PROTECTING CONTINUAL REASONING WITH DEPTH-INFORMED TRUST REGIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual reinforcement learning with verifiable rewards requires reasoning language models to acquire new capabilities without erasing behaviors learned from earlier tasks. This setting is especially challenging when replay is disallowed: after a task ends, the learner must preserve useful prior behavior using only compact statistics rather than old prompts or trajectories. We introduce **DToP** (**D**epth-informed **T**rust Region **o**ver **P**arameters), a replay-free continual adaptation method that summarizes past reasoning behavior through complementary functional sensitivities and constrains future optimization within the resulting anisotropic parameter-space geometry. At each task boundary, DToP first identifies *late-settling* token states, whose intermediate-layer next-token distributions approach the final-layer prediction only at relatively deep layers. For these selected states, DToP measures the parameter sensitivity of the local decision margin between the sampled token and its strongest alternative, and compresses the resulting gradients into a low-rank second-moment memory. On the remaining candidate states, DToP estimates diagonal policy Fisher information to capture broader distributional sensitivity. The two memories are scale-normalized and combined into a joint trust-region metric. During subsequent tasks, DToP does not constrain optimizer steps independently. Instead, each proposed update is projected so that the *cumulative* parameter displacement from the previous task endpoint remains inside the boundary-centered trust region. This explicitly controls the accumulation of many individually small updates along historically sensitive directions. We evaluate DToP on a heterogeneous continual reasoning benchmark spanning multiple reasoning domains. The results provide encouraging evidence that DToP can better balance the acquisition of new capabilities with the retention of previously learned ones, while retaining only compact parameter statistics and requiring no inference-time modification.

1 INTRODUCTION

Reasoning language models are increasingly adapted to new tasks through reinforcement learning with verifiable rewards (RLVR), using automatically checkable rewards to improve reasoning capabilities (Shao et al., 2024; Guo et al., 2025). When tasks arrive sequentially, learning a new capability can interfere with those acquired earlier (Luo et al., 2026a;b). We study this acquisition–retention challenge under a replay-free constraint: after each task, the learner may retain compact parameter statistics but cannot revisit historical prompts or responses.

Existing continual-learning methods follow three strategies. Replay revisits stored data; in continual RLVR, Continual Prompt Replay reuses earlier prompts (Luo et al., 2026a), which a replay-free learner cannot do. Regularization methods penalize parameter changes weighted by an importance estimate such as the Fisher information (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018), and Continual Policy Optimization applies a similar penalty to limit policy drift (Luo et al., 2026b). Gradient-projection methods keep each update orthogonal to subspaces retained from earlier tasks (Farajtabar et al., 2020; Saha et al., 2021; Wang et al., 2023). These methods summarize past behavior with a single notion of sensitivity over all token decisions, and they either constrain each update separately or penalize displacement softly, rather than bounding how far the parameters move over a whole task.

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.

Preserving past behavior does not require protecting every token equally. Some tokens determine critical reasoning steps or final answers (Lin et al., 2025; Wang et al., 2025), while others admit multiple valid continuations. We use layer-wise prediction dynamics (nostalgebraist, 2020; Belrose et al., 2023) to identify *deep-thinking tokens*, whose predictions settle only at late layers and often occur at key reasoning steps and answers (Chen et al., 2026). Changing these decisions could disrupt previously learned reasoning, motivating targeted protection through decision-margin sensitivity. For the remaining states, policy Fisher information (Amari, 1998; Kakade, 2001) captures broader distributional sensitivity without preserving individual tokens.

These complementary sensitivities identify which parameter changes may affect previous behavior, but preserving that behavior also requires controlling how such changes accumulate. During continual adaptation, successive optimizer updates may individually be small yet collectively produce substantial movement along sensitive directions. We therefore consider not only the sensitivity of each direction, but also the cumulative displacement along it, whereas trust regions in policy optimization bound each update separately (Schulman et al., 2015; 2017). This motivates a boundary-centered, anisotropic trust region that limits accumulated parameter movement according to historical sensitivity while allowing greater flexibility in less sensitive directions.

Based on these observations, we introduce **DToP** (**D**epth-informed **T**rust Region **o**ver **P**arameters), which protects two groups of candidate token states differently. At each task boundary, selected late-settling states contribute a low-rank memory of decision-margin gradients, retaining cross-parameter directions sensitive to changes in recorded token preferences. The remaining states contribute a diagonal policy-Fisher memory capturing coordinatewise distributional sensitivity. DToP accumulates and scale-normalizes both memories to define an anisotropic trust region. During subsequent training, it projects optimizer proposals to keep cumulative displacement from the previous task endpoint within this region. This limits movement along retained margin-gradient directions, following the interference-avoidance intuition of orthogonal-gradient continual learning (Farajtabar et al., 2020; Saha et al., 2021). After consolidation, DToP discards historical prompts, responses, and token states, requiring neither task-specific routing nor inference-time modifications.

We evaluate DToP on a six-task continual algorithmic reasoning stream with Qwen3-8B in two task orders and on a ten-task stream with Qwen3-4B, measuring both acquisition of each new task and retention of previously learned capabilities throughout sequential RLVR. We compare against standard sequential adaptation, Fisher-based EWC (Kirkpatrick et al., 2017), TreeLoRA (Qian et al., 2025), and replay-based Continual Prompt Replay (Luo et al., 2026a), with multi-task training as a non-continual reference, under matched RLVR training budgets, while explicitly accounting for differences in access to historical data.

Contributions. Our contributions are threefold:

- We introduce a state-dependent view of retention for continual reasoning, using late-settling behavior to distinguish targeted decision-margin sensitivity from broader policy sensitivity across token states.
- We compress these complementary sensitivities into a low-rank margin memory and a diagonal Fisher memory, which together define a compact anisotropic geometry over the trainable parameters.
- We introduce a boundary-centered projection that enforces an explicit trust-region budget on cumulative within-task displacement in the learned joint sensitivity geometry, providing a constrained alternative to penalty-based parameter regularization.

2 Related Work

Continual reinforcement post-training. Reinforcement fine-tuning can preserve prior capabilities better than supervised fine-tuning in several settings (Lai et al., 2026; Shenfeld et al., 2026), but sequential RLVR still faces interference. Continual Reasoning Gym introduces Continual Prompt Replay to revisit earlier tasks (Luo et al., 2026a), while Continual Policy Optimization regularizes parameter movement to limit historical policy drift (Luo et al., 2026b). These studies motivate replay-free retention during sequential RLVR.

Replay-free continual adaptation. Importance-based methods constrain parameter changes using Fisher information, optimization trajectories, or output sensitivity (Kirkpatrick et al., 2017; Schwarz et al., 2018; Zenke et al., 2017; Aljundi et al., 2018). Gradient-projection methods avoid retained subspaces associated with earlier tasks (Farajtabar et al., 2020; Saha et al., 2021). Related approaches structure or combine low-rank updates to reduce interference (Wang et al., 2023; Liang & Li, 2024; Biswas et al., 2026; Qiao & Mahdavi, 2026); TreeLoRA organizes adapters using task-gradient similarity (Qian et al., 2025). DToP constructs its geometry from token-level decision-margin and policy sensitivities, and projects cumulative displacement into the resulting trust region.

Internal prediction dynamics and trust regions. Intermediate-layer readouts expose how next-token predictions evolve through a model (nostalgebraist, 2020; Belrose et al., 2023). DTT uses settling depth to characterize deep-thinking tokens (Chen et al., 2026). DToP uses this signal to allocate targeted protection. Fisher geometry and trust-region methods provide tools for controlling local policy changes (Amari, 1998; Kakade, 2001; Schulman et al., 2015; 2017); our boundary-centered constraint applies historical sensitivity to cumulative within-task displacement.

3 Continual Reasoning with Verifiable Rewards

We study a sequence of T reasoning tasks. Task t supplies a prompt distribution $\mathcal{D}_t$ and a verifier $v_t(x, y) \in \{0, 1\}$, where x is a prompt and y a generated response. A model with a frozen backbone and trainable LoRA (Hu et al., 2022) parameters $\phi \in \mathbb{R}^P$ defines a policy $\pi_\phi(y \mid x)$. On each task, the objective is

$$J_t(\phi) = \mathbb{E}_{x\sim\mathcal{D}_t,\, y\sim\pi_\phi(\cdot|x)}[v_t(x, y)]. \tag{1}$$

We optimize this objective with GRPO (Shao et al., 2024), which compares rewards within groups of responses to the same prompt. The continual-learning objective is to acquire each new capability while retaining correctness on previous tasks.

Only current-task data are available for training. At task boundaries, the learner may retain parameter statistics, but not historical prompts, responses, or token states; evaluation uses separate held-out sets. A single adapter is used throughout, without task identifiers or inference-time routing. Let ϕ_t be the parameters at the end of task t and ϕ_0 the initialization. A token state $s = (x, y_{<j})$ consists of a prompt x and the first $j-1$ tokens of a response y, and $\pi_\phi(a \mid s)$ is the probability that the next token is $a \in \mathcal{V}$. The token that actually comes next in the response, y_j, is the *recorded token* of s.

4 DToP: Protecting Decisions during Continual Adaptation

4.1 Overview

At each task boundary, DToP partitions token states by settling depth and accumulates decision-margin and policy-Fisher memories. These memories shape a trust region that constrains cumulative adapter displacement on subsequent tasks. Figure 1 illustrates the procedure.

4.2 Partitioning token states at task boundaries

At the end of task t, DToP builds a candidate pool $\mathcal{P}_t = \{s_i\}_{i=1}^{n_t}$ from *verifier-correct* responses (x, y) in the final training batch, where $x \sim \mathcal{D}_t$ and $v_t(x, y) = 1$. Each state $s_i = (x, y_{<i})$ takes the prompt and the prefix before a selected position i of one such response; its recorded token y_i is the token y_j that follows. All states are re-evaluated at the task endpoint ϕ_t (Appendix A.2). These decisions are not equally worth protecting, so DToP looks for the ones the model finds hardest to make.

