Structure Over Scale: Rethinking Adaptation for Reinforcement Learning with Verifiable Rewards

Allan Kazakov [1] **Abdurrahman Javat** [1]

Abstract

The standard justification for Full Fine-Tuning (FFT) in Reinforcement Learning with Verifiable Rewards (RLVR) rests on a reasonable intuition: reasoning requires expressive weight updates that Low-Rank Adaptation (LoRA) cannot provide. We show this intuition identifies the wrong variable. Through a systematic rank sweep under GRPO, we document *rank collapse*—a discontinuous performance cliff where increasing LoRA rank beyond a threshold causes catastrophic, irrecoverable policy failure on moderate batch sizes, a phenomenon absent from the SFT literature. Spectral analysis reveals the mechanism: in a sparse binary reward landscape, unconstrained high-rank adapters allow the optimizer to satisfy rewards through degenerate solutions, bypassing coherent reasoning entirely. FFT exhibits the same pathology in milder form—achieving *lower* effective rank in its learned weight updates than structured adapters using less than 0.6% of the parameters. Expressivity is not the bottleneck; structure is. Structured adapters that constrain *which* high-rank solutions are reachable by gradient descent consistently outperform both LoRA and FFT, and do so more sharply as base-model pre-training scale increases—a pattern we term the *Model Maturity Hypothesis*, supported by behavioral replication across three architecturally independent models and by spectral signatures in frozen base weights that predict adaptation behavior before training begins. The operative question for RLVR is not whether to use LoRA or FFT, but what structure to impose over the update manifold.

[1]Department of Artificial Intelligence, Bahçeşehir University, Istanbul, Türkiye. Correspondence to: Allan Kazakov <allan.kazakov@bahcesehir.edu.tr>, Abdurrahman Javat <abdurrahman.javat@bahcesehir.edu.tr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

1.1. The RLVR Efficiency Problem

The dominant recipe for training reasoning-capable large language models (LLMs) pairs a base model pre-trained on trillions of tokens with a Reinforcement Learning with Verifiable Rewards (RLVR) post-training stage (DeepSeek-AI, 2025). Within this stage, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has emerged as the standard method for reasoning: it eliminates a separate value-function critic by normalizing rewards within a group of sampled rollouts, halving the memory footprint compared with PPO.

Despite GRPO's efficiency, the field faces a persistent dichotomy when choosing how many model parameters to update. Leading laboratories default to *Full Fine-Tuning* (FFT) (DeepSeek-AI, 2025), accepting massive memory and gradient-synchronization overhead under the assumption that reasoning updates require high intrinsic dimensionality. Academic work instead adopts *Low-Rank Adaptation* (LoRA) (Hu et al., 2021), citing evidence that rank $r{=}16$ achieves performance parity with FFT when hyperparameters are carefully optimized (Schulman & Lab, 2025). The field has thus settled into a binary choice: expensive correctness or affordable approximation.

We present experimental evidence hypothesizing that *both* assumptions underlying this binary are incorrect, and that the resolution lies in structured high-rank tensor adaptation.

1.2. Two Surprising Findings

Finding 1 — Rank Collapse in RLVR. The SFT literature predicts that increasing LoRA rank yields diminishing returns (Hu et al., 2021; Albert et al., 2025). Our LoRA rank sweep on Qwen 3 8B trained with GRPO on DeepMath-Hard reveals a qualitatively different phenomenon:

Nominal rank r	8	16	32	64	128	256
Accuracy	78.1%	76.2%	77.3%	73.1%	**4.7%**	**2.3%**

Performance peaks at $r{=}8$ and *collapses catastrophically* above $r{=}64$ on moderate batch sizes. SVD analysis (Figure 2) reveals the mechanism: despite their high *nominal*

Table 1. Performance split across three base models. All results on DeepMath-Hard (Seed 42). QuanTA is competitive on ∼15T-token models but is the clear winner on the 36T-token model.

Model	Pre-training	Best LoRA	Best QuanTA	Δ
Apertus 8B	∼15T	2.3%	3.5%	+1.2%
Llama 3.1 8B	∼15.6T	2.7%	3.9%	+1.2%
Qwen 3 8B	∼36T	78.1%	84.0%	+5.9%

rank, the r=128 and r=256 checkpoints achieve near-zero *effective rank* (participation ratio)—the RL optimizer concentrates gradient energy into fewer than ten singular directions while ignoring all remaining degrees of freedom. Unconstrained high-rank parameters therefore provide destructive flexibility rather than beneficial expressivity in the RLVR regime. This is not predicted by, nor analogous to, any known SFT behavior.

Finding 2 — The Model Maturity Hypothesis. Alongside rank collapse, we observe a systematic divergence in which adapter family performs best, and this divergence correlates cleanly with base-model pre-training scale. Table 1 summarizes the pattern across three architecturally independent models.

The two ∼15T models not only agree in direction but also exhibit a qualitative signature absent from Qwen 3. We attribute this behavioral bifurcation to pre-training scale: a model trained on 36T tokens has converged toward flatter singular value spectra in its frozen weights, meaning that activating its latent reasoning circuits requires updates distributed across many singular directions—precisely what QuanTA's Matrix Product Operator (MPO) structure provides and what LoRA's rank constraint prevents. We present this as a well-supported hypothesis rather than an established causal law: the two groups are cleanly separated, but architecture, tokenizer, and data mixture co-vary with pre-training scale and cannot be fully disentangled with the current model set.

1.3. Why Structure, Not Rank Alone

Both findings point to the same underlying principle. Rank collapse demonstrates that *unconstrained* high-rank updates are harmful in RLVR. The model maturity effect demonstrates that *constrained low-rank* updates are also inadequate for sufficiently mature models. The resolution is structured high-rank.

1.4. Contributions

This paper makes five concrete contributions:

1. **Rank Collapse in RLVR.** The first systematic documentation of catastrophic policy collapse at LoRA $r \geq 128$ in the GRPO framework on moderate batch sizes, with a spectral mechanistic explanation grounded in effective-rank diagnostics—a phenomenon qualitatively distinct from the SFT literature's prediction of diminishing returns.
2. **The Full Fine-Tuning Effective-Rank Paradox.** Evidence that FFT achieves lower effective rank in learned ΔW matrices than structured adapters using less than 0.6% of the parameters, demonstrating that structural inductive bias—not parameter count—is the operative factor in the RLVR regime.
3. **Comprehensive Empirical Study with Spectral Diagnostics.** The first head-to-head comparison of LoRA (rank sweep $r \in \{8, 16, 32, 64, 128, 256\}$), DoRA, QuanTA, and FFT under a cold-start GRPO pipeline, with multi-seed validation and SVD-based spectral analysis.
4. **The Model Maturity Hypothesis.** A well-supported, falsifiable hypothesis linking base-model pre-training scale to optimal adaptation strategy, substantiated by behavioral replication across three architecturally independent models spanning ∼15T to 36T pre-training tokens and by a pre-training spectral signature that predicts adaptation behavior before any fine-tuning occurs.
5. **Practical Guidance for Efficient RLVR.** We show that the industry FFT-vs-LoRA binary is a false dichotomy, that vanilla LoRA above $r = 32$ is unsafe in cold-start GRPO, and that cosine scheduling with warmup is recommended at moderate batch sizes—with constant schedules causing immediate policy collapse regardless of adapter choice.

