the novelty variant reaches **36.0**% (+13pp over SFT and +7pp over the matched vanilla GSPO baseline), as shown in Figure 2 and Table 3. Notably, the novelty gain grows with k, from a modest gap at pass@1 to +7pp at pass@32, showing that it increases the capability ceiling rather than sharpening a dominant path. The same pattern holds on OMEGA Explorative (Sun et al., 2025b) and HMMT (Figure 5; Table 3). Finally, on larger-grid puzzle evaluations, both RL variants raise pass@k over SFT across the full sampling budget on Bridges 8×8 and Undead 5×5, suggesting the learned strategies extrapolate across grid scale. The same recipe on Qwen3-4B-Instruct-2507, trained one grid size up so that its stronger SFT checkpoint starts in the same trainable band and evaluated on tiers its base does not saturate, gives gains both in domain and on held-out hard mathematics, with the novelty bonus ahead of vanilla GSPO at every k on every suite (Figure 2, bottom row; Appendix O).

The remainder of the paper asks why puzzle-only training transfers to hard mathematics: what reasoning behaviors does SFT introduce, what does vanilla RL amplify or suppress, and which changes accompany the sampled ceiling gains observed above? Section 4 introduces the primitive-

Table 1: Primitive vocabulary used by the span classifier.

Primitive	Meaning	Primitive	Meaning
PLAN	High-level approach.	SETUP	Notation or auxiliaries.
ENUMERATE	Cases or candidates.	HYPOTHESIZE	Tentative claim to test.
COMPUTE	Calculation or manipulation.	CHECK	Verify against constraints.
BACKTRACK	Reject and return.	SUMMARIZE	Consolidate partial results.
OTHER	Filler or formatting.		

level analysis, Sections 4.1–4.3 apply it to the recipe stages, and Section 5 uses the resulting diagnosis to derive and test the novelty bonus.

4 Analysis Framework: Reasoning Primitives

Aggregate metrics such as response length and token entropy reveal that training changes the policy (Shypula et al., 2025; Deshpande et al., 2025), but not *what* reasoning behaviours change. We therefore analyse chain-of-thought traces as sequences of reasoning primitives: short, semantically coherent spans that play canonical roles such as planning, computing, checking, hypothesising, or backtracking. Because traces can reach ~ 25k tokens, we split each trace into spans at paragraph boundaries and discourse markers (e.g., *Case 1*, *Wait*, *Alternatively*, *Going back*), yielding contiguous spans of 80–250 tokens (full algorithm in Appendix I). Each span is then classified with its preceding context, producing one primitive sequence per rollout.

Table 1 gives the nine-class vocabulary. We developed the labels using a jury of LLM judges (DeepSeek-V3.2 and DeepSeek-Chat (Liu et al., 2025)), iterating until inter-judge agreement reached 86% on a held-out span set, and distilled the judges into a nine-way span classifier that reaches 0.80 macro-F1 against judge consensus, since judge-based labelling is impractical at scale. We validate the classifier against blind human annotation rather than judge labels alone: on the two primitives our claims rest on, BACKTRACK and CHECK, it reaches 75% precision and 75% recall against the primary annotator, that precision does not shift across the checkpoints being compared, and the two are rarely confused with each other (Appendix N). Finally, for each checkpoint we track three statistics of the primitive sequence: primitive episode counts per trace, the CHECK→BACKTRACK transition count per trace, and the uninterrupted-chain fraction, the share of a trace's spans in its longest stretch with no BACKTRACK (Appendix J). Tracking these across checkpoints lets us identify which reasoning patterns each training stage amplifies or suppresses.

4.1 Puzzle SFT Induces a Reasoning Vocabulary

Before any RL, we ask what puzzle SFT alone contributes to reasoning behaviour. Applying our primitive classifier (Section 4), we compare base, DSR-teacher, and post-SFT rollouts on held-out puzzles, and base versus post-SFT rollouts on OlymMATH-Hard, for both model families. If puzzle-only SFT transfers, the math rollouts should acquire the same primitive vocabulary even though no mathematics appears in the SFT corpus. Figure 3 tests this at the primitive level.

Puzzle SFT induces the vocabulary used later. On held-out puzzles, post-SFT rollouts in both families inherit the DSR teacher's primitive profile, showing that puzzle SFT induces the teacher's reasoning style in addition to improving accuracy; backtracking in particular rises well above both the base model and the teacher, from 0.1 to 7.4 episodes per trace for OLMo3-7B on Bridges 8×8 and from 3.1 to 6.4 for Qwen3-4B on Bridges 10×10, against 2.2 in the teacher corpus (Figure 3, top). The transfer test asks whether this vocabulary appears out of domain. It does, in both families, from different starting points. OLMo3-7B's base has almost no structured vocabulary on OlymMATH-Hard, 2.8 primitive episodes per trace and no backtracking; after puzzle SFT its mathematics traces carry 19.6 episodes per trace with the same primitives as its puzzle traces. Qwen3-4B's base already reasons with the full set on mathematics; puzzle SFT raises this from 32 to 54 episodes per trace, roughly doubling planning, enumeration, and hypothesis episodes (Figure 3, bottom). Since the SFT corpus contains only puzzle traces, these changes indicate that puzzle SFT induces a reusable

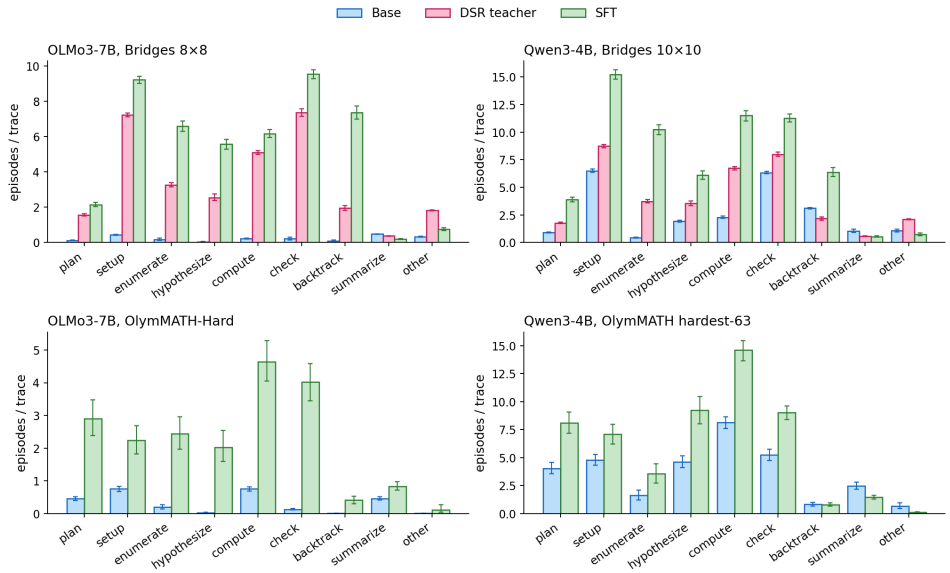


Figure 3: Primitive episodes per trace for the base model, the DSR teacher corpus, and the model after puzzle SFT, on held-out puzzles (top) and held-out mathematics (bottom), for both families. Puzzle SFT raises backtracking above both the base and the teacher and installs a reusable reasoning-primitive vocabulary that carries over to its mathematics traces.

reasoning vocabulary. Having established that puzzle SFT makes these primitives available, we next ask how vanilla RL reallocates them.

4.2 What Vanilla RL Changes: Longer Uninterrupted Chains

What does vanilla RL change relative to puzzle SFT? Our answer is not that it discovers a new reasoning vocabulary. The SFT model already has the relevant primitives; vanilla GSPO instead changes how strongly the model commits to one part of that vocabulary. On OlymMATH-Hard, the dominant shift is toward longer uninterrupted execution: once the model is on a line of reasoning, it carries that line forward without stepping back to an earlier state. The same shift appears on harder puzzles and in both model families.

From primitives to uninterrupted chains. Frequency alone does not tell us whether these primitives are chained together, so we next measure the chain structure directly. We define the *uninterrupted-chain fraction* of a trace as the share of its spans that fall in the longest stretch containing no BACKTRACK: a value of 1.0 means the model pursued a single line of reasoning from start to finish without ever abandoning it, while a value of 0.5 means the longest line it pursued before backtracking was half the trace. Unlike a count of primitives, the fraction does not depend on how long the trace is, which lets us compare models whose response lengths differ (Appendix J). On OlymMATH-Hard, the fraction rises from 0.93 after SFT to 0.98 after vanilla GSPO for OLMo3 and from 0.85 to 0.90 for Qwen3 (Figure 4(a)). The shift is far larger on held-out puzzles, where SFT traces still backtrack often: from 0.55 to 0.93 on Bridges 8×8 for OLMo3 and from 0.56 to 0.82 on Bridges 10×10 for Qwen3. The effect is developmental: on the puzzle training distribution, the fraction rises from 0.63 to 0.93 early in vanilla GSPO training (Figure 4(c)).