To measure this, we follow the layer-wise readout of Chen et al. (2026). Let $H_i^{(\ell)}$ be the hidden state at the last prefix position of s_i after layer ℓ of the L layers; the final normalization $\mathcal{N}$ and output projection W_{out} turn it into an intermediate next-token distribution. We define

$$p_i^{(\ell)} = \text{softmax}\big(W_{\text{out}}\mathcal{N}(H_i^{(\ell)})\big), \qquad \kappa_i = \min\big\{\ell \in \{1, \dots, L\} : \text{JSD}(p_i^{(\ell)}, p_i^{(L)}) \le \tau\big\}, \tag{2}$$

where JSD is the Jensen–Shannon divergence, which measures the difference between the intermediate and final token distributions, and $\tau > 0$ is a closeness threshold. Thus κ_i is the *first layer*

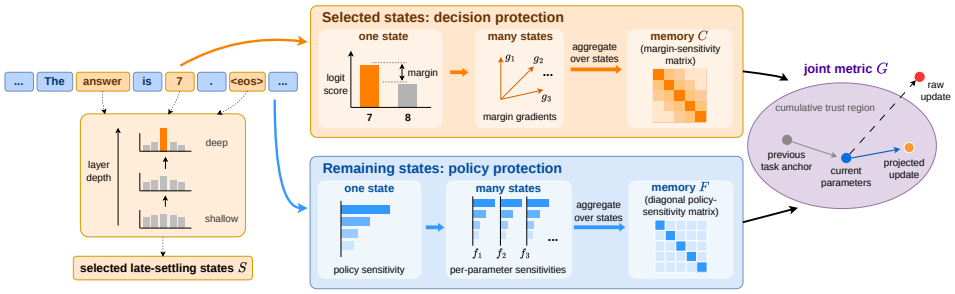


Figure 1: Overview of DToP and its boundary-centered cumulative trust-region projection.

whose prediction is sufficiently close to the final prediction. A larger κ_i means later settling, which serves as a selection signal for potentially consequential token decisions. Chen et al. (2026) call these *deep-thinking tokens*; they occur at key reasoning steps and answers, and responses with more of them are more often correct (see also Chuang et al., 2024; Schuster et al., 2022). DToP therefore protects these decisions directly.

Concretely, a state is deep-thinking if $\kappa_i \geq \lceil \beta L \rceil$ for a relative depth $\beta \in (0, 1)$. Across the candidate pool $\mathcal{P}_t$, *DToP keeps at most the K eligible states with the largest κ_i and calls this set $\mathcal{S}_t$.* The remaining candidate states form $\mathcal{R}_t = \mathcal{P}_t \setminus \mathcal{S}_t$, including eligible states excluded by the cap.

4.3 Protecting Decisions and Next-Token Distributions

Selected states $\mathcal{S}_t$ mark late-settling decisions in verified-correct responses. To preserve their recorded-token preferences during later tasks, DToP stores decision-margin gradients in $C_{\leq t}$. This memory retains sensitive cross-parameter directions, along which the trust region limits cumulative movement (Sec. 4.4).

Decision-margin sensitivity and memory. For $s_i \in \mathcal{S}_t$, denote by $z_{\phi,a}(s_i)$ the output-layer logit of vocabulary token a. The strongest alternative to the recorded token y_i at the endpoint is $y_i^- = \arg\max_{a \in \mathcal{V} \setminus \{y_i\}} z_{\phi_t,a}(s_i)$. Holding this competitor fixed, define the margin and its endpoint gradient:

$$m_i(\phi_t) = z_{\phi_t,y_i}(s_i) - z_{\phi_t,y_i^-}(s_i), \qquad g_i = \nabla_\phi m_i(\phi_t). \tag{3}$$

The margin measures the recorded token's preference over its strongest endpoint competitor. Writing $m_i(\phi)$ for the same fixed-competitor gap at any adapter ϕ, a small displacement δ changes it by

$$m_i(\phi_t + \delta) - m_i(\phi_t) = g_i^\top \delta + o(\|\delta\|_2). \tag{4}$$

Thus g_i identifies directions in which new-task learning could change this recorded decision. Restricting $(g_i^\top \delta)^2$ discourages large changes to the gap, with greater freedom in less sensitive directions. A recorded token need not be the endpoint argmax, so controlling its gap does not by itself guarantee that the token remains preferred.

DToP summarizes these gradients in an uncentered second moment, retaining parameter interactions in its leading directions. Let $\mathcal{I}_t^M$ contain the accepted margin observations at boundary t, $n_t^M = |\mathcal{I}_t^M|$, and $N_{\leq t}^M = N_{\leq t-1}^M + n_t^M$. DToP first computes their uncentered second moment at ϕ_t, then merges it with the historical memory:

$$\bar{C}_t = \frac{1}{n_t^M} \sum_{i \in \mathcal{I}_t^M} g_i g_i^\top, \qquad C_{\leq t} = \mathrm{Top}_r\left(\frac{N_{\leq t-1}^M C_{\leq t-1} + n_t^M \bar{C}_t}{N_{\leq t}^M} \right), \tag{5}$$

$$\delta^\top C_{\leq t} \delta \approx \frac{1}{N_{\leq t}^M} \sum_{i \in \mathcal{I}_{\leq t}^M} (g_i^\top \delta)^2. \tag{6}$$

Here $\mathcal{I}_{\leq t}^M$ denotes the conceptual collection of accepted observations across boundaries, counting repeated observations separately; individual historical gradients are not retained. The approximation in Eq. (6) reflects repeated rank truncation. Its quadratic energy tracks the mean squared

predicted change in stored logit gaps under δ. We retain the leading r spectral directions as $C_{\le t} = U_{\le t}^\top \Lambda_{\le t} U_{\le t}$, where $U_{\le t}$ has at most r orthonormal rows and $\Lambda_{\le t}$ is nonnegative diagonal. The leading directions summarize combinations of parameters with large observed margin sensitivity, allowing us to retain this information without storing all past gradients. We do not subtract the gradient mean, since a direction shared across states can itself be important to their decisions (Appendix A).

The remaining candidate states $\mathcal{R}_t$ also carry useful predictions, including deep-thinking states excluded by the selection cap, but we treat them as less critical than the selected decisions and protect them more loosely. Instead of each recorded token's margin, we discourage drift in their next-token distributions, and we accept a coarser memory for it: the margin memory keeps cross-parameter directions of the deep-thinking gradients, whereas for $\mathcal{R}_t$ we keep only the diagonal $f_{\le t}$ of the policy Fisher, a single P-vector that drops parameter interactions. Both memories combine statistics measured at the current endpoint with summaries from earlier task boundaries; token states are discarded after consolidation.

Policy Fisher sensitivity and memory. For $s_i \in \mathcal{R}_t$, define

$$p_{i,a} = \pi_{\phi_t}(a \mid s_i), \qquad g^\pi_{i,a} = \nabla_\phi \log \pi_{\phi_t}(a \mid s_i), \quad a \in \mathcal{V}.$$

The diagonal policy Fisher at this state is the probability-weighted mean of the squared log-probability gradients:

$$f_i = \sum_{a \in \mathcal{V}} p_{i,a} \, (g^\pi_{i,a})^2, \tag{7}$$

where a vector squared denotes *elementwise* squaring, so $f_i \in \mathbb{R}^P$. The full Fisher $F_i^{\text{full}} = \sum_a p_{i,a} g^\pi_{i,a} (g^\pi_{i,a})^\top$ appears in the local policy-KL expansion:

$$\mathrm{KL}(\pi_{\phi_t}(\cdot \mid s_i) \,\|\, \pi_{\phi_t+\delta}(\cdot \mid s_i)) = \frac{1}{2} \delta^\top F_i^{\text{full}} \delta + o(\|\delta\|_2^2). \tag{8}$$

We approximate F_i^{full} by $\mathrm{diag}(f_i)$, keeping coordinatewise sensitivity while omitting parameter interactions. Unlike the margin memory's retained low-rank directions, this branch provides a broad coordinatewise summary of distributional sensitivity. Similar diagonal Fisher approximations are standard in continual learning and Fisher-based optimization (Kirkpatrick et al., 2017; Schwarz et al., 2018; Kim et al., 2022); here the approximation reduces persistent storage from a $P \times P$ matrix to a P-vector. Its energy is a local KL surrogate, not an exact KL bound. We sample up to K_F states uniformly without replacement from $\mathcal{R}_t$ and estimate f_i with the stratified vocabulary estimator in Appendix A.3.

Let $\mathcal{I}_t^F$ be the accepted Fisher observations, with $n_t^F = |\mathcal{I}_t^F|$ and $N_{\le t}^F = N_{\le t-1}^F + n_t^F$. As with the margin branch, DToP first averages current-task estimates and then merges them with the historical memory:

$$\bar{f}_t = \frac{1}{n_t^F} \sum_{i \in \mathcal{I}_t^F} \widehat{f}_i, \qquad f_{\le t} = \frac{N_{\le t-1}^F f_{\le t-1} + n_t^F \bar{f}_t}{N_{\le t}^F}, \tag{9}$$

$$\delta^\top \mathrm{diag}(f_{\le t}) \delta = \frac{1}{N_{\le t}^F} \sum_{i \in \mathcal{I}_{\le t}^F} \delta^\top \mathrm{diag}(\widehat{f}_i) \delta. \tag{10}$$

Here $\mathcal{I}_{\le t}^F$ is the conceptual collection of accepted Fisher observations. If $n_t^F = 0$, the branch keeps its previous memory and count. The two counts are separate, so a missing observation in one branch does not dilute the other. Neither memory requires retaining historical prompts or token states after consolidation. The two branches complement each other: $C_{\le t}$ targets changes to selected gaps, whereas $f_{\le t}$ discourages broader distributional drift at the remaining states, including in directions outside the retained margin subspace.