2. Background and Related Work

2.1. Reinforcement Learning with Verifiable Rewards

RLVR uses ground-truth outcomes—such as mathematically verified answers—as the reward signal, avoiding the over-optimization risks of learned reward models (DeepSeek-AI, 2025). Shao et al. (2024) introduced Group Relative Policy Optimization (GRPO), which eliminates the memory cost of a separate critic by computing per-output advantages within a group of G completions sampled from the current policy:

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G) + \epsilon}, \tag{1}$$

where $r_i \in \{0, 1\}$ is the binary correctness reward for the i-th completion. The resulting high variance in advantage estimates—inherent to sparse binary rewards—is a defining property of the RLVR optimization landscape and, as we

show in Section 4, a key driver of rank collapse in unconstrained adapters. We operate in the cold-start (Base-to-RL) regime (Liu et al., 2025), in which the model must simultaneously learn output structure and reasoning logic from the reward signal alone, without any SFT warmup. We adopt the Dr. GRPO token-level loss normalization (Liu et al., 2025) to prevent penalizing longer Chain-of-Thought completions.

2.2. Parameter-Efficient Fine-Tuning

LoRA. Hu et al. (2021) freezes the pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ and inject a trainable low-rank update $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$:

$$W = W_0 + \Delta W, \qquad \Delta W = BA. \tag{2}$$

The method is grounded in the Intrinsic Dimension Hypothesis (Aghajanyan et al., 2020), which posits that downstream adaptation lies in a low-dimensional subspace—a characterization established in SFT settings whose validity in RLVR our experiments directly challenge.

DoRA. Liu et al. (2024) decompose each weight matrix into a magnitude scalar and a directional component, applying a LoRA update only to the direction:

$$W = \|W_0\|_c \cdot \frac{W_0 + \Delta W}{\|W_0 + \Delta W\|_c}, \tag{3}$$

where $\|\cdot\|_c$ denotes column-wise norm. By forcing the optimizer to treat magnitude and direction as independent degrees of freedom, this decomposition acts as a structural constraint on the update manifold—one that, as our spectral analysis shows, produces substantially higher effective rank than LoRA at identical nominal rank. DoRA is therefore not simply a LoRA variant with better hyperparameters: it is a structurally constrained adapter whose inductive bias places it in the same mechanistic category as QuanTA, achieved via a different parameterization.

QuanTA. Chen et al. (2024) parameterize ΔW as a Matrix Product Operator (MPO), a tensor network structure adapted from quantum many-body physics. For a weight matrix with reshaped local indices $\mathbf{i} = (i_1, \ldots, i_n)$ and $\mathbf{j} = (j_1, \ldots, j_n)$, the update takes the form:

$$\Delta W(\mathbf{i}, \mathbf{j}) = \sum_{\boldsymbol{\alpha}} \prod_{k=1}^{n} C^{(k)}_{\alpha_{k-1}\alpha_k}(i_k, j_k), \tag{4}$$

where $C^{(k)} \in \mathbb{R}^{\chi\times\chi\times d_k\times d_k}$ are learnable core tensors and χ is the bond dimension. The MPO structure decouples adaptation rank from parameter count: ΔW can approximate a full-rank transformation while parameters scale linearly in the number of modes n. We evaluate two configurations: a parameter efficient setting ($d = 4$, decomposition $[16, 8, 8, 4]$, 7.7M parameters) and a high-capacity setting ($d = 3$, decomposition $[16, 16, 16]$, 35.3M parameters). BOFT (Liu et al., 2023) and MoRA (Jiang et al., 2024) impose alternative structural constraints; we defer a detailed theoretical, not empirical, comparison to Appendix J.

2.3. Related Work

Rank and expressivity in SFT. The SFT literature consistently finds that increasing LoRA rank yields diminishing returns, with performance plateauing well below the full-rank regime (Hu et al., 2021; Liu et al., 2024)—a pattern explained by the Intrinsic Dimension Hypothesis (Aghajanyan et al., 2020). Our findings establish a qualitative distinction: in RLVR, higher rank without structural constraint does not yield diminishing returns, it yields catastrophic policy collapse.

Stability and scheduling in RLVR. Schulman & Lab (2025) report that Base-to-RL fine-tuning with LoRA $r = 16$ achieves parity with FFT under optimized hyperparameters, and that constant learning-rate schedules are stable at large batch sizes. We show that neither finding generalizes to resource-constrained regimes: at a global batch size of 64, constant schedules cause immediate policy collapse, and LoRA parity with FFT depends critically on which structured alternative is used as the comparison point. Notably, their setup reports no rank collapse at high LoRA rank—a discrepancy we attribute to batch size and address in Section 4.1.

PEFT for RL reasoning. Liu et al. (2025) analyze cold-start RLVR failure modes and motivate the Dr. GRPO loss correction we adopt; DeepSeek-AI (2025) demonstrate strong reasoning via RLVR but rely on FFT throughout. To our knowledge, no prior work has conducted a systematic comparison of structured high-rank adapters against LoRA and FFT in the same RLVR pipeline, nor provided the spectral mechanistic analysis we present here.

3. Methodology

3.1. Base Models and Pre-training Scale

We study three dense transformer models spanning a wide range of pre-training scale (Table 2), using the *base* checkpoint throughout. Base checkpoints ensure the model must learn both output formatting and reasoning logic from the reward signal alone—the cold-start regime in which adapter expressivity is most likely to be a binding constraint.

Table 2. **Base models evaluated.** All models are dense transformers at the 8B parameter scale; pre-training token counts are taken from the respective technical reports.

Model	Family	Pre-training Scale	Vocab Size
Apertus 8B	Independent	∼15T tokens	131K
Llama 3.1 8B	Meta	∼15.6T tokens	128K
Qwen 3 8B	Alibaba	∼36T tokens	150K

3.2. Adapter Configurations

Table 3 summarizes all configurations evaluated, applied to the same set of linear projection modules in `bfloat16` throughout. The LoRA rank sweep ($r \in \{8, 16, 32, 64, 128, 256\}$) and DoRA rank sweep ($r \in \{8, 16, 32, 64\}$) are conducted on Qwen 3 8B only; FFT is similarly restricted to Qwen 3 8B due to compute constraints. We additionally evaluate QuanTA with target-module ablations (excluding key/value projections and excluding only the value projection); full ablation results are in Appendix F.

3.3. Benchmark Selection and Data Curation

We train and evaluate on five benchmarks; full curation details are in Appendix B. **DeepMath-Hard** applies a difficulty filter ($\geq 8.5/10$) to DeepMath-103K (He et al., 2025), yielding 5,399 hardest problems concentrated in abstract mathematics; this is our primary discriminative benchmark. **Skywork-Hard** retains the 6,702 problems from Skywork-OR1 (Zeng et al., 2024) on which DeepSeek-R1-Distill-Qwen-32B achieves a pass rate below 19% (Score ≥ 13), providing a frontier stress test. **Enigmata** (Chen et al., 2025) is a logic and puzzle reasoning benchmark. We draw 10,000 problems while preserving the dataset's original difficulty distribution, and evaluate the model on the official published evaluation set. **MATH** (Hendrycks et al., 2021) serves as a standard reference benchmark. We additionally probe out-of-distribution generalization by evaluating DeepMath-Hard-trained checkpoints on the merged **AIME 2025–2026** (Art of Problem Solving, 2025a;b; 2026a;b) competition set (60 problems, Avg@32).

3.4. Reinforcement Learning Protocol

We train all models with GRPO (Shao et al., 2024) using group size $G = 8$, KL coefficient $\beta = 0$, the Dr. GRPO token-level loss (Liu et al., 2025), and a global batch size of 64; vLLM (Kwon et al., 2023) accelerates generation via PagedAttention. The reward is a composite $R_{\text{total}} = R_{\text{acc}} + R_{\text{fmt}}$: R_{acc} uses depth-aware `\boxed{}` extraction with `math_verify` semantic equivalence (1.0 correct, 0.0 otherwise), and R_{fmt} applies shaped penalties for structural degeneration (empty reasoning chains, repetition, hallucinated turns), bounded below R_{acc} to prevent reward hacking. Full reward weights and hyperparameters are in Appendix A.