On OlymMATH-Hard, this shift tracks the vanilla-RL gain: `pass@32` rises from 23.0% at SFT to 29.0%, a $+6$pp gain that accompanies the move toward uninterrupted chains above. This gain represents a ceiling extension: in contrast to the convergence pattern reported by Yue et al. (2025), puzzle RL adds genuinely new solvable problems rather than sharpening the SFT distribution. Paired inspection of SFT and vanilla-GSPO rollouts on problems solved by both checkpoints gives the same interpretation: vanilla GSPO usually improves *execution* of an approach rather than changing the *set* of approaches attempted (Appendix M, Part A). The novelty-bonus variant later breaks this trend,

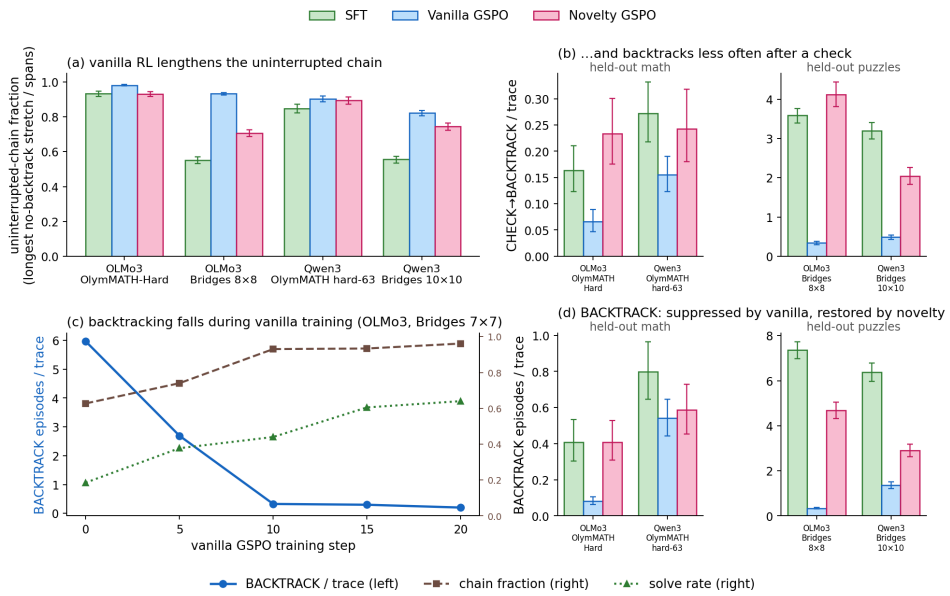


Figure 4: **(a)** Uninterrupted-chain fraction, **(b)** CHECK→BACKTRACK transitions per trace, and **(d)** BACKTRACK episodes per trace, on one held-out mathematics set and one held-out puzzle grid per family; **(c)** backtracking and puzzle solve rate across vanilla-GSPO training on the puzzle training distribution. Vanilla GSPO lengthens the chains and suppresses backtracking; the novelty bonus reduces the backtrack suppression.

gaining another $+7$pp `pass@32` while its uninterrupted-chain fraction moves back toward the SFT value; the next section shows what vanilla RL suppresses.

4.3 What Vanilla RL Suppresses

The shift toward uninterrupted chains in Section 4.2 comes with a cost. As vanilla GSPO commits more strongly to a single line of reasoning, it suppresses the backtracking that puzzle SFT introduced. On OlymMATH-Hard, BACKTRACK falls from 0.41 to 0.08 episodes per trace (-80%) for OLMo3 and from 0.80 to 0.54 (-32%) for Qwen3; on held-out puzzles it falls from 7.4 to 0.3 on Bridges 8×8 and from 6.4 to 1.4 on Bridges 10×10 (Figure 4(d)). On the puzzle training distribution itself the suppression is sharper and visibly developmental: across vanilla-GSPO training, BACKTRACK falls by 95% while puzzle solve rates rise substantially (Figure 4(c)). The novelty bonus restores part of the suppressed backtracking in both families (Figure 4(d)), fully on OLMo3 math and partially on puzzles. Vanilla RL therefore does not merely lengthen its chains; it also prunes the recovery move that would let the model recover when a direct line of reasoning fails.

Backtracks that follow a check. A backtrack that immediately follows a check is the natural place to look for the model acting on its own verification, so we track the CHECK→BACKTRACK transition separately from how often the model checks or backtracks in general. It falls on every dataset (Figure 4(b); the additional held-out sets are in Appendix F): from 0.16 to 0.07 per trace on OlymMATH-Hard and from 3.6 to 0.3 on Bridges 8×8 for OLMo3, and from 0.27 to 0.16 and from 3.2 to 0.5 for Qwen3. Vanilla GSPO keeps verifying, but it less often changes course right after a check. The novelty bonus restores the transition on all four datasets, to 0.23 and 4.2 for OLMo3 and 0.24 and 2.0 for Qwen3, including Qwen3 OlymMATH-Hard, where the raw backtrack count rises only marginally. In both families, CHECK is the most common step preceding a backtrack after SFT and after novelty GSPO (Appendix G).

Why does vanilla RL suppress its own backtracking? A simple explanation is that, on the puzzle training distribution, vanilla RL makes the model better at recognizing and executing the right approach. Once the policy can usually identify a productive route, the highest-reward behavior is not

to search over alternatives; it is to carry that route farther through setup, computation, and checking. This is the same pressure that produces the longer uninterrupted chains of Section 4.2. In that regime, backtracking introduces detours without often changing the outcome, so continued on-policy training can make it less likely. Figure 4(c) is consistent with this interpretation: as puzzle accuracy rises, the policy gives up the recovery move that would help when its first approach is wrong.

But hard math is not the training distribution. That tradeoff is useful only when the first approach is usually right. Hard math often violates this assumption: a promising route may lead to a dead end, and solving requires trying another representation, case split, or construction. A model whose backtracking has been suppressed can still execute a chosen approach deeply, but it is less able to recover when that approach fails. The novelty GSPO variant, also trained from the SFT checkpoint, preserves the BACKTRACK primitive at its SFT level on OLMo3 math and raises OlymMATH-Hard `pass@32` to 36.0%, compared with 29.0% for vanilla GSPO and 23.0% for SFT (Table 3). We see similar gains on math for Qwen3 with a similar pattern in the primitives: on HMMT, `pass@32` rises from 68.4% for vanilla GSPO to 73.5% under the novelty bonus (Appendix O).

Section 5 introduces a targeted intervention designed to preserve the backtracking this section showed vanilla RL suppresses, and shows that the restored backtracking accompanies the held-out math gain in both families.

5 A Frozen-Reference Novelty Bonus Preserves Recovery

Sections 4.2 and 4.3 give us a design target: preserve the backtracking that vanilla RL suppresses while still training on the same verifiable puzzle reward. We do this by adding a reward bonus for correct rollouts whose reasoning is unusually surprising under the frozen SFT reference model. The bonus is applied only to verifier-correct rollouts and normalised within the correct rollouts on the same prompt: it rewards alternative successful reasoning paths, not arbitrary low-probability text.

Bonus definition. For each rollout r on prompt p, we evaluate the generated tokens under the frozen SFT reference π_{SFT}, with token loss $\mathrm{NLL}_t(r) = -\log \pi_{\mathrm{SFT}}(r_t \mid r_{<t}, p)$. We define the rollout novelty score as the mean of the top-k largest token NLLs:

$$s(r) = \frac{1}{k} \sum_{t \in \mathrm{top}_k(\mathrm{NLL}(r))} \mathrm{NLL}_t(r), \qquad k = 100. \tag{1}$$

For correct rollouts within each prompt group, we z-score and clip the novelty scores before adding them to the verifier reward:

$$z_p(r) = \mathrm{clip}\left(\frac{s(r) - \mu_p}{\sigma_p},\ -2,\ 2 \right), \qquad R_{\mathrm{novelty}}(r) = R_{\mathrm{base}}(r) + \alpha z_p(r) \mathbb{1}[r \text{ correct}]. \tag{2}$$

where μ_p and σ_p are computed over correct rollouts for prompt p. We set $\alpha = 0.1$. Thus incorrect rollouts receive no novelty credit, and correct rollouts receive more credit when their reasoning is more novel relative to the SFT reference than other correct rollouts on the same prompt. Pseudocode is in Appendix K.

Why frozen-reference top-k NLL? The frozen reference keeps the novelty target stable. If novelty were ranked under the current policy, as in unlikeliness-style rewards (He et al., 2025), the target would drift toward whatever the model is already doing well, weakening the diversity signal as training progresses. The top-k statistic makes the signal local enough to matter at our trace lengths: with $\sim$25k-token rollouts, mean NLL has near-zero within-group variance because most tokens are high-probability formatting or structural choices. Using only the $k = 100$ highest-NLL tokens isolates rare decision points and amplifies the within-group signal by 10–15$\times$ (Appendix L).

Effect. The novelty-bonus variant starts from the same SFT model as vanilla GSPO and uses the same data, algorithm, training budget, and number of optimisation steps. It raises OlymMATH-Hard `pass@32` to $\mathbf{36.0}\%$, compared with 29.0% for vanilla GSPO and 23.0% for the shared SFT starting point (Table 3; Figure 2). Mechanistically, it restores the backtracking that vanilla GSPO suppressed: on OlymMATH-Hard, BACKTRACK returns to its SFT level (0.41 episodes per trace)

Table 2: Backtracking statistics and `pass@32`, vanilla GSPO → novelty bonus, on the held-out mathematics sets of both families. Where the novelty bonus raises backtracking after a check, held-out accuracy rises with it.