4.4 Projecting cumulative adaptation into a joint trust region

The previous task endpoint as center. After task t, DToP retains $C_{\le t}$ and $f_{\le t}$. The trust region for task $t+1$ is centered at the adapter endpoint ϕ_t. The center and memories remain fixed throughout task $t+1$.

Let $\varphi_{t+1,k}$ be the adapter after k optimizer updates on task $t+1$, with $\varphi_{t+1,0} = \phi_t$. When the task is clear, abbreviate $\varphi_{t+1,k}$ as φ_k and let $\Delta\varphi_k$ be the base optimizer's proposed parameter increment. The cumulative displacement from the previous endpoint is $c_k = \varphi_k - \phi_t$. Applying the proposal unchanged would give $b_k = c_k + \Delta\varphi_k$. Thus b_k includes all earlier within-task updates, not only the latest proposal (Figure 2).

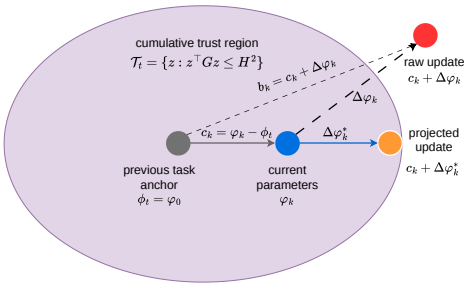


Figure 2: DToP applies the closest feasible update $\Delta\varphi_k^*$.

A joint sensitivity metric. Define the Fisher matrix $F = \mathrm{diag}(f_{\le t}) + \mu I_P$, with a small $\mu > 0$, and write $C = C_{\le t}$. Because the two memories have different scales, DToP calibrates them using the *first optimizer proposal* $\Delta\varphi_{\mathrm{cal}}$ after the boundary:

$$q_F = \Delta\varphi_{\mathrm{cal}}^\top F \Delta\varphi_{\mathrm{cal}}, \qquad q_M = \Delta\varphi_{\mathrm{cal}}^\top C \Delta\varphi_{\mathrm{cal}}, \qquad G = \frac{F}{q_F} + \rho\frac{C}{q_M}. \tag{11}$$

With valid positive calibration energies, each normalized memory has unit energy along that proposal, and $\rho \ge 0$ sets their relative weight *along the calibration direction*, not uniformly in parameter space. The resulting $G \in \mathbb{R}^{P\times P}$ is fixed for task $t+1$. It defines an anisotropic cost for moving the adapter: directions with large recorded sensitivity consume more of the trust-region budget. Positivity checks and fallback behavior are in Appendix A.5.

Projection of cumulative displacement. The trust region is the set of displacements $\mathcal{T}_t = \{z : z^\top G z \le H^2\}$, where $H > 0$ is its horizon. If the proposed total displacement b_k lies inside this region, DToP applies $\Delta\varphi_k$ unchanged. Otherwise it finds the closest feasible update:

$$\Delta\varphi_k^* = \arg\min_{u\in\mathbb{R}^P} \tfrac{1}{2}\|u - \Delta\varphi_k\|_2^2 \qquad \text{subject to} \quad (c_k+u)^\top G (c_k+u) \le H^2. \tag{12}$$

The objective keeps the applied increment as close as possible to the optimizer's proposal; the constraint requires the resulting adapter to remain in the historical-sensitivity region. The projected update is applied as $\varphi_{k+1} = \varphi_k + \Delta\varphi_k^*$. Because the constraint bounds the *entire* displacement $c_k + u$ rather than the update u alone, many small updates cannot accumulate outside the region.

With $b_k = c_k + \Delta\varphi_k$, the projection has a closed form.

Proposition 1 (Cumulative projection)**.** *Suppose $G \succeq 0$, $H > 0$, and ϕ_t and G remain fixed during task $t+1$. Problem (12) has the unique solution*

$$\Delta\varphi_k^* = (I_P + \lambda G)^{-1} b_k - c_k, \tag{13}$$

with $\lambda = 0$ if $b_k^\top G b_k \le H^2$, and otherwise the unique $\lambda > 0$ for which $(c_k+\Delta\varphi_k^*)^\top G (c_k+\Delta\varphi_k^*) = H^2$. Starting from $\varphi_0 = \phi_t$, every iterate produced by projection therefore satisfies $c_k^\top G c_k \le H^2$, independently of the number of optimizer updates.

The proof and an efficient scalar search for λ are in Appendix A.6. Along an eigenvector of G with eigenvalue γ_j, the proposed displacement is scaled by $1/(1+\lambda\gamma_j)$: directions with greater historical sensitivity are attenuated more strongly. Projection can therefore pull the adapter back toward its center rather than merely shorten the latest step. The diagonal-plus-low-rank structure of G permits efficient evaluation (Appendix A.5).

Why the trust region mitigates forgetting. To see what the constraint in Eq. (12) protects, build G from the exact second moment of the margin gradients and the full policy Fisher of each Fisher state, before any approximation. Dropping the nonnegative damping term, the constraint then implies

$$\frac{\rho}{q_M N_{\le t}^M} \sum_{i\in\mathcal{I}_{\le t}^M} \big(g_i^\top (c_k+u)\big)^2 + \frac{1}{q_F N_{\le t}^F} \sum_{j\in\mathcal{I}_{\le t}^F} \sum_{a\in\mathcal{V}} p_{j,a} \big((g_{j,a}^\pi)^\top (c_k+u)\big)^2 \le H^2. \tag{14}$$

Each term is a squared inner product between the total displacement from ϕ_t and a gradient stored at a previous boundary. With a small horizon H, the displacement must therefore stay nearly orthogonal to the margin gradients of past deep-thinking states and to the log-probability gradients of the remaining states; for the former, to first order (Eq. (4)), the recorded decisions change little. Orthogonality is a well-established way to limit interference with previously learned knowledge, both in continual learning, where updates avoid the gradient subspaces of earlier tasks (Farajtabar et al., 2020; Saha et al., 2021; Wang et al., 2023), and in orthogonal finetuning (Qiu et al., 2023; 2025). DToP applies it to the cumulative displacement and weights each direction by its measured sensitivity. In practice, DToP keeps the margin sum in the rank-r memory $C_{\leq t}$ and replaces each full Fisher by its diagonal, which protects the remaining states more loosely (Sec. 4.3).

The first task uses ordinary GRPO because no historical memory exists. Subsequent tasks use joint projection when both memories and calibration are valid. Algorithm 1 summarizes the full procedure.

5 Experiments

We evaluate the acquisition–retention balance across task orders, the two memories, state selection, and a longer stream; projection behavior and cost are in Appendix B.6 and B.8.

5.1 Experimental Setup

Benchmark and training. Our primary benchmark is a six-task stream from CRG Algorithmic-10 (Luo et al., 2026a), built on Reasoning Gym (Stojanovski et al., 2025): `base_conversion`, `binary_alternation`, `rotate_matrix`, `caesar_cipher`, `isomorphic_strings`, and `string_manipulation`. We evaluate this canonical order and its exact reverse using a frozen Qwen3-8B backbone (Yang et al., 2025) and a single trainable LoRA adapter. Training uses GRPO with binary verifier rewards. Replay-free methods use only current-task data. We also evaluate a ten-task Qwen3-4B stream in both orders. Methods share task-specific training and rollout budgets unless otherwise specified; full configurations are in Appendix B.2.

Evaluation. We evaluate each task-end policy on fixed held-out examples. The six-task setting uses 64 examples per task (384 in total); the ten-task setting uses 606 examples, with 30 for `jugs` and 64 for each remaining task. Let $R_{i,j}$ be the accuracy on task j after training task i. We follow the stage-wise evaluation of Luo et al. (2026a) and report acquisition (Acq), final average accuracy (Final), and backward transfer (BWT; Lopez-Paz & Ranzato, 2017):

$$\text{Acq} = \frac{1}{T}\sum_{t=1}^{T} R_{t,t}, \qquad \text{Final} = \frac{1}{T}\sum_{t=1}^{T} R_{T,t}, \qquad \text{BWT} = \frac{1}{T-1}\sum_{t=1}^{T-1}(R_{T,t} - R_{t,t}).$$

Acq measures how well each task is learned when first encountered, while BWT measures how that performance changes after learning subsequent tasks. Final captures the resulting performance over the whole stream. All three metrics give each task equal weight, and higher values are better. Comparisons use matched evaluation items. Six-task results are reported as mean $\pm$ standard deviation over three training seeds; the ten-task evaluation uses one seed.

Comparators. We compare with sequential GRPO (Seq), Fisher-based parameter regularization (EWC; Kirkpatrick et al., 2017), and TreeLoRA (Qian et al., 2025), adapted to GRPO. These baselines use no replay. Continual Prompt Replay (CPR; Luo et al., 2026a) provides a replay-enabled reference by replacing part of each batch with stored historical prompts. Multi-task reinforcement learning (MTRL) has access to all tasks from the start and uses the same total update budget, with equal prompt counts per task. It serves as a non-continual reference. All methods use a single LoRA adapter; baseline implementation details are in Appendix B.1.