Scheduling and cold-start stability. Schulman & Lab (2025) report that constant learning-rate schedules are stable in large-batch Base-to-RL training. We find this does *not* generalize to our regime (Appendix C): at batch size 64, constant schedules cause immediate policy collapse across all methods, which we attribute to the elevated variance of GRPO advantage estimates at small batch sizes. A **cosine decay schedule with warmup ratio 0.1** and minimum learning rate ratio 0.15 is suggested for stable cold-start GRPO at moderate batch sizes, regardless of adapter choice.

4. Experiments and Results

4.1. Rank Collapse: The LoRA Rank Sweep in RLVR

Table 4 presents a systematic LoRA rank sweep on Qwen 3 8B trained on DeepMath-Hard. Performance peaks at $r = 8$ (78.1%) and collapses catastrophically between $r = 64$ (73.1%) and $r = 128$ (4.7%); at $r = 256$, the policy learns nothing (2.3%). This is qualitatively distinct from the SFT literature's prediction of diminishing returns: the RL optimizer does not plateau at high rank, it finds degenerate solutions that satisfy the binary reward signal without learning a coherent reasoning policy. We provide the spectral mechanistic explanation in Section 4.4.

Batch size as a moderating factor. Schulman & Lab (2025) report no rank collapse at $r = 128$ using the same base model, which appears to contradict our finding. We attribute this to batch size: at large batch sizes, GRPO advantage estimates are averaged over substantially more rollouts per update, reducing the variance that drives degenerate energy concentration in unconstrained high-rank adapters. At global batch size 64, this variance is high enough to make unconstrained high-rank freedom actively harmful. Rank collapse is therefore not a universal property of RLVR but a batch-size-conditional phenomenon.

4.2. Main Comparison: LoRA, DoRA, QuanTA, and Full Fine-Tuning

Table 5 presents results for all primary methods on Qwen 3 8B (seed 42); multi-seed DeepMath-Hard and Skywork-Hard results are in Appendix E.

Three findings stand out. **(i)** Both QuanTA $d = 3$ and DoRA $r = 16$ reach **84.0%** on DeepMath-Hard, outperforming FFT by 6.7 percentage points and LoRA $r = 8$ by 5.9 points—with QuanTA using 22% fewer parameters than DoRA (35.3M vs. 45.0M) and leading on Skywork-Hard (19.4% vs. 17.0%), where the inversion of the DoRA-QuanTA ordering is consistent with QuanTA's higher ef-

Table 3. **Adapter configurations.** Trainable parameter counts reported for Qwen 3 8B. "All linear"† denotes all attention projection and feed-forward linear modules with KV excluded.

Method	Configuration	Target Modules	Trainable Params
LoRA	$r \in \{8, 16, 32, 64, 128, 256\}$	All linear	22M–∼700M
DoRA	$r \in \{8, 16, 32, 64\}$	All linear	23M–180M
QuanTA $d = 4$	$[16, 8, 8, 4]$	All linear†	7.7M
QuanTA $d = 3$	$[16, 16, 16]$	All linear†	35.3M
FFT	Full model	All params	∼8B

Table 4. **LoRA rank sweep on Qwen 3 8B, DeepMath-Hard (seed 42).** Performance peaks at $r = 8$ and collapses at $r \geq 128$, a discontinuous cliff with no analogue in the SFT literature.

Method	Nominal Rank	Trainable Params	DeepMath-Hard
LoRA	$r = 8$	22M	78.1%
LoRA	$r = 16$	43.6M	76.2%
LoRA	$r = 32$	87.2M	77.3%
LoRA	$r = 64$	174M	73.1%
LoRA	$r = 128$	348M	4.7%
LoRA	$r = 256$	696M	2.3%

fective rank providing a clearer advantage as task difficulty increases. That two structurally distinct adapters converge to identical performance on DeepMath-Hard is itself evidence that the operative variable is the structural constraint they share, not any property specific to either parameterization. **(ii)** FFT scores 77.3% on DeepMath-Hard, below both structured adapters and LoRA $r = 8$ despite ∼200× more trainable parameters; the spectral explanation is in Section 4.4. On MATH, all methods saturate near 78.9% with the exception of DoRA $r = 16$ (75.6%), given that multi-seed results are unavailable for this configuration on MATH, we do not draw conclusions from this single-seed anomaly. Within vanilla LoRA, $r = 8$ consistently outperforms $r = 16$ across both MATH and DeepMath-Hard, confirming that the rank collapse dynamic depresses performance even below the catastrophic threshold. Multi-seed averages (seeds 42–44) show meaningful run-to-run variance: DoRA $r = 16$ achieves 74.4% ± 10.9% and QuanTA $d = 3$ no_kv achieves 75.1% ± 9.4% on DeepMath-Hard; structured adapters consistently outperform LoRA and FFT on average despite this (Appendix I). Out-of-distribution generalization to AIME 2025–2026 (Avg@32) follows the same structured-adapter advantage—DoRA $r = 16$ (11.0%) > QuanTA $d = 3$ (10.4%) > LoRA $r = 16$ (10.0%)—consistent with higher effective rank producing reasoning circuits less brittle to novel problem phrasings; full fine-tuning results are deferred to Appendix G.

4.3. The Model Maturity Hypothesis

Table 1 presents results across all three base models. Both ∼15T models score near floor on DeepMath-Hard under any adapter configuration; Qwen 3 8B reaches 78–84%. The behavioral contrast extends beyond absolute accuracy: on MATH, Llama 3.1 8B exhibits seed-level instability under QuanTA $d = 4$ absent from LoRA—seed 43 collapses entirely (0.8%), while LoRA $r = 16$ is stable across all three seeds (30.2% ± 3.7%); full multi-seed results are in Appendix E. Qwen 3 8B shows no such instability under any structured adapter on any benchmark.

Figure 1 offers a structural explanation that precedes any fine-tuning. The singular value decay of frozen linear layers reveals a monotonic ordering: both ∼15T models decay steeply and nearly identically despite differing in tokenizer, hidden activation function, and training data mixture, while Qwen 3 8B's curve sits markedly flatter throughout. That two architecturally independent models converge to the same spectral signature at the same pre-training data volume rules out any single family-specific confound and implicates data volume as the primary driver. The proposed mechanism follows directly: a model with flat singular value spectra encodes information more uniformly across its weight space; activating its latent reasoning circuits requires updates distributed across many singular directions simultaneously, which low-rank adapters cannot provide but structured high-rank adapters can. More details can be found in Appendix H.

4.4. Spectral Analysis: Why Structure Beats Scale

Figure 2 reports the mean effective rank (participation ratio $\rho = \exp(H(\mathbf{p}))$, where $\mathbf{p}$ is the distribution over squared singular values of ΔW and H is Shannon entropy) averaged across all target layers for every method.

Three findings emerge. **(i) Rank collapse is visible at the**

Table 5. **Main results on Qwen 3 8B (seed 42).** †QuanTA $d = 3$ uses the `no_kv` target-module configuration; see Appendix F for full ablation. Best result per benchmark is **bolded**.