Family	Evaluation	n	BACKTRACK / trace	CHECK→BACKTRACK / trace	`pass@32` (%)
OLMo3-7B	OlymMATH-Hard	100	$0.08 \rightarrow 0.41$	$0.07 \rightarrow 0.23$	$29.0 \rightarrow 36.0$
OLMo3-7B	HMMT	123	$0.17 \rightarrow 0.97$	$0.14 \rightarrow 0.63$	$45.5 \rightarrow 49.6$
Qwen3-4B	OlymMATH-Hard (hardest 63)	63	$0.54 \rightarrow 0.59$	$0.16 \rightarrow 0.24$	$46.6 \rightarrow 48.9$
Qwen3-4B	HMMT post-cutoff	63	$0.66 \rightarrow 0.87$	$0.19 \rightarrow 0.44$	$68.4 \rightarrow 73.5$

and the CHECK→BACKTRACK transition rises above it (0.16 at SFT, 0.07 under vanilla GSPO, 0.23 under the bonus); on Qwen3-4B the transition returns to its SFT level on the hardest-63 problems (0.27, 0.16, 0.24; Figure 4(b)). On the held-out puzzle grids BACKTRACK returns to about two thirds of its SFT level for OLMo3 (Bridges 8×8) and to about half for Qwen3 (Bridges 10×10; Figure 4(d)). This restoration comes with shorter uninterrupted chains: the chain fraction moves back toward its SFT value on every dataset in Figure 4(a). The bonus therefore does not improve by committing still harder to a single line of reasoning; it trades some of that commitment for recovered ability to try and abandon alternatives. We close by asking whether the restored backtracking accompanies the transfer gain.

Restored backtracking accompanies the held-out math gain. When the novelty bonus restores backtracking, does the held-out math gain come with it? Table 2 puts the two side by side on the four held-out mathematics evaluations of Figure 2: for each family, the vanilla and novelty checkpoints evaluated on OlymMATH-Hard and on HMMT. Each row compares the same problems across the two RL runs, which start from the same SFT model and differ only in the reward bonus. In all four rows, moving from vanilla GSPO to the novelty bonus raises the CHECK→BACKTRACK transition per trace and raises `pass@32`. The raw backtrack count rises in three of the four, the exception being Qwen3 OlymMATH-Hard, where it rises only marginally. Appendix M, Part B, gives two full trace pairs showing the same pivot in text: the novelty run backtracks after a failed check and reaches an answer that the vanilla run commits past.

6 Conclusion and Limitations

We post-trained models from two families with no mathematics in the post-training stages studied here and found that puzzle-only SFT and RL can raise the sampled ceiling on hard mathematics. The primitive analysis suggests a more specific lesson than "more reasoning helps." Puzzle SFT makes a reusable vocabulary of planning, checking, and backtracking available out of domain; vanilla GSPO then commits to longer uninterrupted chains of reasoning, but at the cost of suppressing backtracking. The frozen-reference novelty bonus improves transfer while bringing that backtracking back toward its SFT level. More broadly, cross-domain RL transfer may depend not only on how far a model can pursue an approach, but also on whether training preserves the ability to abandon that approach when it fails. Taken together, our contribution is an empirical test of whether and how RL-trained LLMs can turn a shared problem-solving vocabulary into measurable out-of-domain math gains.

Limitations. Our results cover two base models at two scales, OLMo3-7B-Instruct-SFT and Qwen3-4B-Instruct-2507; we did not train beyond the 7B scale because of compute cost, and the novelty-bonus hyperparameters were tuned on the first model only. Our held-out evaluations are competition mathematics with verifiable final answers. The same transfer may reach other domains that reward search behaviour, such as competitive programming and multi-step planning in agentic tasks, but we leave that to future work. Extended discussion is in Appendix B.

AI use statement

In this work, we used generative AI tools for editing and polishing prose, for assisting with analysis and plotting code, and for literature search. All AI-assisted code was reviewed and tested by the authors, and all AI-suggested references were checked against the original sources. An LLM judge was additionally used as a labelling aid inside the method itself (the span-classifier training pipeline

of Section 4 and Appendix N); its labels were validated against blind human annotation as reported there. We did not use generative AI tools to generate research ideas, experimental results, or figures from data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work improves out-of-domain reasoning of two small open models using verifier-only puzzle reinforcement learning. The standard dual-use considerations for reasoning LLMs apply, and the method does not introduce capability surfaces beyond those already accessible from the base model. All training data are synthetically generated constraint-satisfaction puzzles with programmatic verifiers; no human-subject data or personal information is involved. The human span-labelling study (Appendix N) was performed by the authors and colleagues on model-generated text only. Funding sources are disclosed in the acknowledgments of the camera-ready version.

Reproducibility statement

Both base models are publicly available (`allenai/OLMo-3-7B-Instruct-SFT` and `Qwen/Qwen3-4B-Instruct-2507`), as is the RL algorithm (GSPO). Full training hyperparameters for both stages and both families are listed in Appendix H, the novelty-bonus algorithm is given in Appendix K, and the span-segmentation procedure and classifier training are described in Appendix I with metric definitions in Appendix J. Detailed per-benchmark accuracy tables that back every headline number appear in Appendix C and Appendix O, and compute requirements are stated in Appendix A. Training scripts, the reward implementation, evaluation harness, and classifier checkpoints will be released upon publication.

References

Haoyue Bai, Yiyou Sun, Wenji Hu, Shi Qiu, Maggie Ziyu Huan, Peiyang Song, Robert Nowak, and Dawn Xiaodong Song. How and why llms generalize: A fine-grained analysis of llm reasoning from cognitive behaviors to low-level patterns. *ArXiv*, abs/2512.24063, 2025. URL `https://api.semanticscholar.org/CorpusID:284350292`.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025. doi: 10.48550/arXiv.2505.23281. URL `https://arxiv.org/abs/2505.23281`.

Jiangjie Chen, Qianyu He, Siyu Yuan, Aili Chen, Zhicheng Cai, Weinan Dai, Hongli Yu, Jiaze Chen, Xuefeng Li, Qiying Yu, Hao Zhou, and Mingxuan Wang. Enigmata: Scaling logical reasoning in large language models with synthetic verifiable puzzles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL `https://openreview.net/forum?id=fmnxunacr4`.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025b.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *AAAI Conference on Artificial Intelligence*, 2025a. URL `https://api.semanticscholar.org/CorpusID:279410424`.

Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Nilabjo Dey, Yonghao Zhuang, Yuheng Zha, et al. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025b.

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. *ArXiv*, abs/2505.22617, 2025. URL `https://api.semanticscholar.org/CorpusID:278959427`.

Vijeta Deshpande, Debasmita Ghose, John D. Patterson, Roger Beaty, and Anna Rumshisky. Diverse, not short: A length-controlled data selection strategy for improving response diversity of language models. *arXiv preprint arXiv:2505.16245*, 2025. URL https://arxiv.org/abs/2505.16245.

Benjamin Estermann, Luca A Lanzendörfer, Yannick Niedermayr, and Roger Wattenhofer. Puzzles: A benchmark for neural algorithmic reasoning. *Advances in Neural Information Processing Systems*, 37:127059–127098, 2024.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STaRs. *arXiv preprint arXiv:2503.01307*, 2025. COLM 2025.

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.

Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. *arXiv preprint arXiv:2506.02355*, 2025. EMNLP 2025.

Chuxuan Hu, Yuxuan Zhu, Antony Kellermann, Caleb Biddulph, Suppakit Waiwitlikhit, Jason Benn, and Daniel Kang. Breaking barriers: Do reinforcement post training gains transfer to unseen domains? In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=mvLhN0veUd.

Z Hu, Y Wang, Y He, J Wu, et al. Rewarding the rare: Uniqueness-aware reinforcement learning for creative problem solving. *arXiv preprint arXiv:2601.08763*, 2026b.

Maggie Ziyu Huan, Yuetai Li, Tuney Zheng, XiaoYu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *ArXiv*, abs/2507.00432, 2025. URL https://api.semanticscholar.org/CorpusID:280146966.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. *ArXiv*, abs/2510.03222, 2025. URL https://api.semanticscholar.org/CorpusID:281830220.

Joongwon Kim, Anirudh Goyal, Liang Tan, Hanna Hajishirzi, Srini Iyer, and Tianlu Wang. Astro: Teaching language models to reason by reflecting and backtracking in-context. *ArXiv*, abs/2507.00417, 2025. URL https://api.semanticscholar.org/CorpusID:280128357.

Ming Li, Chenrui Fan, Yize Cheng, Soheil Feizi, and Tianyi Zhou. Schoenfeld's anatomy of mathematical reasoning by language models. *arXiv preprint arXiv:2512.19995*, 2025.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

George Pólya. *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press, Princeton, NJ, 1945.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Alan H Schoenfeld. *Mathematical problem solving*. Elsevier, 2014.

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

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of llm generated content. *arXiv preprint arXiv:2504.12522*, 2025. URL https://arxiv.org/abs/2504.12522.

Herbert A Simon and Allen Newell. Human problem solving: The state of the theory in 1970. *American psychologist*, 26(2):145, 1971.

Yuda Song, Julia Kempe, and Rémi Munos. Outcome-based exploration for LLM reasoning. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2025. URL https://openreview.net/forum?id=VORSpYLBJ6.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, and Ji-Rong Wen. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*, 2025a.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization. *arXiv preprint arXiv:2506.18880*, 2025b. doi: 10.48550/arXiv.2506.18880. URL https://arxiv.org/abs/2506.18880.

Simon Tatham. Simon tatham's portable puzzle collection. https://www.chiark.greenend.org.uk/~sgtatham/puzzles/, 2024.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. Base model: allenai/OLMo-3-7B-Instruct-SFT.

Christian Walder and Deep Karkhanis. Pass@ k policy optimization: Solving harder reinforcement learning problems. *arXiv preprint arXiv:2505.15201*, 2025.

Zhen Hao Wong, Jingwen Deng, Runming He, et al. LogicPuzzleRL: Cultivating robust mathematical reasoning in LLMs via reinforcement learning on logic puzzles. *arXiv preprint arXiv:2506.04821*, 2025.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-RL: Unleashing LLM reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2a36EMSSTp.