5.2 Main Results

Preserving prior capabilities while learning new tasks. DToP achieves the highest mean Final accuracy among replay-free methods in both orders (Table 1), improving over the strongest baseline

by 2.1 points in canonical order and 6.9 points in reverse order. The distinguishing benefit is retention at comparable acquisition. In reverse order, for example, DToP and EWC acquire new tasks almost equally well (70.6% and 70.7%), but DToP has substantially less negative BWT (-1.4 versus -9.9). The same pattern holds against TreeLoRA in canonical order. Thus, DToP retains more of the capabilities it acquires without a substantial reduction in new-task learning.

Consistent gains across task orders. Task order changes which replay-free baseline is strongest: TreeLoRA leads in canonical order, whereas EWC leads in reverse order. DToP improves over both in each order, with positive Final gains in every training seed. Its advantage over sequential GRPO also grows from 4.2 to 17.1 points under order reversal, where sequential GRPO exhibits greater forgetting. Per-task acquisition and retention results are provided in Appendix B.4.

Comparison with access to historical data. CPR achieves the best canonical-order BWT, but allocates part of its training budget to earlier prompts and has lower acquisition. DToP achieves higher mean Final accuracy in both orders while devoting its training batches to the current task. Its final performance is also close to the multi-task reference, which sees all tasks from the start.

Table 1: Six-task Qwen3-8B results (%), mean $\pm$ standard deviation over three seeds. Bold: best replay-free result. MTRL does not depend on the task order.

	Canonical order			Reverse order		
Method	Final ↑	Acq ↑	BWT ↑	Final ↑	Acq ↑	BWT ↑
SEQ	62.5 ± 3.1	68.0 ± 1.1	-6.6 ± 2.8	52.3 ± 5.2	63.5 ± 2.7	-13.3 ± 4.1
EWC	63.0 ± 0.9	68.6 ± 0.8	-6.7 ± 1.9	62.5 ± 1.2	$\mathbf{70.7 \pm 2.4}$	-9.9 ± 2.0
TreeLoRA	64.6 ± 2.3	$\mathbf{69.4 \pm 1.3}$	-5.8 ± 1.3	56.3 ± 3.0	67.4 ± 1.8	-13.2 ± 1.7
DToP	$\mathbf{66.7 \pm 2.2}$	69.0 ± 3.4	$\mathbf{-2.8 \pm 1.4}$	$\mathbf{69.4 \pm 0.8}$	70.6 ± 1.6	$\mathbf{-1.4 \pm 1.7}$
CPR (replay)	66.1 ± 1.3	63.6 ± 1.3	3.0 ± 2.9	66.4 ± 1.4	68.8 ± 1.3	-2.9 ± 0.8
MTRL (multi-task)	68.0 ± 1.7	–	–	68.0 ± 1.7	–	–

5.3 COMPONENT ANALYSIS

We remove either term from Eq. (11), keeping the task stream and training protocol fixed.

Combining the memories gives the highest mean Final accuracy in both orders (Table 2). The clearest benefit of decision-margin protection appears in reverse order, where adding it to Fisher-only improves Final accuracy by 2.6 points, with a gain in every seed. The Fisher term is also useful: removing it reduces canonical-order acquisition by 2.6 points. In canonical order, the joint metric combines higher acquisition with sufficient retention to finish with the best average performance. An additional control for constraint strength is in Appendix B.3.

Table 2: Ablations on the six-task Qwen3-8B stream (%), mean $\pm$ standard deviation over three seeds. Bold: best result. Selection rules use the canonical order only.

	Canonical order			Reverse order
Variant	Final ↑	Acq ↑	BWT ↑	Final ↑
DToP (DTT states)	$\mathbf{66.7 \pm 2.2}$	$\mathbf{69.0 \pm 3.4}$	-2.8 ± 1.4	$\mathbf{69.4 \pm 0.8}$
Memory components				
Fisher-only	66.1 ± 2.5	67.4 ± 3.0	-1.6 ± 0.6	66.8 ± 2.3
Margin-only	65.6 ± 2.7	66.4 ± 3.0	$\mathbf{-0.9 \pm 0.3}$	68.5 ± 1.0
Margin-state selection				
Random	63.9 ± 0.7	65.4 ± 0.7	-1.8 ± 0.5	–
Lowest entropy	58.9 ± 0.9	64.0 ± 0.3	-6.1 ± 0.8	–

5.4 Depth-Informed State Selection

With a limited margin-memory budget, the choice of which decisions to protect becomes important. We compare depth-informed selection with two alternatives that select the same number of states from the same pool: uniform random selection and lowest endpoint entropy. Each rule assigns its selected states to the margin branch and the remaining states to the Fisher candidate pool, so the comparison evaluates how DToP allocates the two forms of protection. The metric and training configuration are shared.

Depth-informed selection yields the best mean Final accuracy (Table 2): 66.7 against 63.9 for random selection and 58.9 for lowest entropy. In paired seed differences, it exceeds lowest entropy by 7.8 ± 1.3 points and random selection by 2.8 ± 1.6 points. The lowest-entropy rule also falls below random selection in every seed. Thus, assigning margin protection to the model's most confident predictions is less effective than using layer-wise settling to guide the partition.

The state diagnostics help explain what distinguishes these choices. In the seed-101 run, depth-selected states settle later and have smaller endpoint margins than the lowest-entropy states (Appendix B.5). Depth therefore directs protection toward decisions that take more layers to settle, while lowest entropy favors decisions with larger existing logit gaps. Together with the accuracy comparison, this supports using internal prediction dynamics to allocate targeted protection beyond what endpoint confidence alone provides.

5.5 Continual Adaptation over Ten Tasks

We next evaluate a ten-task Qwen3-4B stream that adds `binary_matrix`, `cryptarithm`, `jugs`, and `ab` to the six-task set, with its own canonical order and its reverse (Appendix B.7). DToP achieves the highest Final accuracy in both orders, including comparisons with CPR and MTRL (Table 3). Its advantage over sequential GRPO is 11.9 points in canonical order and 17.7 points in reverse order, extending the empirical benefit to a longer stream and a smaller backbone.

Here, DToP's advantage is driven by continued acquisition. It reaches 47.3% acquisition, the highest among the continual methods, compared with 35.4% for sequential GRPO. Their canonical-order BWT is the same, so the higher acquisition carries through to higher Final accuracy.

Table 3: Ten-task Qwen3-4B results (%). Bold: best replay-free result. MTRL does not depend on the task order.

	Canonical order			Reverse order
Method	Final ↑	Acq ↑	BWT ↑	Final ↑
SEQ	19.1	35.4	**−18.1**	12.5
EWC	22.6	39.3	−18.6	19.2
TreeLoRA	23.0	43.9	−23.2	16.9
DToP	**31.0**	**47.3**	**−18.1**	**30.2**
CPR (replay)	26.5	29.4	−3.2	25.1
MTRL (multi-task)	22.8	–	–	22.8

6 Conclusion

We introduced DToP, a replay-free approach that combines depth-informed functional sensitivity with cumulative trust-region control for continual RLVR. Across our continual reasoning experiments, DToP achieves higher mean final accuracy and acquisition than sequential GRPO while retaining only compact parameter statistics, and the selector ablation suggests that layer-wise settling provides a useful signal beyond endpoint confidence. These results support sensitivity-aware geometric constraints as a promising approach to continual reasoning without replay.

REPRODUCIBILITY STATEMENT

The complete DToP formulation and optimization procedure are given in Section 4, with implementation details, boundary construction, estimators, projection, and ablation definitions provided in Appendix A. Training and method hyperparameters are reported in Appendix B.2, and the evaluation protocol and baselines are described in Section 5. An anonymized implementation and experiment configurations are provided with the supplementary material.

AI USE STATEMENT

Generative AI tools were used throughout this project to assist with debugging training and evaluation code, translation, language editing, and improving the clarity and presentation of the methodology and experiments. All technical content, experimental results, and scientific claims were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

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

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. arXiv:2303.08112, 2023. URL https://arxiv.org/abs/2303.08112.

Shristi Das Biswas, Yue Zhang, Anwesan Pal, Radhika Bhargava, and Kaushik Roy. ELLA: Efficient Lifelong Learning for Adapters in Large Language Models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1907–1924. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.84. URL https://aclanthology.org/2026.eacl-long.84/.

Wei-Lin Chen, Liqian Peng, Tian Tan, Chao Zhao, Blake JianHang Chen, Ziqian Lin, Alec Go, and Yu Meng. Think Deep, Not Just Long: Measuring LLM Reasoning Effort via Deep-Thinking Tokens. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.13517.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

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. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Minyoung Kim, Da Li, Shell Xu Hu, and Timothy M. Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073/pnas.1611835114.

Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Qingfu Zhang, Hongbin Liu, Gaofeng Meng, and Fei Zhu. Reinforcement Fine-Tuning Naturally Mitigates Forgetting in Continual Post-Training. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2507.05386.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37906–37918, 2025.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Lirui Luo, Guoxi Zhang, Hongming Xu, Rongqing Li, Cong Fang, and Lifeng Fan. Beyond Forgetting: Diagnosing and Harnessing Shared Reasoning in Continual RLVR. arXiv:2608.18574, 2026a. URL https://arxiv.org/abs/2608.18574.

Mao-Lin Luo, Zhe-Xu Wang, Zi-Hao Zhou, Bo Ye, Jian Zhao, Min-Ling Zhang, and Tong Wei. RL Forgets! Towards Continual Policy Optimization. arXiv:2607.04364, 2026b. URL https://arxiv.org/abs/2607.04364.

nostalgebraist. interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Yu-Yang Qian, Yuan-Ze Xu, Zhen-Yu Zhang, Peng Zhao, and Zhi-Hua Zhou. TreeLoRA: Efficient Continual Learning via Layer-Wise LoRAs Guided by a Hierarchical Gradient-Similarity Tree. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2506.10355.