Method	Config	Params	MATH	DeepMath-Hard	Skywork-Hard
LoRA	$r = 8$	22M	**78.9%**	78.1%	19.0%
LoRA	$r = 16$	43.6M	78.1%	76.2%	15.2%
DoRA	$r = 8$	23.2M	78.1%	81.9%	16.9%
DoRA	$r = 16$	45.0M	75.6%	**84.0%**	17.0%
QuanTA†	$d = 3$	35.3M	**78.9%**	**84.0%**	**19.4%**
FFT	—	∼8B	**78.9%**	77.3%	15.9%

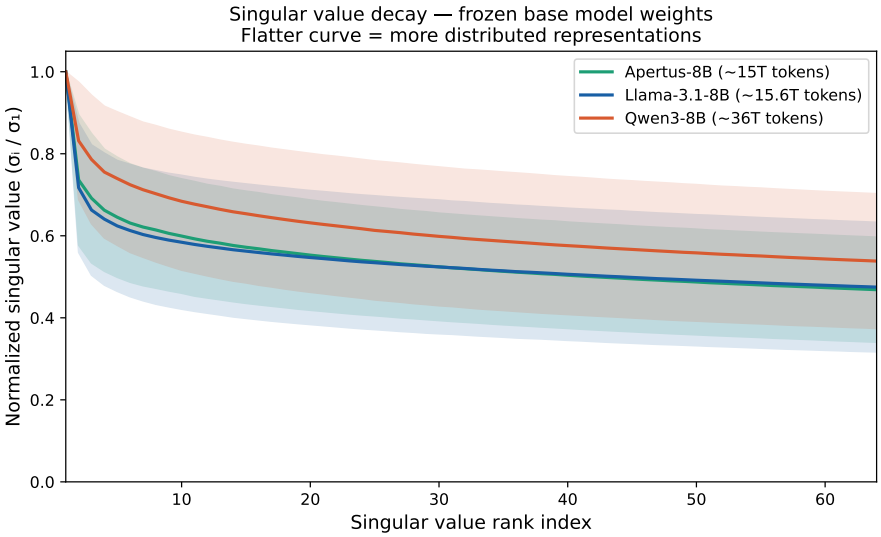

Figure 1. **Singular value decay of frozen base model weights, averaged across sampled linear layers using the top 64 singular values per layer.** Shaded regions denote one standard deviation across layers. The two ∼15T models decay steeply and overlap almost exactly despite differing architectures, ruling out family-specific confounds. Qwen 3 8B's markedly flatter curve indicates more distributed pre-trained representations—and predicts, before any gradient step, that structured high-rank adaptation will be required to activate its latent reasoning circuits.

weight level: collapsed LoRA runs ($r \geq 128$) achieve near-zero effective rank despite high nominal rank, confirming that the RL optimizer concentrates all update energy into a handful of singular directions when given unconstrained high-rank degrees of freedom. **(ii) Structure, not scale, determines effective rank**: FFT with ∼8B parameters achieves mean effective rank ∼400—substantially lower than DoRA $r = 16$ (∼900) and QuanTA $d = 3$ (∼1580) at less than 0.6% of the parameters; unconstrained optimization actively concentrates gradient energy rather than distributing it. **(iii) Structural constraint is the unifying principle**: DoRA's magnitude-direction decomposition and QuanTA's MPO tensor network are mechanistically distinct, yet both achieve high effective rank. This convergence across two independent structural forms constitutes the strongest available evidence that effective rank, produced by structural inductive bias, is the operative variable in RLVR adaptation—not parameter count, nominal rank, or any property specific to tensor decomposition. The performance gap between DoRA and QuanTA opens on Enigmata (Section 4.5), where the less binding ceiling reveals QuanTA's higher effective rank (∼1580 vs. ∼900) as an advantage when task difficulty demands it.

4.5. Domain Generalization: Enigmata

To assess whether the advantage of structured adaptation extends beyond mathematics, we train all primary methods on Enigmata 10K under the same GRPO protocol and evaluate on the held-out Enigmata evaluation set (Table 6).

The held-out ordering—QuanTA (20.6%) > FFT (19.1%) > DoRA (16.6%) > LoRA (11.8%)—is consistent across both training-time and evaluation signals, confirming the finding generalizes beyond mathematical reasoning. QuanTA's advantage over DoRA is larger here (4.0pp) than on DeepMath-Hard (0.0pp at seed 42), consistent with the prediction that higher effective rank provides a clearer advantage when the

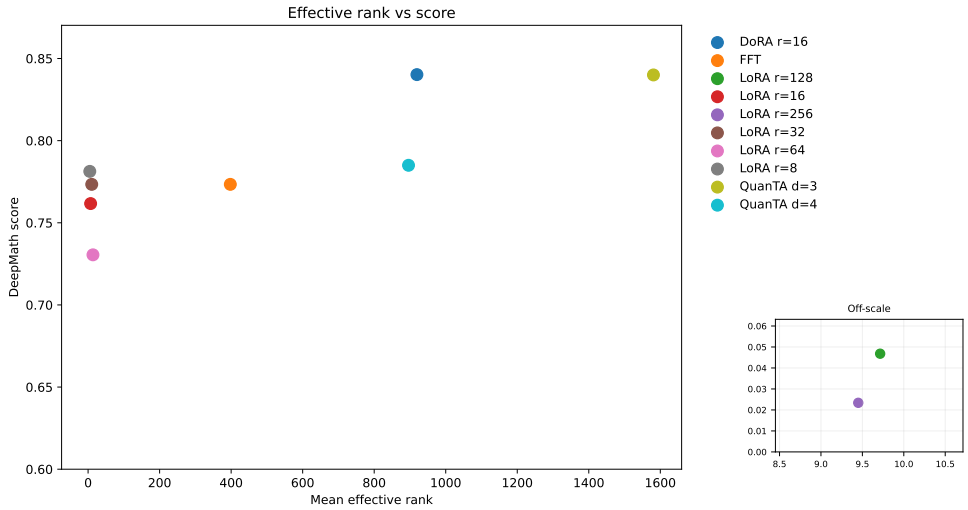

Figure 2. **Mean effective rank versus DeepMath-Hard accuracy (Qwen 3 8B, seed 42).** Performance tracks structured high-rank adaptation rather than trainable parameter count. DoRA and QuanTA achieve high effective rank via structurally distinct mechanisms yet converge to identical accuracy, supporting effective rank as the operative variable rather than any specific parameterization.

Table 6. **Enigmata results (Qwen 3 8B).** Models trained on Enigmata 10K and evaluated on the disjoint Enigmata evaluation set (pass@1).

Method	Config	10K (train)	Eval (pass@1)
LoRA	$r = 8$	3.3%	11.8%
DoRA	$r = 16$	7.7%	16.6%
FFT	—	9.8%	19.1%
QuanTA	$d = 3$ (no_kv)	12.7%	**20.6%**

task ceiling is less binding. The intermediate FFT result (19.1%) falls between the two structured adapters; given that Enigmata results reflect a single seed on a novel and difficult task, we are cautious about interpreting this ordering as meaningful—multi-seed validation would be required to distinguish a genuine effect from the run-to-run variance documented on comparably difficult benchmarks (Appendix E).

5. Discussion and Limitations

Why the RL optimizer causes rank collapse. Sparse binary rewards normalized within groups of $G = 8$ completions produce high-variance advantage estimates that create a pathological opportunity for unconstrained high-rank adapters: rather than distributing update energy across many singular directions to learn a coherent policy, the optimizer concentrates energy into directions that produce reward-correlated surface features without genuine logical deduction. Structural constraints—whether DoRA's magnitude-direction decomposition or QuanTA's MPO tensor network—close off this escape route not by restricting what transformations are representable, but by restricting *which* high-rank solutions are reachable by gradient descent. This unifies all three spectral findings of Section 4.4: rank collapse, the FFT effective-rank paradox, and the convergence of two structurally distinct adapters to similar performance are all consequences of the same dynamic—structural inductive bias functioning as implicit regularization under sparse rewards. This also explains the apparent discrepancy with Schulman & Lab (2025): large-batch training sufficiently reduces advantage variance that the degenerate escape routes available to unconstrained high-rank adapters are never reached within the training horizon.

Model Maturity Hypothesis: evidence and limits. Two independent lines of evidence support the hypothesis. Behaviorally, Apertus and Llama 3.1—architecturally independent models sharing only approximate pre-training data volume—produce identical signatures (near-floor DeepMath-Hard accuracy, QuanTA seed-level instability), ruling out any single family-specific confound. Spectrally, singular value decay of frozen weights before any fine-tuning reveals a monotonic ordering Apertus $\approx$ Llama 3.1 $<$ Qwen 3, predicting the behavioral partition before a single gradient step is taken (Appendix H). We cannot, however, causally isolate pre-training scale from architecture, tokenizer, and data mixture; the hypothesis warrants controlled validation holding architecture fixed across pre-training checkpoints.