Yang Yu. Pass@k metric for RLVR: Diagnostic, not objective. *arXiv preprint arXiv:2511.16231*, 2025.

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. From $f(x)$ and $g(x)$ to $f(g(x))$: Llms learn new skills in rl by composing old ones. *arXiv preprint arXiv:2509.25123*, 2025.

Yang Yue, 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? *arXiv preprint arXiv:2504.13837*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *ArXiv*, abs/2507.18071, 2025. URL https://api.semanticscholar.org/CorpusID:280017753.

A Compute resources

Stage 1 SFT runs on $8\times$B200 for $\approx$9 hr. The two Stage 2 RL variants (vanilla GSPO and the novelty-bonus GSPO) are parallel branches from the same SFT checkpoint and each runs on $8\times$B200 for $\approx$70 hr. End-to-end reproduction of the headline OlymMATH-Hard 36.0% pass@32 number (SFT plus the novelty-bonus branch) is therefore approximately 632 GPU-hours; reproducing both Stage 2 branches together is approximately 1,192 GPU-hours. The full research project, including preliminary pilot runs, abandoned hyperparameter configurations, and ablations not reported in the paper, required roughly an additional 2–3$\times$ the reproduction budget on top.

B Limitations and future work

Our evidence comes from two base models at two scales, OLMo3-7B-Instruct-SFT and Qwen3-4B-Instruct-2507, and we did not train beyond the 7B scale because full-scale runs with long puzzle rollouts are computationally expensive. The novelty bonus has design choices that were tuned on the first model only: $\alpha = 0.1$ and top-$k = 100$ were selected from smaller pilot runs, and we did not re-tune them for the second family. Seed robustness of the OLMo3 headline is in Appendix D. Our held-out evaluations are competition mathematics with verifiable final answers; the same transfer may reach other domains that reward search behaviour, such as competitive programming and multi-step planning in agentic tasks, which we leave to future work. Finally, the span classifier was validated against human annotation on one model family and trace style (Appendix N).

C Detailed accuracy tables

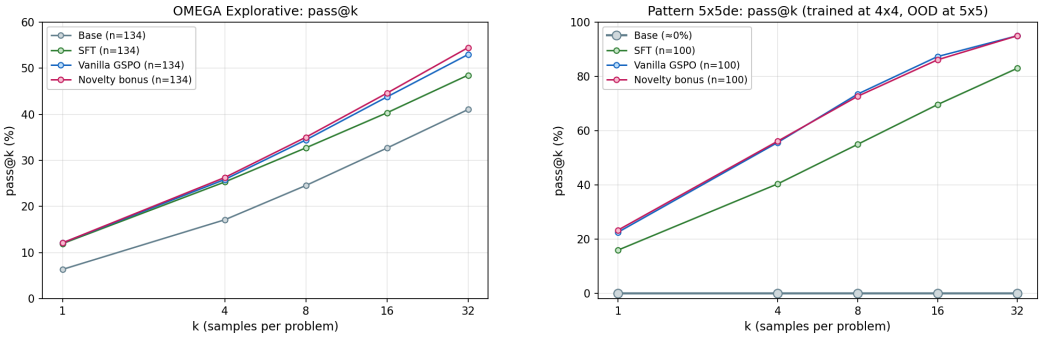
This appendix collects the per-benchmark accuracy tables referenced from Section 3. Table 3 reports held-out hard mathematics at `pass@`k for both families, and Table 4 the two saturating OLMo3 sets. Figure 2 summarises the same data as `pass@`k curves for the body benchmarks.

Table 3: Capability on held-out hard mathematics, both families. `pass@`k over 32 rollouts per problem at temperature 0.6. The recipe trains on constraint-satisfaction puzzles only; none of these sets appears in training.

	@1	@4	@8	@16	@32
OLMo3-7B-Instruct-SFT, OlymMATH-Hard (N=100)					
Base	1.5	5.1	8.3	12.2	16.0
+ Puzzle SFT	2.7	7.5	11.3	16.5	23.0
+ GSPO RL (vanilla)	2.8	8.1	12.6	19.2	29.0
+ GSPO RL with novelty bonus	**3.2**	**9.9**	**16.4**	**25.3**	**36.0**
OLMo3-7B-Instruct-SFT, HMMT (N=123)					
Base	3.3	8.5	12.6	17.5	22.0
+ Puzzle SFT	17.4	31.2	37.2	42.2	46.3
+ GSPO RL (vanilla)	17.4	29.8	35.8	41.2	45.5
+ GSPO RL with novelty bonus	**18.5**	**31.9**	**38.1**	**43.9**	**49.6**
Qwen3-4B-Instruct-2507, OlymMATH-Hard hardest-63					
Base	0.7	2.6	5.2	10.0	18.3
+ Puzzle SFT	6.4	17.6	26.5	36.1	44.4
+ GSPO RL (vanilla)	6.8	18.2	26.8	36.6	46.6
+ GSPO RL with novelty bonus	**9.6**	**23.9**	**32.8**	**41.2**	**48.9**
Qwen3-4B-Instruct-2507, HMMT post-cutoff (N=63)					
Base	37.4	50.3	55.9	61.0	66.7
+ Puzzle SFT	36.4	49.8	55.1	60.1	65.1
+ GSPO RL (vanilla)	37.2	52.8	58.7	63.9	68.4
+ GSPO RL with novelty bonus	**38.4**	**54.1**	**60.4**	**66.7**	**73.5**

Table 4: OLMo3-7B on the two saturating mathematics sets. AIME25 and OlymMATH-Easy saturate quickly with sampling at this scale, so RL gains there are necessarily small. `pass@1` and `pass@32` over 32 rollouts per problem.

	AIME25 (N=30)		OlymMATH-Easy (N=100)	
	@1	@32	@1	@32
Base	9.4	34.0	3.2	33.9
+ Puzzle SFT	27.0	73.1	14.2	72.0
+ GSPO RL (vanilla)	**29.6**	75.4	**15.9**	**73.0**
+ GSPO RL with novelty bonus	28.5	**76.7**	15.4	71.0



(a) OMEGA Explorative (N=134) (b) Pattern 5×5

Figure 5: OLMo3-7B `pass@`k on two further held-out sets: **(a)** OMEGA Explorative and **(b)** Pattern 5×5, one grid size above the training size. Both RL arms improve over SFT on both sets.

D SEED ROBUSTNESS OF THE OLMO3 HEADLINE

Both OLMo3 RL arms in Table 3 are single training runs. To check that the headline OlymMATH-Hard result does not depend on the seed, we retrained each arm from the same SFT checkpoint with the same script, changing only the data seed, and evaluated OlymMATH-Hard `pass@32` with the paper's protocol (32 rollouts per problem, temperature 0.6, same scorer). Every novelty run scores above every vanilla run (Table 5).

Table 5: OLMo3-7B, OlymMATH-Hard `pass@32` (%) by training seed.

Arm	Training seed	`pass@32`
vanilla GSPO	default (paper)	29.0
vanilla GSPO	7	26.0
novelty bonus	default (paper)	36.0
novelty bonus	7	33.0

E ADDITIONAL BENCHMARK `pass@`k CURVES

Figure 2 reports OlymMATH-Hard, HMMT and two held-out larger-grid puzzles for OLMo3-7B, and the corresponding hard-tier evaluations for Qwen3-4B. Figure 5 gives two further OLMo3-7B evaluations, OMEGA Explorative and Pattern 5×5. OlymMATH-Easy and AIME25 are reported in Table 4.

F BACKTRACKING ACROSS ALL HELD-OUT SETS

Table 6 reports the three statistics of Section 4.3 on every held-out set we evaluated with 32 rollouts per problem, for both families: BACKTRACK episodes per trace, CHECK→BACKTRACK transitions per trace, and the uninterrupted-chain fraction. Vanilla GSPO lowers the first two and raises the third on every row. The novelty bonus moves every statistic back toward its SFT value on the six OLMo3 rows and the two Qwen3 puzzle rows; on the two Qwen3 mathematics rows the CHECK→BACKTRACK transition recovers on both, and on HMMT the count and the chain fraction recover as well.

Table 6: Backtracking statistics on all held-out sets, read SFT → vanilla GSPO → novelty bonus. Means over 32 rollouts per problem; n is the number of problems. * marks a change from the previous column whose paired per-problem bootstrap 95% CI (5,000 resamples) excludes zero.