Fuli Qiao and Mehrdad Mahdavi. Merge before Forget: A Single LoRA Continual Learning via Continual Merging. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.23017.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zeju Qiu, Weiyang Liu, Adrian Weller, and Bernhard Schölkopf. Orthogonal finetuning made scalable. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31934–31951, 2025. doi: 10.18653/v1/2025.emnlp-main.1627.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2103.09762.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015.

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

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537, 2018.

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:2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's Razor: Why Online Reinforcement Learning Forgets Less. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.04259.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning Gym: Reasoning environments for reinforcement learning with verifiable rewards. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 63612–63655. Curran Associates, Inc., 2025. doi: 10.52202/085713-1913. URL https://doi.org/10.52202/085713-1913.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.01939.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal Subspace Learning for Language Model Continual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.715. URL https://aclanthology.org/2023.findings-emnlp.715/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995, 2017.

A Additional Method Details

A.1 Notation and complete procedure

Table 4 collects the notation used in Sections 3–4. Task index t refers to a completed task at its boundary; within the subsequent task $t+1$, k indexes optimizer updates. We use φ_k as shorthand for $\varphi_{t+1,k}$ only when the current training task is clear, with $\varphi_{t+1,0} = \phi_t$.

Table 4: Notation for DToP. All gradients are with respect to the single trainable LoRA adapter; the backbone is frozen.

Symbol	Meaning
$\phi \in \mathbb{R}^P$, ϕ_t	Adapter vector and endpoint after task t
φ_k, $\Delta\varphi_k$	Current adapter and base-optimizer proposal within the next task
c_k, b_k	Current displacement $c_k = \varphi_k - \phi_t$ and proposed displacement $b_k = c_k + \Delta\varphi_k$
s_i, y_i, $\mathcal{V}$	Prompt–prefix state, its recorded next token, and vocabulary
$\mathcal{P}_t$, $\mathcal{S}_t$, $\mathcal{R}_t$	Correct-response candidate states, selected margin states, and their complement
$\mathcal{J}_t^F$	Fisher states sampled from $\mathcal{R}_t$
$H_i^{(\ell)}$, L, κ_i	Intermediate prefix representation, layer count, and first-crossing depth
m_i, g_i	Fixed-competitor logit gap and its endpoint gradient
$g_{i,a}^{\pi}$, f_i, $\widehat{f}_i$	Token log-probability gradient, diagonal Fisher, and its estimator
$\bar{C}_t$, $\bar{f}_t$	Accepted margin and Fisher sensitivity summaries from task t
$C_{\le t}$, $f_{\le t}$	Cumulative rank-r margin memory and Fisher vector through task t
n_t^M, n_t^F	Numbers of accepted observations at boundary t
$N_{\le t}^M$, $N_{\le t}^F$	Numbers of accepted margin and Fisher observations through task t
$\Delta\varphi_k^*$, q_F, q_M, G	Projected update, calibration energies, and joint metric
H, ρ	Trust-region horizon and additive margin weight

Algorithm 1 (DToP)**.**

1. Initialize the shared adapter at ϕ_0, empty memories $f_{\le 0}, C_{\le 0}$, and counts $N_{\le 0}^F = N_{\le 0}^M = 0$.
2. For tasks $t = 1, \dots, T$:
   (a) Start at ϕ_{t-1} and reset optimizer moments and calibration. If both memories $f_{\le t-1}, C_{\le t-1}$ exist, use the first GRPO optimizer proposal to calibrate G. Enable joint projection for this task if calibration succeeds.
   (b) Train on the current task. When projection is enabled, process every proposal, including the calibration proposal, using Eq. (12) with fixed center ϕ_{t-1} and metric G. Otherwise, apply the base-optimizer proposals unchanged.
   (c) Set the endpoint to ϕ_t. If $t < T$, construct $\mathcal{P}_t$, select $\mathcal{S}_t$, and sample Fisher states from its complement $\mathcal{R}_t$. Merge accepted observations into $C_{\le t}, f_{\le t}$ using separate counts; a branch with no new observations retains its memory and count.
   (d) After consolidation, discard boundary responses, states, and individual gradients. Each subsequent task uses ϕ_t as its center, including when neither branch adds an observation.
3. Return ϕ_T. Inference uses only the ordinary backbone and shared adapter, not the consolidation memories or task-specific routing.

A.2 Candidate construction and state partition

At each nonfinal boundary, we take at most eight verifier-correct responses (x, y) from the final training batch of task t, with $x \sim \mathcal{D}_t$ and $v_t(x, y) = 1$. We allocate up to 64 token positions approximately uniformly across these responses. Each selected position j of a response (x, y) yields a state $s_i = (x, y_{<j})$ whose recorded token y_i is the next token y_j. The pool $\mathcal{P}_t$ contains these states. We re-evaluate each full prompt and generated prefix at ϕ_t to compute endpoint sensitivities. The responses are reused from the final training batch, so their sampling policy is the policy used for that batch; consolidation requires no additional generation.

For each s_i, the hidden representation $H_i^{(\ell)}$ is taken at the last prefix position after layer ℓ. The first layer satisfying $\mathrm{JSD}(p_i^{(\ell)}, p_i^{(L)}) \le \tau$ defines κ_i in Eq. (2). We use natural logarithms, so the

threshold in Table 5 is measured in nats. Thus, κ_i records the first threshold crossing. We apply the relative-depth threshold, rank eligible states by κ_i, and select the deepest K as $\mathcal{S}_t$; ties use a deterministic order. States excluded by the cap, including otherwise eligible late-settling states, remain in $\mathcal{R}_t$. We draw up to K_F distinct Fisher states uniformly from $\mathcal{R}_t$ using a boundary-specific random generator independent of training and rollout randomness. The two branches therefore receive disjoint states. Figure 3 illustrates settling depths and the selection rule on two held-out responses.

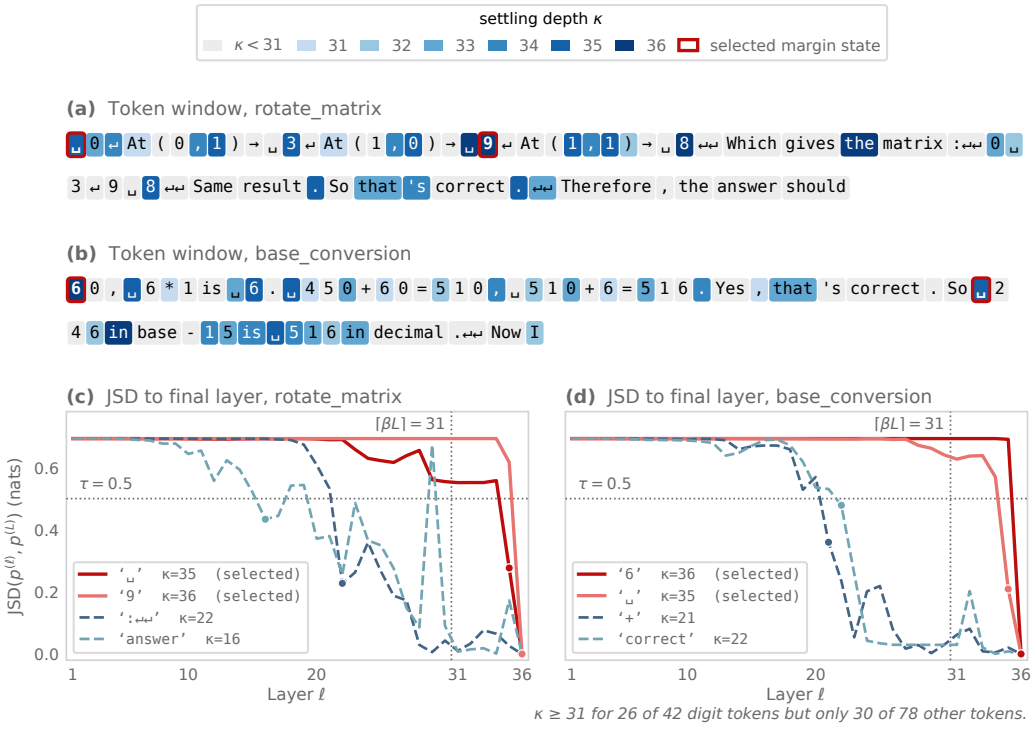


Figure 3: Held-out illustrations of deep-thinking token selection after `rotate_matrix` and `base_conversion` (canonical order, training seed 101). For each example, we re-apply the rule to one correct held-out response: among 64 evenly spaced candidate positions, select at most 8 with $\kappa \geq 31$, ranked by κ. These illustrate the rule on held-out states, separate from training-time consolidation. (a, b) Tokens colored by settling depth; boxes mark selected candidates. (c, d) Layer-to-final JSD for selected states and the two earliest-settling tokens; κ is the first crossing of $\tau = 0.5$.

For each margin state, the recorded token and its strongest endpoint competitor define the fixed-competitor logit gap in Eq. (3). The gradient is taken directly at ϕ_t, without a per-state normalizer. A margin observation is accepted only when its gap and gradient are finite, regardless of whether the recorded token is the endpoint argmax. A Fisher observation is accepted only when all score gradients required by its estimator are finite. Missing observations in one branch leave the other branch's count and memory unaffected. The next task is still centered at ϕ_t if neither branch accepts an observation. Boundary responses and prefix states are discarded once consolidation finishes; the final task requires no further consolidation.