Practical takeaways. (1) **Replace the FFT-vs-LoRA binary** with the question of which structured adapter suits the model maturity and compute budget. (2) **Avoid vanilla LoRA above** $r = 32$ in cold-start GRPO at moderate batch sizes—the collapse cliff is steep and irrecoverable. (3) **Co-**

sine scheduling with warmup is recommended at batch size 64; constant schedules cause immediate policy collapse regardless of adapter choice.

Limitations.

- **Scale and architecture.** All experiments are at 8B dense transformer scale; MoE architectures and 70B+ models remain untested.
- **Maturity confounds.** Pre-training scale co-varies with architecture, tokenizer, and data mixture; the three-model comparison cannot disentangle these effects, and the hypothesis remains correlational.
- **Batch size boundary.** Rank collapse is documented at global batch size 64 and may not manifest at larger batch sizes (Schulman & Lab, 2025); the precise threshold at which unconstrained high-rank adaptation transitions from harmful to benign remains uncharacterized, and our guidance against vanilla LoRA above $r = 32$ should be interpreted as conditional on resource-constrained regimes.

6. Conclusion

The field has long framed the choice between Full Fine-Tuning and LoRA as a tradeoff between performance and compute, but our results suggest this framing misidentifies the operative variable. What determines the quality of a learned reasoning policy in RLVR is not how many parameters are updated, but what structure is imposed over the update manifold: unconstrained optimization—whether via high-rank LoRA or full fine-tuning—does not produce maximally expressive updates under sparse rewards, it produces degenerate ones. That two structurally distinct adapters exhibit high effective rank and similar performance outcome suggests the finding is general—any constraint that restricts *which* high-rank solutions gradient descent can reach provides the same protection, regardless of its specific form.

This reframing dissolves the FFT-vs-LoRA dichotomy and replaces it with a more productive question: for a given model and task, what is the appropriate structure over the update manifold? Our evidence suggests the answer depends on pre-training scale—less mature models benefit from the regularization of low-rank constraints, while highly mature models require structured high-rank expressivity to unlock their latent reasoning potential. As base models continue to scale, we expect this demand to grow. The spectral diagnostic framework introduced in our work provides a model-agnostic tool for measuring it. Code can be found here.

References

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning, 2020. URL https://arxiv.org/abs/2012.13255.

Albert, P., Zhang, F. Z., Saratchandran, H., Rodriguez-Opazo, C., van den Hengel, A., and Abbasnejad, E. Randlora: Full-rank parameter-efficient fine-tuning of large models. *arXiv preprint arXiv:2502.00987*, 2025.

Art of Problem Solving. 2025 AIME I problems. https://artofproblemsolving.com/wiki/index.php/2025_AIME_I_Problems, 2025a. AoPS Wiki. American Invitational Mathematics Examination I problems and answer key.

Art of Problem Solving. 2025 AIME II problems. https://artofproblemsolving.com/wiki/index.php/2025_AIME_II_Problems, 2025b. AoPS Wiki. American Invitational Mathematics Examination II problems and answer key.

Art of Problem Solving. 2026 AIME I problems. https://artofproblemsolving.com/wiki/index.php/2026_AIME_I_Problems, 2026a. AoPS Wiki. American Invitational Mathematics Examination I problems and answer key.

Art of Problem Solving. 2026 AIME II problems. https://artofproblemsolving.com/wiki/index.php/2026_AIME_II_Problems, 2026b. AoPS Wiki. American Invitational Mathematics Examination II problems and answer key.

Chen, J., He, Q., Yuan, S., Chen, A., Cai, Z., Dai, W., Yu, H., Yu, Q., Li, X., Chen, J., Zhou, H., and Wang, M. Enigmata: Scaling logical reasoning in large language models with synthetic verifiable puzzles, 2025. URL https://arxiv.org/abs/2505.19914.

Chen, Z., Dangovski, R., Loh, C., Dugan, O., Luo, D., and Soljačić, M. QuanTA: Efficient high-rank fine-tuning of llms with quantum-informed tensor adaptation. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 92210–92245. Curran Associates, Inc., 2024. doi: 10.52202/079017-2928.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

He, Z., Chen, Y., Liang, T., et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025. URL https://arxiv.org/abs/2504.11456.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *CoRR*, abs/2103.03874, 2021. URL https://arxiv.org/abs/2103.03874.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jiang, T., Huang, S., Luo, S., et al. MoRA: High-rank updating for parameter-efficient fine-tuning, 2024. URL https://arxiv.org/abs/2405.12130.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32100–32121. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Liu, W., Qiu, Z., Feng, Y., et al. Parameter-efficient orthogonal finetuning via butterfly factorization, 2023. URL https://arxiv.org/abs/2311.06243.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

Schulman, J. and Lab, T. M. Lora without regret. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250929. URL https://thinkingmachines.ai/blog/lora/.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Zeng, L., Zhong, L., Zhao, L., et al. Skywork-Math: data scaling laws for mathematical reasoning in large language models — the story goes on, 2024. URL https://arxiv.org/abs/2407.08348.

A. Implementation Details

To ensure complete reproducibility of our "Cold Start" RLVR experiments, we detail the exact computational environment, hyperparameters, and reward definitions used.

A.1. Infrastructure and Environment

All experiments were conducted on 4xH100 node. To maintain consistency across runs, we utilized a custom Singularity container. The precise hardware and software stack specifications are listed in Table 7.

Table 7. **Computational Infrastructure.**

Component	Specification
Compute Hardware	4× NVIDIA H100 (64GB) node
Operating System	Ubuntu 24.04 LTS (via Singularity)
CUDA Driver	Version 12.8.0
Precision	bfloat16 (Training & Inference)
Deep Learning Framework	PyTorch 2.9.0 (with Custom torchao 0.13.0)
Training Orchestration	Axolotl 0.13.0
Inference Engine	vLLM 0.11.1 (PagedAttention enabled)
Acceleration Kernels	FlashAttention 2.8.3, Xformers 0.0.33
Distributed Strategy	PyTorch DDP (DistributedDataParallel)

A.2. Hyperparameter Configuration

We utilized the Axolotl training framework with GRPO trainer by TRL integrated into Axolotl. Table 8 details the global optimization parameters held constant across runs to ensure fair comparison. Table 9 details the specific configurations for the adaptation methods and task-specific adjustments.

Table 8. **Global RL Training Hyperparameters.** Parameters chosen to stabilize the "Cold Start" regime with a moderate global batch size.

Parameter	Value
RL Algorithm	Group Relative Policy Optimization (GRPO)
Loss Function	Dr. GRPO (Token-level unbiased)
Group Size (G)	8 generations per prompt
KL Coefficient (β)	0.0 (Pure Outcome-Based RL)
Reward Scaling	False (Raw scores used)
Global Batch Size	64
Optimizer	AdamW (Fused)
Weight Decay	0.0
Scheduler	Cosine Decay
Warmup Ratio	0.1 (10% of steps)
Min LR Ratio	0.15
Num Epochs	1

A.3. Reward Definitions and Verification

In the "Cold Start" regime, the model must simultaneously learn formatting (XML tags) and reasoning logic. We utilized a composite reward function $R_{total} = R_{format} + R_{accuracy}$ with equal weighting ($w = 1.0$).

Table 9. **Method-Specific and Task-Specific Configurations.** We highlight the drastic difference in parameter count between LoRA and QuanTA.