Family	Evaluation	n	BACKTRACK / trace	CHECK→BACKTRACK / trace	chain fraction
OLMo3-7B	OlymMATH-Hard	100	$0.41 \to 0.08^* \to 0.41^*$	$0.16 \to 0.07^* \to 0.23^*$	$0.93 \to 0.98^* \to 0.93^*$
OLMo3-7B	OlymMATH-Easy	100	$0.31 \to 0.08^* \to 0.37^*$	$0.15 \to 0.06^* \to 0.24^*$	$0.94 \to 0.98^* \to 0.94^*$
OLMo3-7B	AIME25	30	$0.29 \to 0.05^* \to 0.24^*$	$0.15 \to 0.04^* \to 0.16^*$	$0.94 \to 0.99^* \to 0.95^*$
OLMo3-7B	HMMT	123	$0.81 \to 0.17^* \to 0.97^*$	$0.37 \to 0.14^* \to 0.63^*$	$0.87 \to 0.96^* \to 0.87^*$
OLMo3-7B	Bridges 8×8	100	$7.36 \to 0.34^* \to 4.69^*$	$3.59 \to 0.34^* \to 4.12^*$	$0.55 \to 0.93^* \to 0.71^*$
OLMo3-7B	Undead 5×5	100	$5.64 \to 0.36^* \to 2.51^*$	$2.93 \to 0.35^* \to 2.12^*$	$0.55 \to 0.94^* \to 0.72^*$
Qwen3-4B	OlymMATH-Hard (hardest 63)	63	$0.80 \to 0.54^* \to 0.59$	$0.27 \to 0.16^* \to 0.24^*$	$0.85 \to 0.90^* \to 0.89$
Qwen3-4B	HMMT post-cutoff	63	$0.91 \to 0.66^* \to 0.87^*$	$0.36 \to 0.19^* \to 0.44^*$	$0.84 \to 0.89^* \to 0.86^*$
Qwen3-4B	Bridges 10×10	50	$6.37 \to 1.36^* \to 2.90^*$	$3.20 \to 0.49^* \to 2.04^*$	$0.56 \to 0.82^* \to 0.74^*$
Qwen3-4B	Undead 5×5	50	$7.77 \to 2.12^* \to 5.01^*$	$3.45 \to 0.72^* \to 3.40^*$	$0.47 \to 0.79^* \to 0.65^*$

G WHICH PRIMITIVE PRECEDES A BACKTRACK

Figure 6 decomposes the BACKTRACK episodes of Section 4.3 by the primitive that immediately precedes them, for the same four held-out datasets and three checkpoints as Figure 4. Each bar is the mean number of X →BACKTRACK episode transitions per trace. In both families, CHECK is the most common predecessor after SFT and after novelty GSPO, and the CHECK→BACKTRACK transition is the one the novelty bonus raises on every dataset (to or above its SFT level for OLMo3).

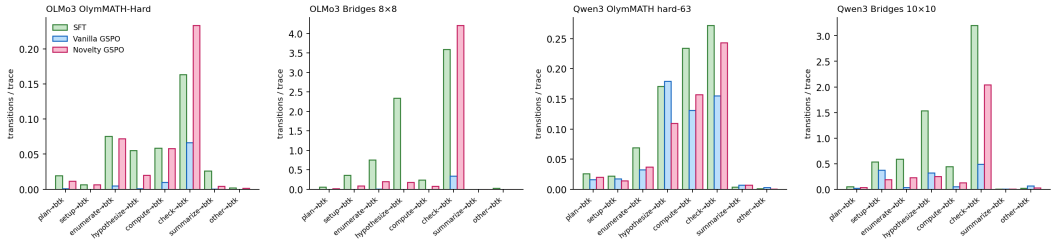


Figure 6: Transitions into BACKTRACK per trace, split by the preceding primitive, on OlymMATH-Hard and Bridges 8×8 (OLMo3-7B) and OlymMATH-Hard hardest-63 and Bridges 10×10 (Qwen3-4B).

H TRAINING HYPERPARAMETERS

Tables 7–8 give the OLMo3-7B-Instruct-SFT settings. The Qwen3-4B-Instruct-2507 runs (Appendix O) use the same RL settings (LoRA rank 64 / α 128, learning rate 5×10^{-5}, batch 128, temperature 0.8, 3,000 / 28,000-token prompt and response lengths, one epoch) and differ only in the base model, the SFT initialisation (its epoch-8 puzzle-SFT checkpoint), and training grids one size up. Both stages train LoRA adapters (rank 64, all-linear); stage-specific α values are in Tables 7–8.

Both runs use n_{gen}=8 on 8×B200 GPUs. All other RL hyperparameters are identical.

Verifier reward components. The exact term is binary: it is 1 iff the extracted `<answer>` verifies against the puzzle solution and 0 otherwise. The completion reward is a dense board-completion

Table 7: Stage 1 (SFT) hyperparameters. The SFT corpus contains rejection-sampled DSR puzzle traces only; we keep traces whose final answer verifies against the puzzle scorer and discard the rest. Epoch 5 is selected as the SFT endpoint (best held-out puzzle pass@1).

Setting	Value
Base model	OLMo-3-7B-Instruct-SFT
LoRA rank / α / target	64 / 64 / `all-linear`
Optimizer / learning rate	AdamW / 2×10^{-4}
Train batch size	32 (8 GPUs $\times$ micro-batch 1×4 gradient-accum. steps)
Max sequence length	32,000 (corpus filtered to $\leq$ 28k tokens)
Epochs / endpoint	8 trained / epoch 5 used downstream
Precision	bfloat16 training; fp32 LoRA merge

Table 8: Stage 2 (RL) hyperparameters. The vanilla-GSPO baseline and the novelty-bonus variant share all of these settings; the only difference is whether the novelty bonus (Section 5, α=0.1, top-k=100, z-clip 2) is added on top of the GSPO reward.

Setting	Value
Init / KL reference	SFT epoch 5, fp32-merged checkpoint
LoRA rank / α / target	64 / 128 / `all-linear`
Optimizer / learning rate	AdamW / 5×10^{-5} (cosine, min ratio 0.3)
Train batch size	128 prompts
Rollouts per prompt	8
PPO mini-batch / epochs	1024 / 4
Sampling	T=0.8, top-p=1.0
Max prompt / response length	3,000 / 28,000 tokens
GSPO clip ratios	ϵ_{low}=2.5×10^{-3}, ϵ_{high}=5×10^{-3}
KL coefficient β	10^{-3} (low-variance KL)
Reward	exact + completion reward (p=3, changed-cell weight 2) + 0.1·format reward
Epochs over data / total steps	1 (no data recycling) / 70 steps
GPUs	8×B200

score used before exact solving is common: after strict shape checks, it computes the fraction of grid cells matching the ground-truth solution, gives cells that differ from the initial puzzle board weight 2, and raises the weighted fraction to power $p = 3$. The format reward is an XML-structure reward in $[0, 1]$: it gives 0.25 each for exactly one closed `<reasoning>` tag, exactly one closed `<answer>` tag, reasoning appearing before answer, and no duplicate opening tags. Thus the table's 0.1·format-reward contribution is at most 0.1 and is intended only to stabilize response formatting.

I SPAN SEGMENTATION AND CLASSIFIER TRAINING

Span segmentation algorithm. Each reasoning trace is processed as follows. First, the reasoning block is extracted from `<reasoning>` tags if present; otherwise the full response is used. The text is then split at paragraph boundaries (double newlines) and further split within each paragraph at discourse markers detected by a multi-line, case-insensitive regex covering: numbered steps (*Case N*, *Step N*, *Option N*); sequential connectives (*First*, *Second*, *Then*); hypothetical openers (*Suppose*, *Assume*, *Let's*); verification openers (*Check*, *Verify*, *Confirm*); self-correction markers (*Wait*, *Actually*, *Hmm*); redirection markers (*Instead*, *Alternatively*, *Going back*); conclusion markers (*Therefore*, *Hence*, *Thus*); and failure markers (*Contradiction*, *This doesn't work*). Short fragments below 80 tokens are merged into an accumulation buffer until the threshold is reached; long fragments above 250 tokens are split at sentence boundaries (period, exclamation mark, or question mark followed by whitespace). A trailing buffer below 40 tokens is merged into the previous span rather than emitted as a standalone fragment. Token counts during merging use a character-based estimate (character count divided by 4) for speed; the base model's tokenizer is invoked only for the final token count on each emitted span.

Span classifier training. We fine-tune Qwen3-1.7B (Qwen Team, 2025) as a nine-way sequence classifier over the primitive vocabulary of Table 1. Training data consists of 35k judge-labeled spans: 17k from puzzle reasoning traces and 18k from mathematics reasoning traces. Labels were produced by a jury of LLM judges (DeepSeek-V3.2 and DeepSeek-Chat (Liu et al., 2025)), iterating prompt refinements until inter-judge agreement reached 86% on a held-out span set. The resulting classifier reaches 0.80 macro-F1 against judge consensus on a held-out test set. All OLMo3 primitive sequences in Sections 4.1–5 use this classifier; the Qwen3 classifier is trained the same way on Qwen3 rollouts (Appendix O).

J METRIC DEFINITIONS

Let $s = (s_1, \ldots, s_L)$ denote the span-level primitive sequence of a trace (one label per span, Appendix I), and let $e = (e_1, \ldots, e_M)$ denote its *episode* sequence, obtained by collapsing each maximal run of identical adjacent labels into one episode. Episode counts are the unit of every per-trace primitive count in the paper; the chain fraction is computed on spans.

Primitive episodes per trace. For a primitive p, its count in a trace is $|\{m : e_m = p\}|$. The paper reports means over the 32 rollouts of each problem and then over problems.

CHECK→BACKTRACK transitions per trace. The number of adjacent episode pairs (e_m, e_{m+1}) with $e_m = \text{CHECK}$ and $e_{m+1} = \text{BACKTRACK}$. Because episodes collapse repeated labels, a run of several CHECK spans followed by a backtrack counts once.

Uninterrupted-chain fraction. The length of the longest contiguous run of spans containing no BACKTRACK, divided by L:

$$\text{chain}(s) \;=\; \frac{1}{L}\,\max_{i \le j} \big\{\, j - i + 1 \;:\; s_i, \ldots, s_j \neq \text{BACKTRACK} \,\big\}.$$

A trace that never backtracks has $\text{chain}(s) = 1$; a trace whose only backtrack sits at its midpoint has $\text{chain}(s) \approx 0.5$. The fraction depends on where backtracks fall, not on how many spans the trace has, so it is comparable across checkpoints whose response lengths differ.

Paired per-problem bootstrap. For a contrast between two checkpoints, each statistic is first averaged over the rollouts of each problem, the per-problem differences are formed on the problems common to both checkpoints, and the mean difference is resampled $5{,}000$ times over problems. A change is marked significant when the 95% percentile interval excludes zero.