A.3 Stratified Fisher estimation

For each sampled Fisher state $s_i \in \mathcal{J}_t^F$, let $A_i \subset \mathcal{V}$ be its two highest-probability next tokens under π_{ϕ_t}, and set $r_i = 1 - \sum_{a \in A_i} p_{i,a}$ for the remaining probability mass. DToP computes the two head

contributions exactly and estimates the tail from six conditional actions:

$$\widehat{f}_i = \sum_{a \in A_i} p_{i,a} (g^{\pi}_{i,a})^2 + \frac{r_i}{6} \sum_{k=1}^{6} (g^{\pi}_{i,b_{i,k}})^2, \qquad b_{i,k} \sim p_i(\cdot \mid a \notin A_i). \tag{15}$$

As in Eq. (7), vector squares are elementwise. In exact arithmetic the conditional tail expectation recovers the omitted mass, making $\widehat{f}_i$ unbiased for f_i. Uniform sampling of Fisher states from $\mathcal{R}_t$ then gives

$$\widehat{f}^{\,\mathrm{new}}_t = \frac{1}{|\mathcal{J}^F_t|} \sum_{i:\, s_i \in \mathcal{J}^F_t} \widehat{f}_i, \qquad \mathbb{E}[\widehat{f}^{\,\mathrm{new}}_t \mid \mathcal{P}_t, \mathcal{S}_t] = \frac{1}{|\mathcal{R}_t|} \sum_{i:\, s_i \in \mathcal{R}_t} f_i. \tag{16}$$

The implementation omits tail mass at most 10^{-12}, so exact unbiasedness applies before that cutoff.

A.4 Memory consolidation and margin compression

Let $\mathcal{I}^M_t \subseteq \mathcal{S}_t$ and $\mathcal{I}^F_t \subseteq \mathcal{J}^F_t$ be the accepted observations at boundary t. For each branch $A \in \{M, F\}$, let $n^A_t = |\mathcal{I}^A_t|$ and $N^A_{\le t} = N^A_{\le t-1} + n^A_t$. For nonempty branches, first form $\bar{C}_t = (n^M_t)^{-1} \sum_{i:\, s_i \in \mathcal{I}^M_t} g_i g_i^\top$ and $\bar{f}_t = (n^F_t)^{-1} \sum_{i:\, s_i \in \mathcal{I}^F_t} \widehat{f}_i$, respectively, and then update the memories as

$$f_{\le t} = \frac{N^F_{\le t-1} f_{\le t-1} + n^F_t \bar{f}_t}{N^F_{\le t}}, \tag{17}$$

$$C_{\le t} = \mathrm{Top}_r\left(\frac{N^M_{\le t-1} C_{\le t-1} + n^M_t \bar{C}_t}{N^M_{\le t}} \right). \tag{18}$$

If a branch receives no accepted observations, it retains its previous memory and count. The two separate counts prevent one missing branch from diluting the other. Margin gradients from earlier tasks are not recomputed: their retained information is already in $C_{\le t-1}$. The margin moment is uncentered, because subtracting its mean would eliminate a shared direction even if all observed decisions were sensitive to that direction.

Let

$$M_t = \frac{N^M_{\le t-1} C_{\le t-1} + \sum_{i:\, s_i \in \mathcal{I}^M_t} g_i g_i^\top}{N^M_{\le t}} \tag{19}$$

be the pre-truncation, positive-semidefinite matrix at boundary t. If its eigenvalues are $\sigma_1 \ge \cdots \ge \sigma_P \ge 0$, its best rank-$r$ spectral approximation is $C_{\le t} = \mathrm{Top}_r(M_t)$:

$$\|M_t - C_{\le t}\|_2 = \sigma_{r+1}, \qquad \|M_t - C_{\le t}\|_F^2 = \sum_{j>r} \sigma_j^2, \tag{20}$$

with

$$0 \le \delta^\top (M_t - C_{\le t}) \delta \le \sigma_{r+1} \|\delta\|_2^2. \tag{21}$$

This truncation retains the leading observed sensitivity directions. Proposition 2 accounts for truncation errors accumulated across boundaries.

Writing $C_{\le t-1} = U^\top \Lambda U$, form the stacked matrix

$$V_t = \begin{bmatrix} \sqrt{N^M_{\le t-1} / N^M_{\le t}}\, \Lambda^{1/2} U \\ (g_i^\top / \sqrt{N^M_{\le t}})_{i:\, s_i \in \mathcal{I}^M_t} \end{bmatrix}. \tag{22}$$

The top r right singular vectors of V_t, together with the squared singular values, give the exact incremental update in Eq. (18) for the previous retained memory and new gradients.

A.5 Projection, calibration, and missing observations

During task $t+1$, keep the center ϕ_t and the memories $C = C_{\le t}$ and $F = \mathrm{diag}(f_{\le t}) + \mu I_P$ fixed. For the first optimizer proposal $\Delta\varphi_{\mathrm{cal}}$ of each task with both memories available, calculate

$q_F = \Delta\varphi_{\text{cal}}^\top F \Delta\varphi_{\text{cal}}$ and $q_M = \Delta\varphi_{\text{cal}}^\top C \Delta\varphi_{\text{cal}}$. If both memories exist and these energies are finite and pass the positive-energy threshold, construct $G = F/q_F + \rho C/q_M$; the same proposal is then tested against the calibrated trust region. Calibration is performed once per task, even when the preceding boundary adds no new observations, and G is held fixed until the following boundary. The center is ϕ_t even when no new observations were accepted. Joint projection is enabled only when both memories are available and calibration succeeds; otherwise the task uses the base-optimizer proposals unchanged. A branch with no new observations provides its accumulated memory whenever one is available.

For an optimizer proposal $\Delta\varphi_k$, let $c_k = \varphi_k - \phi_t$, $b_k = c_k + \Delta\varphi_k$, and substitute $y = c_k + u$ into Eq. (12). Then

$$\min_y \frac{1}{2}\|y - b_k\|_2^2 \quad \text{subject to} \quad y^\top G y \le H^2. \tag{23}$$

The feasible set is convex, and the Euclidean objective is strictly convex. The KKT stationarity equation is $(I_P + \lambda G)y = b_k$. A feasible b_k uses $\lambda = 0$; otherwise $\lambda > 0$ is found by the scalar equation

$$Q(\lambda) = \sum_{j=1}^{P} \frac{\gamma_j \widetilde{b}_j^2}{(1 + \lambda\gamma_j)^2} = H^2, \tag{24}$$

where γ_j and $\widetilde{b}_j$ denote eigenvalues of G and coordinates of b_k in its eigenbasis. This produces Eq. (13).

The joint metric has diagonal-plus-low-rank form $G = D + U^\top \Gamma U$, where

$$D = \frac{\operatorname{diag}(f_{\le t} + \mu\mathbf{1})}{q_F}, \qquad \Gamma = \frac{\rho\Lambda}{q_M}. \tag{25}$$

Let $r_t \le r$ be the number of retained rows of U. For a candidate λ, set $A = I_P + \lambda D$ and $W = (\lambda\Gamma)^{1/2} U$. Woodbury's identity gives

$$(I_P + \lambda G)^{-1} b_k = A^{-1} b_k - A^{-1} W^\top (I_{r_t} + W A^{-1} W^\top)^{-1} W A^{-1} b_k. \tag{26}$$

Projection therefore uses diagonal operations, an $r_t \times r_t$ solve, and scalar search over λ. Persistent method state is $O(rP)$; a scalar-search evaluation costs $O(Pr^2 + r^3)$.

When calibration is valid, $\Delta\varphi_{\text{cal}}^\top G \Delta\varphi_{\text{cal}} = 1 + \rho$. The Fisher-only ablation uses $G = F/q_F$, and the margin-only ablation uses $G = C/q_M$. These are the operators evaluated in the component ablation; their results include the effect of each operator's calibration. Every task starts at its preceding endpoint, so $c_0 = 0$ even when neither branch accepted new observations. When joint projection is enabled, every iterate, including the initial one, lies in the region. Optimizer state and calibration reset between tasks; the memories persist. Projection changes the applied proposal, not the base optimizer's internal state.

A.6 Properties of the projected update

The projection and calibration procedure above has three useful properties. The first two hold for any positive-semidefinite metric, including the singular metric used by the margin-only ablation. The third concerns the joint metric under fixed calibration. The first two lemmas together prove Proposition 1.

Lemma 1 (Euclidean projection onto the trust region)**.** *Let $H > 0$, $G \succeq 0$, and $b_k = c_k + \Delta\varphi_k$. The projection in Equation (12) has a unique solution. If $b_k^\top G b_k \le H^2$, then $\Delta\varphi_k^* = \Delta\varphi_k$. Otherwise, there is a unique $\lambda^* > 0$ such that*

$$Q(\lambda^*) = H^2, \qquad Q(\lambda) = b_k^\top (I_P + \lambda G)^{-1} G (I_P + \lambda G)^{-1} b_k,$$

and $\Delta\varphi_k^* = (I_P + \lambda^* G)^{-1} b_k - c_k$. For $\lambda > 0$, $Q(\lambda) \le \|b_k\|_2^2/(4\lambda)$; consequently, the root is bracketed by $[0, \|b_k\|_2^2/(4H^2)]$.