CONFIGURATION	LORA / DORA ($r = 16$)	QUANTA ($d = 4$)	QUANTA ($d = 3$)
RANK / DIMS	$r = 16, \alpha = 32$	$d = 4$, FEATS $= [16, 8, 8, 4]$	$d = 3$, FEATS $= [16, 16, 16]$
TARGET MODULES	ALL LINEAR	ALL LINEAR	ALL LINEAR
DROPOUT	0.0	0.0	0.0
BIAS	NONE	NONE	NONE
TRAINABLE PARAMS (LLAMA 3.1 8B)	41.9M	6.9M	N/A
TRAINABLE PARAMS (QWEN 3 8B)	43.6M	7.7M	35.3M
Task-Specific Learning Rates & Context			
MATH / DEEPMATH / SKYWORK	LR: 1×10^{-5} | MAX LENGTH: 8,192		

A.3.1. ACCURACY REWARD ($R_{accuracy}$)

To evaluate correctness, we strip all formatting tags (e.g., `<think>`) and extract the final answer using a hierarchical strategy:

1. **Robust Box Extraction:** We scan for the **last** occurrence of `\boxed{...}`, utilizing depth-aware parsing to correctly capture mathematical expressions containing nested braces.
2. **Semantic Verification:** We utilize the `math_verify` library to compare the extracted answer against the ground truth. This handles symbolic equivalences (e.g., $\frac{1}{2} = 0.5$, $x + y = y + x$) that strict string matching would miss.
3. **Score:** Returns **1.0** if semantically equivalent, **0.0** otherwise.

A.3.2. FORMAT SHAPING REWARD (R_{FMT})

To induce Chain-of-Thought reasoning without Supervised Fine-Tuning, we impose structural constraints to penalize degeneration (e.g., empty thoughts, infinite loops). The shaping reward is computed as a weighted sum of satisfied constraints:

$$R_{\text{fmt}}(y) = \sum_i w_i \cdot \mathbb{I}(C_i(y)) \tag{5}$$

Constraints (C_i) and Weights (w_i):

- **Structure ($+0.375$):** The output contains a valid closing `</think>` tag.
- **Answer Existence ($+0.375$):** Non-empty content exists after the reasoning block.
- **Hallucination (-0.75):** The model generates simulated user turns (e.g., "User:").
- **Lazy Thinking (-0.6):** The content inside `<think>...</think>` is empty or null.
- **Repetition (-0.5):** Line-level n-gram repetition exceeds the degeneration threshold.
- **Malformed XML (-0.3):** Tags are missing or unclosed.

Note: The maximum cumulative format bonus (0.75) is bounded below the accuracy reward (1.0) to prevent reward hacking, ensuring the optimizer prioritizes correctness over mere template compliance.

B. Data Curation & "Hardcore" Filtering Strategy

B.1. DeepMath-Hard Construction

We utilized the `DeepMath-103K` dataset (He et al., 2025), which aggregates high-difficulty problems from sources like NuminaMath and MathStackExchange. The original dataset contains a difficulty metadata field ranging from 0.0 to 10.0.

Filtering Logic. We applied a strict threshold of ≥ 8.5, discarding the vast majority of the dataset to focus exclusively on the "long tail" of complexity. As illustrated in Figure 3, this reduces the dataset from $\sim 103,000$ to 5,399 samples.

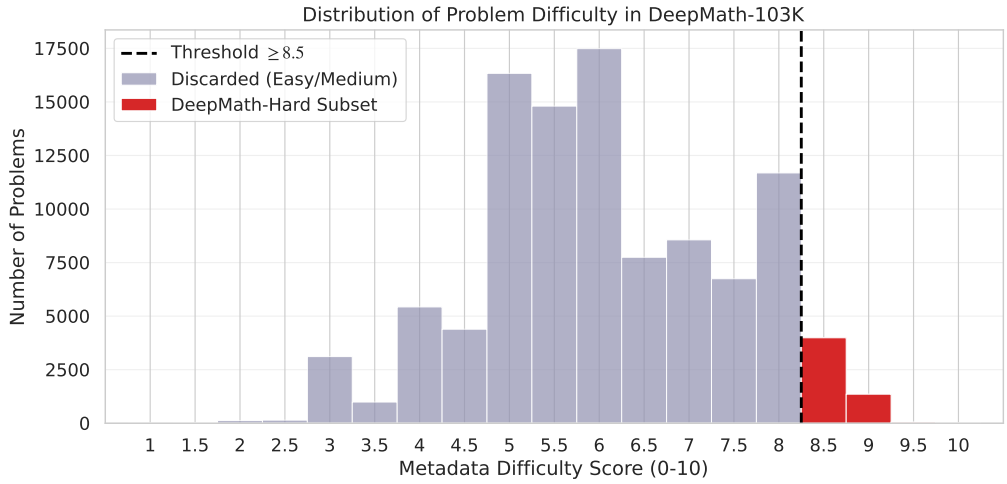

Figure 3. **Difficulty Distribution of DeepMath-103K.** The dashed line represents our cut-off at difficulty score 8.5. We discard the high-volume "easy-medium" mode (grey) to train exclusively on the high-complexity tail (red).

Impact on Topic Distribution. Filtering for difficulty implicitly shifts the topic distribution toward abstract mathematical domains. As shown in Figure 4, the "Hardcore" subset (red) is enriched for fields like *Differential Geometry*, *Abstract Algebra (Group/Field Theory)*, and *Topology*, while simpler procedural topics like standard Calculus are de-emphasized compared to the original distribution (purple). This confirms that the filter selects for semantic reasoning depth.

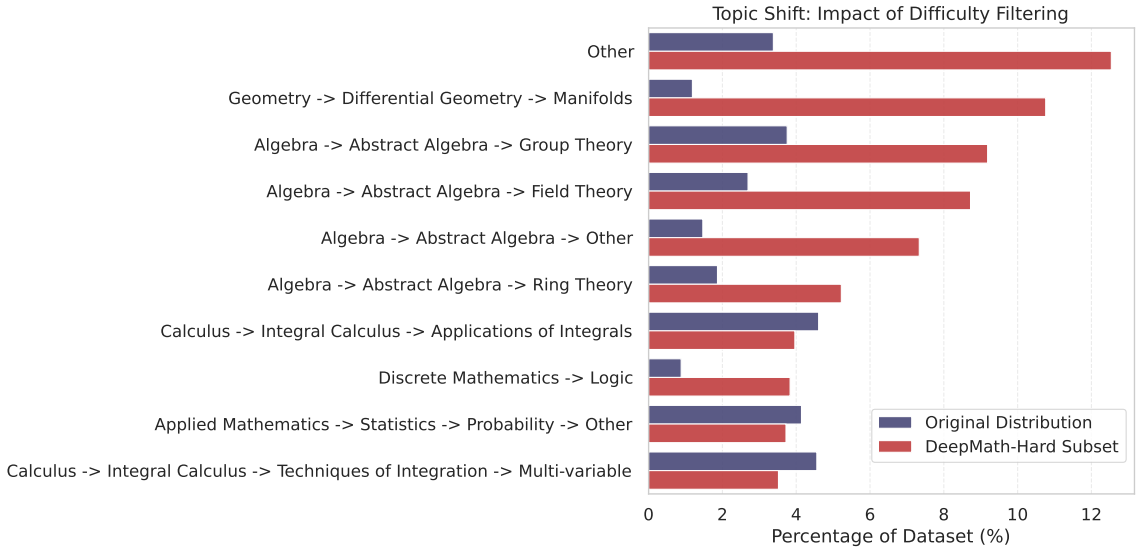

Figure 4. **Topic Shift Analysis.** Filtering for difficulty (≥ 8.5) fundamentally alters the dataset composition, enriching for abstract reasoning domains like Manifolds and Field Theory while suppressing procedural calculus.

B.2. Skywork-Hard and the 32B Judge

For the `Skywork-OR1-RL-Data` (Zeng et al., 2024), we utilized the `train-math-deepscaler` subset. Unlike DeepMath, this dataset includes metadata on how larger models performed on each specific problem.