K NOVELTY-BONUS ALGORITHM

Figure 7 gives the per-step computation that augments the GSPO reward with the frozen-reference novelty bonus described in Section 5. The bonus is reward shaping: it is computed once per rollout under the *frozen* SFT reference policy π_{ref} (no gradients), then added to the last valid token's reward before advantage computation. All other GSPO machinery (clip ratios, KL penalty, optimisation) is unchanged.

The bonus is applied only to correct rollouts ($a_i = 1$): incorrect rollouts already receive zero base reward, so adding novelty there would amount to rewarding incorrect novel reasoning. Within-group z-scoring with the z_{clip}=2 cap bounds $|\alpha z| \le 0.2$, i.e. at most 20% of the maximum base reward, which keeps the bonus shaping rather than dominant.

L WITHIN-GROUP SIGNAL ANALYSIS

The novelty bonus only has signal if SFT perplexity varies meaningfully across correct rollouts of the *same* prompt: with no within-group spread, the z-score in Figure 7 is undefined and the bonus collapses to zero. We measured this directly on 100 puzzle prompts (32 rollouts each, post-SFT) and on 100 OlymMATH-Hard prompts under both SFT and vanilla-GSPO checkpoints. Table 9 reports

Top-k frozen-reference novelty bonus (per training step).

Inputs: rollouts $\{r_i\}_{i=1}^{B}$ grouped by prompt id $\mathrm{uid}(r_i)$; per-token frozen reference log-probs $\log \pi_{\text{ref}}(r_i)$; response masks m_i; correctness flags $a_i \in \{0, 1\}$; per-token rewards R_i; hyper-parameters α=0.1, k=100, z_{clip}=2.

1. For each rollout i, compute the masked per-token NLL $\mathrm{neg_lp}_i = -\log \pi_{\text{ref}}(r_i) \odot m_i$ and the top-k score $s_i = \frac{1}{k}\sum_{t \in \mathrm{TopK}_k(\mathrm{neg_lp}_i)} \mathrm{neg_lp}_{i,t}$.
2. For each prompt group g, let $C_g = \{i \in g : a_i = 1\}$ be the correct rollouts. Skip the group if $|C_g| < 2$ or $\mathrm{std}(\{s_i\}_{i \in C_g}) < 10^{-6}$.
3. Compute within-group statistics $\mu_g, \sigma_g = \mathrm{mean}, \mathrm{std}$ of $\{s_i\}_{i \in C_g}$, and set $z_i = \mathrm{clip}((s_i - \mu_g)/\sigma_g, -z_{\text{clip}}, +z_{\text{clip}})$ for each $i \in C_g$.
4. Add the bonus at the last valid token of each correct rollout: $R_i[\text{last valid}] \mathrel{+}= \alpha \cdot z_i$.
5. Compute GSPO advantages from the modified token-level rewards R as usual.

Figure 7: Pseudocode for the frozen-reference top-k novelty bonus described in Section 5. The bonus is reward shaping: it is computed once per rollout under the frozen SFT reference and added to the last valid token's reward before advantage computation. All other GSPO machinery (clip ratios, KL penalty, optimisation) is unchanged.

the mean within-prompt standard deviation of two candidate per-rollout signals: (i) mean per-token negative log-probability over the full response, and (ii) mean negative log-probability of the top-k most surprising tokens.

Table 9: Within-prompt standard deviation (nats/token) of per-rollout SFT-NLL signals, averaged over 100 prompts. Mean NLL has near-zero spread: at our response lengths ($\sim$25k tokens), per-token NLL is dominated by formatting and connective tokens that share probability across all reasoning paths. Restricting to the top-k most surprising tokens isolates decision-critical positions and amplifies the within-group signal 10–15$\times$.

Signal	SFT puzzles	SFT math	Vanilla GSPO math
Mean NLL (full response)	0.024	0.024	0.024
Top-10 NLL	0.18	0.16	0.16
Top-100 NLL	0.276	0.260	0.245
Top-200 NLL	0.27	0.26	0.24

At a z-score cap of 2 and α=0.1, the mean-NLL signal would yield bonus magnitudes around $\pm 2 \times 0.024 \times 0.1 \approx 0.005$ on a base reward of order 1, below noise. The top-100 signal yields ≈ 0.05, which is where the bonus becomes a meaningful fraction of the base reward. We chose k=100 because the within-group std saturates between k=100 and k=200 (no further amplification), and smaller k becomes noisier across prompts of varying length.

M Paired CoT inspection on OlymMATH-Hard

This appendix reproduces four worked OlymMATH-Hard rollouts to illustrate the mechanisms claimed quantitatively in the body. *Part A* (§4.2) pairs an SFT rollout against a vanilla-GSPO rollout where both are correct, to show that vanilla GSPO improves execution without changing the set of approaches. *Part B* (§5) pairs a vanilla-GSPO rollout that fails against a novelty-bonus rollout that succeeds, on problems vanilla GSPO never solves, to show what the restored backtracking of Section 4.3 looks like in text. Throughout, we score every rollout with `math_verify`, consistent with the scoring used across all main-paper evaluations.

Part A: Execution of a shared approach (§4.2)

We pair, for the same problem, an SFT rollout and a vanilla-GSPO rollout that are both correct and share the same opening primitives. We pick two problems whose traces are typical of each checkpoint: a single verification pass for SFT, several compute–check cycles for vanilla GSPO.

In both cases the SFT and GSPO rollouts converge on the same algebraic reformulation in the same primitive sequence: single-variable substitution yielding a quadratic, then a vertex/endpoint range argument. They differ in how much compute–verify alternation follows: GSPO continues for one to three additional COMPUTE→CHECK cycles (numerical spot checks, boundary tests, sign discussion) before stopping, while SFT terminates after one verification pass. No new approach appears in the GSPO trace that is absent from the SFT trace.

Problem 1 (Algebra, trigonometric optimisation). Let $f(x) = \sin^4 \omega x - \sin \omega x \cos \omega x + \cos^4 \omega x$ with $\omega > 0$. If there exist $a, b \in [0, \pi]$ with $f(a) + f(b) = 9/4$, find the minimum ω. Gold answer: $\omega = 7/12$.

SFT:
[COMPUTE] Use $\sin^4\theta + \cos^4\theta = 1 - 2\sin^2\theta\cos^2\theta$. Let $t = \sin\omega x \cos\omega x$, so $f(x) = 1 - 2t^2 - t$. With $t = \frac{1}{2}\sin(2\omega x)$ and $y = \sin(2\omega x) \in [-1, 1]$ rewrite $f(y) = 1 - \frac{1}{2}y^2 - \frac{1}{2}y$. Critical point at $y = -\frac{1}{2}$, giving $f = 11/8$; endpoints $f(\pm 1) = 0$.
[CHECK] So $\max f = 11/8$ and $2 \cdot 11/8 = 11/4 \geq 9/4$, hence $f(a) + f(b) = 9/4$ is achievable. Need $\sin 2\omega x = -\frac{1}{2}$, i.e. $2\omega x \in \{7\pi/6, 11\pi/6\} \pmod{2\pi}$. For $x \in [0, \pi]$ this requires $2\omega\pi \geq 7\pi/6$, hence $\omega \geq 7/12$. Minimum $\boxed{7/12}$.

Vanilla GSPO:
[COMPUTE] Same substitution $t = \sin\omega x \cos\omega x$; same quadratic $g(t) = -2t^2 - t + 1$. Vertex at $t = -1/4$, $g(-1/4) = 9/8$; endpoints $g(\pm 1/2) \in \{0, 1\}$. Range of g on $[-1/2, 1/2]$ is $[0, 9/8]$.
[CHECK] To hit $f(a) + f(b) = 9/4$ the only option is $g(t_a) = g(t_b) = 9/8$, hence $t_a = t_b = -1/4$. Translating: $\sin 2\omega a = \sin 2\omega b = -\frac{1}{2}$, with $a, b \in [0, \pi]$ giving $\omega \geq 7/12$.
[COMPUTE] Solve $a = 7\pi/(12\omega)$ and $a = 11\pi/(12\omega)$ in $[0, \pi]$; the $11\pi/6$ solution forces $\omega \geq 11/12$, the $7\pi/6$ solution only $\omega \geq 7/12$. Tentative minimum $\omega = 11/12$ [...]
[CHECK] Re-derive at $b = \pi$, $\omega = 11/12$: first attempt gives $t_b = 1/4$ ("Wait, this contradicts earlier calculation. Let me recheck."). Recompute: $\sin(2\omega b) = \sin(11\pi/6) = -\sin(5\pi/6) = -\frac{1}{2}$, so $t_b = -1/4$ after all, $f(b) = 9/8$.
[COMPUTE] Re-examine the smaller bound $\omega = 7/12$: at this ω, $a = b = \pi$ gives $2\omega a = 7\pi/6$, $\sin = -\frac{1}{2}$, $t = -1/4$, $f(\pi) = 9/8$, so $f(a) + f(b) = 9/4$. The pair is allowed because the problem only asks for existence of $a, b \in [0, \pi]$.
[CHECK / ENUMERATE / CHECK / COMPUTE / CHECK / SUMMARIZE] Test $\omega = 1/2$ (below $7/12$): no $x \in [0, \pi]$ satisfies $\sin 2\omega x = -\frac{1}{2}$, confirming the bound is tight from below. Conclude $\boxed{7/12}$.