Proof. The feasible set is nonempty, closed, and convex, and the Euclidean objective is strictly convex. Since $H > 0$, $u = -c_k$ is strictly feasible, so the KKT conditions characterize its unique minimizer. With the multiplier convention of Equation (13), stationarity gives

$$(I_P + \lambda G)(c_k + u) = b_k.$$

A feasible proposal is the minimizer with $\lambda = 0$; otherwise the constraint is tight and $\lambda > 0$. In an eigenbasis of G with eigenvalues $\gamma_j \geq 0$ and coordinates $\widetilde{b}_j$,

$$Q(\lambda) = \sum_j \frac{\gamma_j \widetilde{b}_j^2}{(1+\lambda\gamma_j)^2} \leq \frac{\|b_k\|_2^2}{4\lambda},$$

where the inequality follows from $v/(1+v)^2 \leq 1/4$ for $v \geq 0$. When the proposal is infeasible, $Q(0) > H^2 > 0$; hence Q is continuous and strictly decreasing to zero as $\lambda \to \infty$, and the stated bracket contains its unique root. The diagonal-plus-low-rank computation is specified in Appendix A.5. □

Lemma 2 (Cumulative displacement budget)**.** *Suppose the metric G and center ϕ_t remain fixed throughout task $t+1$, $\varphi_0 = \phi_t$, and every update is projected by Equation (12). Every iterate then satisfies $c_k^\top G c_k \leq H^2$, independently of the number of updates. In contrast, when G has a nonzero direction, a per-step bound $\Delta\varphi_k^\top G \Delta\varphi_k \leq H^2$ permits cumulative G-seminorm kH after k aligned steps.*

Proof. By feasibility of each projected solution, $c_{k+1} = c_k + \Delta\varphi_k^*$ satisfies $c_{k+1}^\top G c_{k+1} \leq H^2$. Since $\varphi_0 = \phi_t$, we have $c_0 = 0$, and induction gives the assertion for every iterate. For the comparison, choose any e with $e^\top G e > 0$ and take $\Delta\varphi_k = He/(e^\top G e)^{1/2}$ at each step. Then each step has G-seminorm H, while the displacement after k such steps has G-seminorm kH. □

Lemma 3 (Nesting under fixed calibration)**.** *Fix $F \succeq 0$, $C \succeq 0$, $q_F, q_M > 0$, and $H > 0$. For $0 \leq \rho \leq \rho'$, let $G_\rho = F/q_F + \rho C/q_M$. Then $G_\rho \preceq G_{\rho'}$, and*

$$\{c : c^\top G_{\rho'} c \leq H^2\} \subseteq \{c : c^\top G_\rho c \leq H^2\}.$$

In particular, $\rho = 0$ recovers the Fisher-only feasible set.

Proof. The difference $G_{\rho'} - G_\rho = (\rho' - \rho)C/q_M$ is positive semidefinite, which gives the inclusion for fixed memories and calibration. □

A.7 Sensitivity control and local implications

Proposition 2 (Control of stored sensitivity)**.** *Suppose both accumulated memories through boundary t are nonempty, calibration is valid ($q_F, q_M > 0$), $\rho > 0$, and projection is enabled during task $t+1$. Let $\widetilde{C}$ be the uncompressed second moment of all accepted historical margin gradients, $C = C_{\leq t}$ the retained approximation, $N^F = N^F_{\leq t}$, $N^M = N^M_{\leq t}$, and $\epsilon = \lambda_{\max}(\widetilde{C} - C) \geq 0$. For each iterate* after projection*, with $c_k = \varphi_k - \phi_t$,*

$$\frac{1}{N^F}\sum_i c_k^\top \operatorname{diag}(\widehat{f}_i) c_k \leq q_F H^2, \tag{27}$$

$$\frac{1}{N^M}\sum_i (g_i^\top c_k)^2 \leq \frac{q_M H^2}{\rho} + \epsilon\|c_k\|_2^2. \tag{28}$$

The sums refer to conceptual historical observations, not gradients or states retained individually in implementation.

Proof. Lemma 2 gives $c_k^\top G c_k \leq H^2$ for every iterate, including $c_0 = 0$. Since both summands of $G = F/q_F + \rho C/q_M$ are positive semidefinite,

$$c_k^\top F c_k \leq q_F H^2, \qquad c_k^\top C c_k \leq \frac{q_M H^2}{\rho}.$$

By the definition of the accumulated diagonal Fisher,

$$\frac{1}{N^F}\sum_i c_k^\top \operatorname{diag}(\widehat{f}_i) c_k = c_k^\top F c_k - \mu\|c_k\|_2^2 \leq q_F H^2,$$

proving Equation (27).

To account for repeated low-rank truncation, let $\widetilde{C}_{\le s}$ be the uncompressed historical second moment through boundary s, $C_{\le s}$ the retained memory, and M_s its pre-truncation matrix. With $E_s = M_s - \mathrm{Top}_r(M_s) \succeq 0$, the incremental update in Equation (18) yields

$$\widetilde{C}_{\le s} - C_{\le s} = \frac{N^M_{\le s-1}}{N^M_{\le s}}(\widetilde{C}_{\le s-1} - C_{\le s-1}) + E_s.$$

At the first nonempty margin boundary, the old weighted error is zero. For $N^M = N^M_{\le t}$, telescoping over boundaries with positive cumulative margin counts gives

$$0 \preceq \widetilde{C} - C = \sum_{s \le t} \frac{N^M_{\le s}}{N^M} E_s \preceq \epsilon I_P, \qquad \epsilon \le \sum_{s \le t} \frac{N^M_{\le s}}{N^M} \|E_s\|_2.$$

The equality $c_k^\top \widetilde{C} c_k = (N^M)^{-1} \sum_i (g_i^\top c_k)^2$ and the bound on $c_k^\top C c_k$ establish Equation (28). In particular, the first compression is exact when no more than r accepted margin gradients have been accumulated. The weighted sum of E_s accounts for truncation errors accumulated across later boundaries. □

Conditions for functional interpretation. The following conditions connect stored-sensitivity control to local changes in policy distributions and token preferences.

Assumption 1 (Local expansions at the previous endpoint)**.** *For each relevant historical state s_i:*

- ***(A1)*** *Policy KL has a second-order expansion at the previous endpoint ϕ_t with full Fisher $F_i^{\mathrm{full}}(\phi_t)$.*
- ***(A2)*** *The margin and margin gradient at ϕ_t match their stored values.*
- ***(A3)*** *The fixed-competitor margin admits a first-order expansion there.*
- ***(A4)*** *The stored competitor remains the strongest alternative along the displacement when interpreting a margin sign change as a decision change.*

A2 holds for observations collected at the current endpoint; for earlier observations it is an additional stability assumption.

Corollary 1 (Local Fisher–KL interpretation)**.** *Under A1, for a sufficiently small feasible displacement $c_k = \varphi_k - \phi_t$,*

$$\frac{1}{N^F} \sum_i \mathrm{KL}(\pi_{\phi_t}(\cdot \mid s_i) \,\|\, \pi_{\varphi_k}(\cdot \mid s_i)) \le \frac{1}{2} q_F H^2 + e_F(c_k) + o(\|c_k\|_2^2),$$

where

$$e_F(c_k) = \frac{1}{2N^F} \sum_i c_k^\top \left(F_i^{\mathrm{full}}(\phi_t) - \mathrm{diag}(\widehat{f}_i) \right) c_k.$$

The local KL bound includes the signed error e_F, which accounts for diagonalization, sampling, and differences between Fisher information at historical endpoints and at the current endpoint.

Proof. Expand each KL at ϕ_t, add and subtract $\mathrm{diag}(\widehat{f}_i)$, average the expansions, and apply Equation (27). □

Corollary 2 (Local changes in stored token preferences)**.** *Assume A2–A4. Fix $m_0 > 0$ and consider the stored margin states with $m_i(\phi_t) \ge m_0$. For their fixed-competitor logit gaps, define*

$$r_i(c) = m_i(\phi_t + c) - m_i(\phi_t) - g_i^\top c, \qquad \eta(c) = \max_{i:\, m_i(\phi_t) \ge m_0} |r_i(c)|.$$

If this subset is nonempty and $\eta(c_k) < m_0$, the fraction of all N^M stored margin states *with an initial margin at least m_0 whose recorded token is subsequently overtaken by its strongest alternative is at most*

$$\frac{q_M H^2/\rho + \epsilon \|c_k\|_2^2}{(m_0 - \eta(c_k))^2}.$$

For a finite collection of differentiable margins satisfying A2–A3, $\eta(c) = o(\|c\|_2)$ as $c \to 0$. The result characterizes local preference stability under A2–A4.

Proof. Under A4, an overtaking event implies $m_i(\phi_t + c_k) \le 0$. Since $m_i(\phi_t) \ge m_0$ and $m_i(\phi_t + c_k) = m_i(\phi_t) + g_i^\top c_k + r_i(c_k)$, we obtain $|g_i^\top c_k| \ge m_0 - \eta(c_k)$. Apply Markov's inequality to the mean squared projections of all stored margin gradients and use Eq. (28). □

Across-task displacement. Historical Fisher weights extend the endpoint-relative bounds to displacement accumulated across valid task boundaries.

Corollary 3 (Accumulated drift across task boundaries)**.** *Let t_0 be a boundary contributing $n_{t_0}^F > 0$ Fisher observations, with mean $f^{(t_0)}$. Suppose each subsequent task projects from its first update. Writing $\|w\|_{(t_0)}^2 = w^\top \operatorname{diag}(f^{(t_0)})w$, every projected iterate $\varphi_{T,k}$ during task $T > t_0$ satisfies*

$$\|\varphi_{T,k} - \phi_{t_0}\|_{(t_0)} \le H \sum_{t=t_0+1}^{T} \sqrt{\frac{N_{\le t-1}^F q_F^{(t)}}{n_{t_0}^F}}.$$

Proof. The running Fisher average satisfies $\operatorname{diag}(f_{\le t-1}) \succeq (n_{t_0}^F / N_{\le t-1}^F) \operatorname{diag}(f^{(t_0)})$. Applying Proposition 2's Fisher bound to each task controls that task's displacement in the fixed $\|\cdot\|_{(t_0)}$ seminorm. Summing across task boundaries and applying the triangle inequality proves the bound. □

Scope. These results characterize stored-state sensitivity and its local functional implications under the stated approximation and stability conditions. The across-task bound accumulates contributions from successive boundaries; retention of task-level accuracy is evaluated empirically.