The 32B Judge Protocol. We utilized the `model_difficulty` metadata field for the

`DeepSeek-R1-Distill-Qwen-32B` model. This field reports a score inversely proportional to the pass rate. We selected a cutoff score of ≥ 13, in the Skywork metadata schema, higher scores indicate higher failure rates for the 32B model.

Distributional Cutoff. Figure 5 visualizes this exact cutoff. The filtering removes the massive volume of "trivial" problems and isolates a cluster of 6,702 problems that consistently stump significantly larger models.

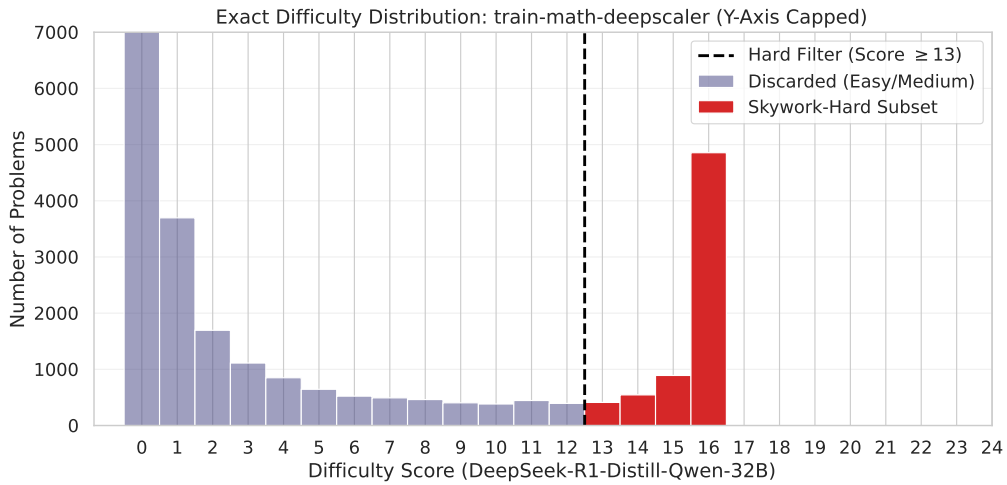

Figure 5. **The 32B Judge Filter on Skywork-Math.** The y-axis is capped to visualize the tail. The vast majority of problems (Scores 0-12) are discarded. We retain only the subset (red) where a 32B parameter model fails consistently (Score ≥ 13).

B.3. Enigmata 10k Train and Eval

We construct a reduced Enigmata training set from the original Enigmata-Data training split released by BytedTsinghua-SIA. The full training split contains 217,541 examples across 38 task names, with a highly non-uniform task distribution. Most tasks contain 6,000 examples each, but several deviate substantially from this mode: `full_crosswords` contains 19,000 examples, `arc_agi` 11,116, `pattern_recognition` 9,000, `hitori`, `light_up`, `natural_language_navigation`, and `tic_tac_toe` 4,000 each, `knights_and_knaves` 3,414, `checkmate_in_one` 2,500, `big_bench_symbolic` 2,000, `hamiltonian_path` 1,095, and `symbolic_hard` 416. Difficulty labels are similarly imbalanced: among rows with valid difficulty annotations, the original training split contains 56,477 easy, 77,451 medium, and 83,197 hard examples.

Shared-task restriction. To ensure that training and evaluation measure performance over the same task family, we first restrict the training pool to tasks shared between the Enigmata training split and the official Enigmata evaluation split. This removes the train-only tasks `arc_agi`, `pattern_recognition`, and `symbol_pattern`. Conversely, the eval-only task `FOLIO` cannot be included because it is absent from the training split. The resulting candidate pool therefore consists of the 35 task names common to both training and evaluation.

Stratified 10k construction. From this shared-task pool, we sample 10,000 examples using task-by-difficulty strata. The sampling procedure enforces two constraints. First, the marginal task distribution of the reduced set matches the relative task frequencies of the original Enigmata training split after restricting to shared tasks. Second, the marginal difficulty distribution matches that of the official evaluation set, rather than the more imbalanced difficulty distribution of the original training split. Concretely, the final Enigmata 10k subset contains 2,930 easy, 3,381 medium, and 3,689 hard examples.

This construction preserves the original train-task mixture as closely as possible while aligning the reduced training set with the evaluation set's difficulty profile. As a result, the Enigmata experiments test generalization over the same task families present at evaluation time without allowing train-only tasks to distort the reduced training distribution. All Enigmata models are trained on this 10k subset and evaluated on the disjoint official Enigmata evaluation set.

C. Constant Learning Rate Collapse

Figure 6 shows the training reward curve of QuanTA $d = 3$ on Llama 3.1 8B trained on MATH dataset under a constant learning rate schedule (all other hyperparameters identical to the main experiments). The policy initially learns, peaking near step 200, before collapsing catastrophically to near-zero reward by step 280—a failure mode we observe consistently across all adapter configurations under constant scheduling at global batch size 64.

We attribute this to the elevated variance of GRPO advantage estimates at moderate batch sizes. Without a decaying learning rate, the optimizer continues to take large gradient steps as advantage estimates become increasingly noisy in later training, destabilizing the policy irreversibly. A cosine decay schedule with warmup ratio 0.1 and minimum learning rate ratio 0.15 eliminates this failure mode across all methods; we therefore adopt it as a fixed component of our protocol rather than a tunable hyperparameter.

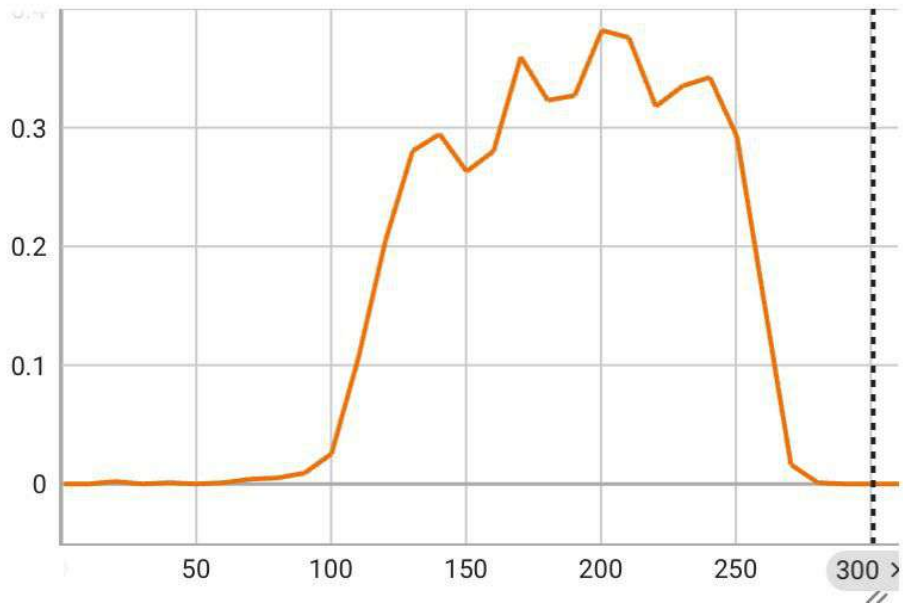

Figure 6. Training reward under a constant learning rate schedule (QuanTA $d = 3$, Llama 3.1 8B, MATH dataset). The policy collapses catastrophically near step 280 despite healthy early progress, illustrating why cosine decay with warmup is suggested in cold-start GRPO at batch size 64. The dashed vertical line marks the early stoppage of training.

D. Full Rank Sweep Results

D.1. LoRA Rank Sweep with Spectral Diagnostics

Table 10 extends the main-body rank sweep (Table 4) with mean effective rank measurements for each configuration, enabling direct comparison between nominal and effective rank.