Both checkpoints reduce f to a single-variable quadratic and identify the vertex condition $\sin 2\omega x = -\frac{1}{2}$. SFT stops after one CHECK pass and asserts the answer; GSPO continues with two additional COMPUTE→CHECK cycles, a numerical self-correction at $b = \pi$ and an explicit lower-bound test at $\omega = 1/2$, before committing to $\omega = 7/12$.

Problem 2 (Algebra, range of a rational expression). If real x, y satisfy $x^2 - y^2 = 4$, find the range of $\dfrac{1}{x^2} - \dfrac{y}{x}$. Gold answer: $(-1, 1)$.

SFT:
[SETUP] $x^2 - y^2 = 4 \Rightarrow x \neq 0$. Let $y = kx$; $x^2(1 - k^2) = 4 \Rightarrow x^2 = 4/(1 - k^2)$, requiring $|k| < 1$.
[COMPUTE] $E = 1/x^2 - y/x = (1 - k^2)/4 - k = -(k^2 + 4k - 1)/4$. Let $f(k) = k^2 + 4k - 1$. Vertex at $k = -2$ (outside $(-1, 1)$); $f(-1) = -4$, $f(1) = 4$, so on $(-1, 1)$ f ranges over $(-4, 4)$.
[SUMMARIZE] Hence $E = -f(k)/4$ has range $(-1, 1)$. (Earlier "± 1.25" boundary computations dismissed in one sentence as miscalculated.) $\boxed{(-1, 1)}$.

Vanilla GSPO:
[SETUP] Note $x^2 - y^2 = (x - y)(x + y) = 4$; rewrite target as $(1 - xy)/x^3$. Let $t = y/x \Rightarrow x^2 = 4/(1 - t^2)$, forcing $|t| < 1$.
[COMPUTE] Target $= (1 - t^2)/4 - t = (1 - 4t - t^2)/4 := f(t)$.
[CHECK] $f'(t) = -t/2 - 1$; $f' = 0$ at $t = -2$, outside $(-1, 1)$, so f is monotonic on

$(-1, 1)$. Limits: $f(t \to 1^-) = -1$, $f(t \to -1^+) = 1$.
[COMPUTE] $f'(0) = -1$, so f is strictly decreasing on $(-1, 1)$; image is $(-1, 1)$ open. Endpoints not attained (would require $x^2 - y^2 = 0$).
[CHECK] Sign of x doesn't matter (both signs give same $f(t)$); $x = 0$ impossible.
[COMPUTE] Spot-check $t = 0$ ($x = \pm 2$, target $= 1/4$); $t = 0.5$ (target -0.3125, matches $f(0.5)$); $t = -0.5$ (target 0.6875, matches $f(-0.5)$).
[CHECK] Limit checks: $t = 0.99 \Rightarrow f \approx -0.985$ (approaches -1); $t = -0.99 \Rightarrow f \approx 1.245$. Reconcile: the supremum is approached only as $t \to -1^+$ and never attained; numeric at $t = -0.99$ is consistent with the decreasing-to-1 picture.
[SUMMARIZE / COMPUTE] Range $\boxed{(-1, 1)}$.

Both checkpoints adopt the substitution $t = y/x$, derive the same quadratic $f(t) = (1 - 4t - t^2)/4$ on $(-1, 1)$, and reach $(-1, 1)$. SFT terminates after one range argument and waves off a sketchy boundary computation as a miscalculation; GSPO continues with two additional COMPUTE→CHECK cycles, numerical spot-checks at $t \in \{0, \pm 0.5\}$ followed by limit checks at $t = \pm 0.99$, before committing.

PART B: BACKTRACKING AFTER A FAILED CHECK (§5)

Section 5 reports that the restored backtracking and the held-out mathematics gain appear together. Below we reproduce the text of two OlymMATH-Hard problems that vanilla GSPO never solves in 32 rollouts and the novelty-bonus variant does, chosen because a BACKTRACK episode in the primitive sequence aligns with a textually visible pivot. The common pattern: vanilla GSPO commits to a first answer that the trace itself flags as suspect; the novelty-bonus variant performs the same opening computation, then backtracks to a different conclusion.

Problem 3 (Algebra, multivariate optimisation). Let x, y, z be positive reals; find the minimum of $f(x, y, z) = \dfrac{(2 + 5y)(3x + z)(x + 3y)(2z + 5)}{xyz}$. Gold answer: $241 + 44\sqrt{30} \approx 481.6$. Of 32 vanilla-GSPO rollouts: 6 box 480 (a loose AM-GM bound), 3 box 784 (the symmetric ansatz), the rest scattered. None match. The novelty-bonus variant's trace 3 succeeds, with one BACKTRACK episode at the pivot.

Vanilla GSPO failure (full trace):
[COMPUTE] Apply AM-GM to each numerator factor independently: $2 + 5y \geq 2\sqrt{10y}$, $3x + z \geq 2\sqrt{3xz}$, $x + 3y \geq 2\sqrt{3xy}$, $2z + 5 \geq 2\sqrt{10z}$. Multiply: numerator $\geq 16\sqrt{10y \cdot 3xz \cdot 3xy \cdot 10z} = 16 \cdot 30xyz = 480xyz$, hence $f \geq 480$.
[CHECK] "To check if equality holds, we need each AM-GM step tight: $2 = 5y \Rightarrow y = 2/5$; $3x = z$, $x = 3y \Rightarrow x = 6/5$, $z = 18/5$; $2z = 5 \Rightarrow z = 5/2$. This contradicts $z = 18/5 \neq 5/2$. Therefore the conditions cannot all be satisfied simultaneously, indicating equality might not be achievable. However, the AM-GM inequalities provide a lower bound, and since the function can approach this bound by choosing values close to those that satisfy the conditions, the minimum is indeed 480." $\boxed{480}$.

Novelty-bonus recovery:
[PLAN] Approach: try AM-GM on each factor; alternatively apply Lagrange multipliers to $\ln f$.
[HYPOTHESIZE] Try the symmetric ansatz $x = y = z = k$. The function reduces to $g(k) = 16(10k + 29 + 10/k)$, $k > 0$.
[COMPUTE → CHECK × 9] $g'(k) = 16(10 - 10/k^2) = 0 \Rightarrow k = 1$, second derivative positive, so $k = 1$ is a local min, $g(1) = 16 \cdot 49 = 784$. Spot-check $g(2) = g(0.5) = 864$, both higher.
[SUMMARIZE] "But is this the global minimum? We assumed $x = y = z$. Maybe there is a lower value when variables are not equal."
[BACKTRACK] Drop the symmetric ansatz. Use $F = \ln f$ and take $\partial F/\partial x = \partial F/\partial y = \partial F/\partial z = 0$ to get the actual critical-point system. Verify $x{=}y{=}z{=}1$ in equation (2): LHS $= 5/7 + 3/4 = 41/28 \neq 1$, so the symmetric ansatz is *not* a critical point, confirming the previous answer was wrong.
[CHECK → COMPUTE → CHECK → SUMMARIZE] Solve the critical-point system: it reduces to $x^2 = yz$, $15y^2 = 2x$, $2z^2 = 15x$. Substituting back gives $x = 1$, $y = \sqrt{2/15}$, $z = \sqrt{30}/2$. Plug in and simplify: $f_{\min} = 241 + 44\sqrt{30}$. $\boxed{241 + 44\sqrt{30}}$.

Both checkpoints reach for AM-GM first; vanilla GSPO multiplies the loose AM-GM bound, observes the equality conditions are inconsistent, and ships the bound anyway. Novelty starts on a stronger AM-GM (symmetric ansatz, giving 784), uses the partial-derivative system to verify whether the ansatz is a critical point, finds it is not, and *backtracks* to a Lagrange-style system whose explicit solution is the gold answer. The BACKTRACK episode marks the moment the novelty trace abandons the ansatz; the GSPO trace contains no such step.

Problem 4 (Geometry, regular tetrahedron). For a regular tetrahedron $ABCD$, M, N are midpoints of edges AB, AC and P, Q are centroids of faces ACD, ABD. Find the angle between lines MP and NQ. Gold answer: $\arccos(7/18)$. Of 32 vanilla-GSPO rollouts: 14 box $\arccos(-7/18)$ (the angle between the *vectors*, not the lines), 3 box $90°$, the rest miscellaneous. None match. The novelty-bonus variant's trace 25 succeeds with three BACKTRACK episodes.

Vanilla GSPO failure (full trace):
`[SETUP → COMPUTE]` Place $A = (1, 1, 1)$, $B = (1, -1, -1)$, $C = (-1, -1, 1)$, $D = (-1, 1, -1)$ (regular tetrahedron, edge $\sqrt{8}$). Midpoints $M = (1, 0, 0)$, $N = (0, 0, 1)$. Centroids $P = (-1/3, 1/3, 1/3)$, $Q = (1/3, 1/3, -1/3)$.
`[COMPUTE]` $\overrightarrow{MP} = (-4/3, 1/3, 1/3)$, $\overrightarrow{NQ} = (1/3, 1/3, -4/3)$. $\overrightarrow{MP} \cdot \overrightarrow{NQ} = -7/9$, $|\overrightarrow{MP}| = |\overrightarrow{NQ}| = \sqrt{2}$.
`[CHECK / final commit]` $\cos\theta = (-7/9)/2 = -7/18$. $\boxed{\arccos(-7/18)}$.