A.8 Ablation definitions

Fisher-only removes the margin term, giving $G = F/q_F$; Margin-only removes the Fisher term, giving $G = C/q_M$. The entropy-selected variant replaces the DTT margin-state selector with a cardinality-matched lowest-entropy selector and samples Fisher states from the corresponding complement. The random variant draws the same number of margin states uniformly from the candidate pool and samples Fisher states from the resulting complement.

B Additional Experimental Details

B.1 Baseline implementation

All methods train a single LoRA adapter across the task stream. Sequential GRPO applies the base update without a continual-learning constraint. EWC adds to the GRPO objective a diagonal-Fisher quadratic penalty toward every previous task endpoint ($\lambda = 1$). Each task's Fisher is estimated from squared GRPO policy gradients collected during training on that task.

TreeLoRA was proposed for supervised continual fine-tuning. We follow its official implementation, which regularizes the adapter's A matrices toward earlier tasks chosen per block by a K-D tree and a lower-confidence-bound bandit. The original scales the regularizer to a fraction $0.5\,t/T$ of the supervised loss value. Since the GRPO surrogate is close to zero on-policy and changes sign, we apply the same fraction to the gradient norm instead.

CPR replaces part of each batch with stored prompts from earlier tasks and generates fresh responses for the same GRPO update. It retains observed training prompts and verifier metadata. MTRL interleaves all tasks from the start, with equal prompt counts per task and the same total number of updates as a continual stream. Its results therefore do not depend on the continual task order.

B.2 Training and method configuration

Table 5 collects the training and DToP settings. All methods use AdamW (Loshchilov & Hutter, 2019); CPR replaces part of the fixed prompt budget with stored prompts. Method hyperparameters are unchanged across tasks unless a stated ablation varies them.

Table 5: Training and DToP configuration. The ten-task setting differs in backbone, task set and order, and updates per task.

Quantity	Value	Description
Optimization		
Backbone	Qwen3-8B	Main six-task setting
Backbone	Qwen3-4B	Ten-task setting
LoRA rank / α	16 / 32	Trainable adapter configuration
Optimizer	AdamW	Applied to GRPO updates
Learning rate	10^{-5}	Fixed across tasks
Prompts / update	8	Total prompts per GRPO update
Responses / prompt	5	Sampled responses per prompt
Updates / task	40 / 50	Six-task / ten-task settings
DToP		
K	8	Maximum margin states per boundary
τ	0.5 nats	Settling-depth JSD threshold
β	0.85	Relative-depth threshold
K_F	8	Maximum Fisher states per boundary
r	8	Rank of the compressed margin memory
ρ	0.5	Additive margin weight
μ	10^{-8}	Fisher damping
H	5.0	Trust-region horizon
Head / tail actions	2 / 6	Exact head actions / conditional tail samples
Candidate positions / responses	64 / 8	Maximum positions / boundary responses
Energy threshold	10^{-20}	Minimum positive branch energy
Bisection steps	40	Maximum KKT scalar-search iterations

B.3 Additional component controls

The Fisher-only and Margin-only variants correspond to removing, respectively, the margin or Fisher contribution from the joint metric in Eq. 11. Both retain the base adaptation protocol.

For fixed memories and calibration, the joint region is contained in the Fisher-only region at the same H (Lemma 3). We also compare the methods at a matched calibration-proposal budget: joint calibration gives energy $1+\rho$, while Fisher-only gives energy 1. Using $H/\sqrt{1+\rho} = 5/\sqrt{1.5} \approx 4.08$ for Fisher-only therefore matches the fraction of the budget used by its calibration proposal. On canonical-order seed 101, DToP achieves 69.0% Final accuracy, compared with 64.6% for this control (Table 6). The advantage thus persists in this single-seed comparison at a matched calibration-proposal budget.

Table 6: Fisher-only control with a matched calibration-proposal budget, canonical-order Qwen3-8B, training seed 101. Final and Acq are accuracies (%); BWT is in percentage points.

Variant	Final	Acq	BWT
Fisher-only ($H = 4.08$)	64.6	71.1	-7.8

B.4 Per-task acquisition and final performance

Table 7 separates acquisition and retention by reporting each task's post-acquisition and final accuracy for training seed 101 in each order. In the reverse order, the gap between DToP and SEQ comes from both terms: after the whole stream DToP keeps `rotate_matrix` at 89.1%

against 34.4%, and right after training it also reaches higher accuracy on `rotate_matrix` and `binary_alternation`.

Table 7: Per-task accuracy (%) on the six-task Qwen3-8B stream, training seed 101. $R_{j,j}$: accuracy on task j right after training it; $R_{T,j}$: after the whole stream. Tasks are listed in training order; for the last task $R_{j,j} = R_{T,j}$.

	SEQ		DToP	
Task	$R_{j,j}$	$R_{T,j}$	$R_{j,j}$	$R_{T,j}$
Canonical order				
`base_conversion`	100.0	100.0	100.0	100.0
`binary_alternation`	42.2	26.6	56.3	50.0
`rotate_matrix`	89.1	81.3	92.2	84.4
`caesar_cipher`	10.9	18.8	12.5	6.3
`isomorphic_strings`	96.9	95.3	98.4	98.4
`string_manipulation`	71.9	71.9	75.0	75.0
Reverse order				
`string_manipulation`	75.0	48.4	73.4	68.8
`isomorphic_strings`	96.9	90.6	98.4	93.8
`caesar_cipher`	17.2	1.6	17.2	14.1
`rotate_matrix`	68.8	34.4	95.3	89.1
`binary_alternation`	14.1	7.8	43.8	56.3
`base_conversion`	95.3	95.3	100.0	100.0

B.5 STATE-SELECTION AND GENERATION DIAGNOSTICS

At matched cardinality, DTT and lowest entropy select different decisions (Table 8, training seed 101). DTT selects states that settle at layers 34–36 and have smaller endpoint margins; lowest entropy selects states at layers 23–34, mostly below the DTT threshold of 31, with larger margins. These diagnostics characterize the state partition induced by each selector, including the complementary Fisher pool and its subsequent calibration.

Table 8: Margin states selected at the five non-final boundaries of the canonical six-task Qwen3-8B stream, training seed 101 (8 per boundary, 40 in total). κ is the settling depth over the $L = 36$ layers; margins are endpoint logit gaps between the recorded token and its strongest competitor.

Margin states	Mean κ (range)	Median margin (min)
DTT (DToP)	35.0 (34–36)	19.9 (1.8)
Lowest entropy	28.8 (23–34)	36.9 (30.5)

After the reverse-order stream, 9.1% of DToP's evaluation responses reach the 4096-token cap, compared with 39.6% for SEQ. Table 9 reports this generation diagnostic after the first and final tasks for training seed 101.

Table 9: Share of the 384 reverse-order evaluation responses that reach the 4096-token response cap, training seed 101.

Responses at the length cap (%), reverse order	SEQ	DToP
After task 1	4.4	7.3
After task T	39.6	9.1

B.6 TRUST-REGION AND OPTIMIZATION DIAGNOSTICS

The projection in Eq. (12) changes a proposal when its cumulative displacement exceeds the trust-region budget. Table 10 reports this behavior for training seed 101. In both orders, 76.4% of

eligible updates require projection; the median correction on projected updates is 4.7–5.5% of the proposal norm. In each task after the first, the constraint becomes active at update 7–16 of 40, after which every update is projected. Thus, cumulative control preserves a feasible displacement through repeated corrections whose median magnitude is small relative to the base proposal. The applied displacement has median margin-energy fractions of 0.095 and 0.031, compared with $\rho/(1+\rho) = 1/3$ for the calibration proposal, indicating reduced relative energy along stored margin directions.

Table 10: Optimization diagnostics on the six-task Qwen3-8B stream, training seed 101, over the 199 eligible updates (those where projection is possible) in tasks 2–6 of each order. Requested energy is $b_k^\top G b_k / H^2$; the correction median uses projected updates. The margin-energy fraction is the ρ-weighted margin contribution divided by total normalized energy of the applied displacement.

Diagnostic	Canonical	Reverse
Eligible updates requiring projection (%)	76.4	76.4
Median requested energy / budget	1.05	1.04
Median $\|\Delta\varphi_k^* - \Delta\varphi_k\|_2 / \|\Delta\varphi_k\|_2$ (%)	5.5	4.7
Median margin-energy fraction	0.095	0.031

B.7 Ten-task evaluation

The longer setting uses Qwen3-4B and 50 updates per task. The canonical task sequence is:

`base_conversion` → `binary_alternation` → `binary_matrix` → `caesar_cipher` → `cryptarithm` → `isomorphic_strings` → `jugs` → `rotate_matrix` → `string_manipulation` → `ab`.

The evaluation contains 606 fixed examples: 64 for each task except `jugs`, which has 30. Final and acquisition metrics are macro-averaged over the ten tasks.

B.8 Efficiency and compute

On the same GPU, DToP adds 5.8% to six-task training time and 4.0% to time per optimizer update; boundary sensitivity estimation and consolidation take 4.0 minutes over the stream (Table 11). Projection has median relative corrections of 4.7–5.5% on projected updates (Appendix B.6). Additional persistent method state totals 1.05 GB: a Fisher diagonal (0.175 GB, FP32), a rank-8 margin basis (0.698 GB, BF16), and the fixed trust-region center ϕ_t (0.175 GB, FP32). These are stored in addition to the adapter and frozen backbone. Inference uses the ordinary backbone and adapter.

Table 11: Efficiency on one NVIDIA H100 80GB GPU, one canonical-order six-task Qwen3-8B run per method (240 optimizer updates each; DToP: training seed 101).

Metric	SEQ	DToP
Six-task training time (h)	5.02	5.31
Time per optimizer update (s)	74.3	77.3
Boundary computation per stream (min)	–	4.0
Additional persistent method state (GB)	–	1.05