Novelty-bonus recovery (three BACKTRACK episodes):
`[SETUP → COMPUTE → CHECK]` Same coordinate system, same vectors, same $\cos\theta = -7/18$. Tentatively $\theta = \arccos(-7/18)$.
`[COMPUTE / SETUP / CHECK]` "However, the angle between two lines is defined as the smallest angle between their direction vectors. So if the calculated angle is obtuse, the actual angle between the lines is the acute supplement [...] $\cos(180° - \theta) = +7/18$, so the angle between the lines is $\arccos(7/18)$. But the problem says 'the angle between MP and NQ'. Usually they mean the angle between the vectors as directed, but sometimes the smallest. Need to check."
`[BACKTRACK #1]` "Maybe I made a mistake in the coordinate scaling. Let me try a different system where the tetrahedron has edge length 2." → Resolve with vertices at $(1/\sqrt{2}, \pm, \pm, \pm)$. Same vectors up to uniform scale; $\cos\theta = -7/18$ again.
`[BACKTRACK #2]` Re-examine the problem statement: "P and Q are the centroids of faces ACD and ABD", yes, what we used. → Recompute in original coordinates: still $-7/9$.
`[BACKTRACK #3]` Try a third coordinate system (edge length 1 with A at the origin). After one arithmetic error, confirm again that the *vector* angle has cosine $-7/18$.
`[CHECK → HYPOTHESIZE → COMPUTE → CHECK / final commit]` "In a regular tetrahedron, take two edges meeting at a vertex; the angle between them is $\arccos(1/3) \approx 54.7°$, which is acute. So the convention takes the acute angle. Probably here also." $\boxed{\arccos(7/18)}$.

Both checkpoints compute the same vectors, the same dot product, and the same intermediate $\cos\theta = -7/18$. Vanilla GSPO commits to the obtuse vector angle the next segment. Novelty spends most of its length checking whether $-7/18$ is "really" the answer: testing two alternative coordinate systems (BACKTRACKs #1 and #3), re-examining the problem statement (BACKTRACK #2), and finally invoking the "angle-between-lines-is-the-smallest-angle" convention. Three BACKTRACK primitives mark three attempts to reconcile the numeric answer with the geometric question; the last produces the convention switch that yields the gold answer. The recovery here is on *interpretation*, not technique; the underlying coordinate-and-dot-product method is unchanged between the two traces.

N HUMAN ANCHOR FOR THE SPAN CLASSIFIER

Section 4 summarises the blind human validation of the span classifier. This appendix gives the protocol and the matrices.

Protocol. A 285-span sheet was drawn stratified at 20 spans per (classifier class × checkpoint) cell and served in a cell-interleaved order, so that any prefix of the sheet is balanced. Annotators saw the span and its surrounding context and pressed one of five buttons (COMPUTE, CHECK, HYPOTHESIZE, BACKTRACK, OTHER), with no access to the checkpoint the span came from, the classifier's label, or the LLM-judge label. Two annotators with a mathematics background contributed the labels used here: the primary annotator labelled 140 spans (∼28 per class, ∼47 per checkpoint), the second labelled the balanced 90-span core set. All numbers below are against the primary annotator unless stated otherwise.

Grouping. The claims in the paper rest on two primitives: BACKTRACK (episode counts, the uninterrupted-chain fraction) and CHECK (the CHECK→BACKTRACK transition). We therefore score the classifier on the binary grouping *backtrack-check* versus *any-other*, which is the distinction those metrics depend on.

Table 10: Human validation of the span classifier ($n = 140$ blind-labelled spans). **(a)** Classifier (rows) vs. primary annotator (columns) on BACKTRACK, CHECK, and any-other. Collapsing the two primitives into one backtrack-check class gives 75% precision, 75% recall, 80% agreement, $\kappa = 0.58$. **(b)** The backtrack-check grouping partitioned by checkpoint, with 95% Wilson CIs: classifier precision does not shift across the checkpoints being compared.

(a) classifier vs. annotator

	BACK	CHECK	other	total	prec.
BACKTRACK	19	3	5	27	70%
CHECK	3	17	9	29	59%
any-other	6	8	70	84	83%
total	28	28	84	140	
recall	68%	61%	83%		

(b) backtrack-check, by checkpoint

checkpoint	n	prec.	recall
SFT	47	74% [51–88]	88% [64–97]
vanilla GSPO	47	79% [57–91]	79% [57–91]
novelty	46	72% [49–88]	62% [41–79]

Confusion between the two primitives. The transition metric requires that BACKTRACK and CHECK are told apart: in Table 10(a) only 3 of 27 classifier BACKTRACK spans and 3 of 29 classifier CHECK spans are the other primitive to the annotator; the remaining disagreements are with any-other.

Inter-annotator agreement. On the 90 spans both annotators labelled, the two humans agree with each other on 56% of spans ($\kappa = 0.44$) at five classes and 78% ($\kappa = 0.55$) on the backtrack-check grouping. The classifier agrees with the primary annotator at 54% ($\kappa = 0.43$) and 80% ($\kappa = 0.59$) on the same 90 spans. The classifier's disagreement with an expert is thus about the size of one expert's disagreement with another, which bounds what any labelling instrument can achieve on this task without a gold standard.

Scope. The design estimates precision within the stratified sample. It does not estimate population rates or population recall, which would require roughly an order of magnitude more labels; reported recall is within-sample only. The instrument is also validated on one model family.

O CROSS-FAMILY REPLICATION ON QWEN3-4B

Section 3 summarises the replication of the recipe on Qwen3-4B-Instruct-2507; Sections 4.2–5 report its primitive-level results alongside OLMo3's. This appendix gives the setup and the per-benchmark tables. Every accuracy here is recomputed from the evaluation rollouts with the unbiased `pass@`k estimator.

Setup. The pipeline is structurally unchanged: puzzle-only rejection-sampled SFT, then two RL branches initialised from the same SFT checkpoint, vanilla GSPO and novelty-bonus GSPO, with identical data, algorithm, hyperparameters and training budget, the bonus being the only difference. No mathematics appears in either stage. The Qwen SFT checkpoint is stronger on puzzles than the OLMo3 one (42.5% `pass@1` on the Bridges 7×7 grids used for OLMo3's RL, against 28.0%), so RL runs one grid size up, where the Qwen SFT starts at a comparable 24% (Bridges 8×8), the

same trainable-band criterion used in Section 3. Because the Qwen base is also much stronger on the standard mathematics sets, the mathematics evaluations use tiers it does not saturate: the 63 OlymMATH-Hard problems the base solves in at most 2 of 64 rollouts, and the HMMT rounds released after the base model (November 2025 and February 2026, 63 problems). Evaluation follows the paper's protocol: 32 rollouts per problem at T=0.6. We did not tune the novelty hyperparameters for this family.

Table 11: Qwen3-4B-Instruct-2507: `pass@`k on held-out harder puzzle grids and hard mathematics. Both RL arms start from the same puzzle SFT checkpoint and differ only in the novelty bonus. Bridges 10×10 and Undead 5×5 are one difficulty rung harder than the training mix; the training grids themselves saturate and cannot differentiate the arms. The four rows marked † are the panels shown in Figure 2.

Dataset	k	Base	+SFT	+vanilla GSPO	+novelty GSPO
Bridges 10×10†	1	0.0	2.6	8.7	**10.6**
	32	0.0	28.0	58.0	**70.0**
Undead 5×5†	1	0.0	0.7	1.0	**4.3**
	32	0.0	12.0	26.0	**60.0**
OlymMATH-Hard (full, N=100)	1	12.0	15.9	15.8	**18.9**
	32	47.8	61.0	65.0	**67.0**
OlymMATH hardest-63†	1	0.7	6.4	6.8	**9.6**
	32	18.3	44.4	46.6	**48.9**
HMMT post-cutoff (N=63)†	1	37.4	36.4	37.2	**38.4**
	32	66.7	65.1	68.4	**73.5**

The novelty arm is ahead of vanilla GSPO at every k on every suite. The margin is largest where the evaluation retains the most headroom: +12.0 and +34.0pp at `pass@32` on the two puzzle rungs, +5.1pp on post-cutoff HMMT, and +2.3pp on the hardest-63 subset, against +2.0pp on the full OlymMATH-Hard set, which the base already half-solves.

Primitive-level replication. We retrained the span classifier on Qwen3-4B rollouts using the same judge protocol as Section 4 and measured the three statistics of Section 4.3 at the same checkpoints (Table 12; the paired-bootstrap significance of each change is in Table 6). Vanilla GSPO lowers BACKTRACK and CHECK→BACKTRACK and raises the uninterrupted-chain fraction on every set; the novelty bonus raises CHECK→BACKTRACK on every set and BACKTRACK on three of the four, the exception being the hardest-63 set, where the count rises only marginally.

Table 12: Qwen3-4B-Instruct-2507: backtracking statistics on the held-out harder puzzle grids (top) and the hard-mathematics suites (bottom), SFT / vanilla GSPO / novelty bonus. Means over 32 rollouts per problem.

Dataset	Statistic	SFT	vanilla GSPO	novelty GSPO
Held-out harder puzzle grids				
Bridges 10×10	BACKTRACK / trace	6.37	1.36	2.90
	CHECK→BACKTRACK / trace	3.20	0.49	2.04
	uninterrupted-chain fraction	0.56	0.82	0.74
Undead 5×5	BACKTRACK / trace	7.77	2.12	5.01
	CHECK→BACKTRACK / trace	3.45	0.72	3.40
	uninterrupted-chain fraction	0.47	0.79	0.65
Hard-mathematics suites				
OlymMATH hardest-63	BACKTRACK / trace	0.80	0.54	0.59
	CHECK→BACKTRACK / trace	0.27	0.16	0.24
	uninterrupted-chain fraction	0.85	0.90	0.89
HMMT post-cutoff	BACKTRACK / trace	0.91	0.66	0.87
	CHECK→BACKTRACK / trace	0.36	0.19	0.44
	uninterrupted-chain fraction	0.84	0.89	0.86