Two aspects of Table 10 are worth emphasizing. First, effective rank increases modestly but consistently from $r = 8$ to $r = 64$ (4.7 $\rightarrow$ 13.6), yet performance is non-monotonic over this range—confirming that effective rank alone does not determine policy quality, and that the structure of the update manifold matters independently of its dimensionality. Second, collapsed runs ($r = 128$, $r = 256$) achieve *lower* effective rank than the best non-collapsed configurations despite their higher nominal rank: the optimizer does not distribute energy across the available directions but instead concentrates it, a signature of the degenerate reward-hacking dynamic described in Section 5.

D.2. DoRA Rank Sweep

Table 11 presents the full DoRA rank sweep on Qwen 3 8B. Unlike LoRA, DoRA does not exhibit catastrophic collapse at any rank tested: performance peaks at $r = 16$ (84.0%) and degrades gracefully to 75.0% at $r = 64$, with no discontinuous cliff.

Table 10. **Full LoRA rank sweep on Qwen 3 8B, DeepMath-Hard (seed 42).** Mean effective rank is the participation ratio $\rho = \exp(H(\mathbf{p}))$ averaged across all 252 target modules. Note the divergence between nominal and effective rank at $r \geq 128$: despite 128 and 256 nominal singular directions, the optimizer concentrates nearly all update energy into fewer than 10 effective directions.

Method	Nominal Rank	Trainable Params	Mean Eff. Rank	DeepMath-Hard
LoRA	$r = 8$	22M	4.7	78.1%
LoRA	$r = 16$	43.6M	7.1	76.2%
LoRA	$r = 32$	87.2M	10.2	77.3%
LoRA	$r = 64$	174M	13.6	73.1%
LoRA	$r = 128$	348M	9.7	4.7%
LoRA	$r = 256$	696M	9.5	2.3%

Table 11. **DoRA rank sweep on Qwen 3 8B, DeepMath-Hard (seed 42).** Graceful degradation at high rank contrasts sharply with LoRA's catastrophic collapse, implicating DoRA's magnitude-direction decomposition as a structural regularizer that closes off the degenerate escape routes available to vanilla LoRA.

Method	Nominal Rank	Trainable Params	DeepMath-Hard
DoRA	$r = 8$	23.2M	81.9%
DoRA	$r = 16$	45.0M	**84.0%**
DoRA	$r = 32$	90.0M	80.9%
DoRA	$r = 64$	180M	75.0%

The asymmetry between LoRA and DoRA at high rank is mechanistically informative. DoRA's decomposition of each weight matrix into independent magnitude and direction components (Equation 3) introduces a structural constraint that LoRA lacks: even at high nominal rank, the optimizer cannot simultaneously rescale feature magnitudes and rotate directions without each component resisting the other's contribution. This implicit coupling acts as a regularizer on the update manifold, preventing the energy concentration that produces rank collapse in unconstrained LoRA.

D.3. Efficiency Summary

Table 12. **Efficiency comparison, Qwen 3 8B on DeepMath-Hard (seed 42).** Training time reflects a full single-epoch run on 4×H100. QuanTA $d = 3$ is the Pareto-optimal choice: identical peak accuracy to DoRA $r = 16$ at 22% fewer parameters and 52% less wall-clock time.

Method	Config	Trainable Params	Training Time	DeepMath-Hard
LoRA	$r = 8$	22.0M	13.1h	78.1%
LoRA	$r = 16$	43.6M	12.4h	76.2%
DoRA	$r = 16$	45.0M	15.6h	84.0%
QuanTA	$d = 4$	7.7M	12.9h	78.5%
QuanTA	$d = 3$	35.3M	7.5h	84.0%

QuanTA $d = 3$'s speed advantage over both $d = 4$ and DoRA despite its larger parameter count reflects the greater parallelism of the $[16, 16, 16]$ MPO tensor contraction relative to DoRA's per-column normalization overhead.

E. Multi-Seed Results

All results in this appendix use seeds 42, 43, and 44. Standard deviations reflect genuine stochasticity in cold-start GRPO at batch size 64, not measurement noise; single-seed comparisons from the main body should be interpreted in light of the variance reported here.

E.1. Qwen 3 8B — DeepMath-Hard

Table 13. **Multi-seed DeepMath-Hard results, Qwen 3 8B.** All structured adapter configurations show substantial run-to-run variance consistent with the difficulty of cold-start GRPO on a highly filtered corpus. No method achieves a statistically clean win on mean accuracy; the primary signal is that structured adapters consistently outperform LoRA and FFT across all three seeds individually.

Method	Config	Seed 42	Seed 43	Seed 44	Mean	Std
DoRA	$r = 16$	84.0%	76.6%	62.5%	74.4%	10.9%
DoRA	$r = 16$ (no_kv)	80.9%	77.0%	60.9%	72.9%	10.4%
QuanTA	$d = 3$ (linear)	78.5%	79.3%	64.5%	74.1%	8.4%
QuanTA	$d = 3$ (no_v_proj)	80.1%	78.5%	63.7%	74.1%	9.0%
QuanTA	$d = 3$ (no_kv)	84.0%	76.2%	65.2%	75.1%	9.6%

QuanTA $d = 3$ (no_kv) achieves the highest mean (75.1%) and is the Pareto-optimal configuration in the main body results, though its advantage over DoRA $r = 16$ (74.4%) and the other QuanTA variants (74.1%) is well within one standard deviation. Seed 44 is consistently the lowest-performing seed across all methods, suggesting a particular initialization is unfavorable for cold-start GRPO on this corpus rather than any method-specific fragility.

E.2. Qwen 3 8B — Skywork-Hard

Table 14. **Multi-seed Skywork-Hard results, Qwen 3 8B.** High problem difficulty produces substantial reward variance throughout training, which propagates into evaluation variance. Differences between methods are within noise; the primary signal is that all methods remain functional on this benchmark, unlike the catastrophic collapses observed at high LoRA rank.

Method	Config	Seed 42	Seed 43	Seed 44	Mean	Std
LoRA	$r = 8$	19.0%	13.2%	21.4%	17.9%	4.3%
QuanTA	$d = 4$	19.6%	11.2%	18.5%	16.4%	4.6%
QuanTA	$d = 3$	19.4%	14.5%	21.7%	18.5%	3.6%

QuanTA $d = 3$ achieves the highest mean (18.5%) and the lowest standard deviation (3.6%) across methods, suggesting marginally more stable optimization on this benchmark.

E.3. Llama 3.1 8B — MATH

In Table 15, the seed-43 collapse of QuanTA $d = 4$ on Llama 3.1 8B (0.8%, effectively random) is not an outlier to be discarded—it is the central prediction of the Model Maturity Hypothesis. A less mature model lacks the distributed pre-trained representations that QuanTA's high effective rank is designed to activate; when the initialization happens to place the optimizer in an unfavorable basin, there is insufficient signal in the sparse binary reward to recover. LoRA's lower effective rank inadvertently regularizes against this failure mode by constraining the optimizer to a manifold from which recovery is more tractable. Qwen 3 8B exhibits no analogous instability under QuanTA on any benchmark or seed, which constitutes the contrastive evidence for the maturity partition.

F. QuanTA Target-Module Ablation

Table 16 reports DeepMath-Hard results for QuanTA across three target-module configurations: `target_linear` (all linear modules), `no_v_proj` (all linear modules except the value projection), and `no_kv` (all linear modules except key and value projections).

Table 15. **Multi-seed MATH results, Llama 3.1 8B.** LoRA $r = 16$ is stable across all seeds; QuanTA $d = 4$ collapses entirely on seed 43 (0.8%), producing high variance and a substantially lower mean. This asymmetry—stable under LoRA, volatile under structured high-rank adaptation—is the primary behavioral signature of the Model Maturity Hypothesis for less mature models.

Method	Config	Seed 42	Seed 43	Seed 44	Mean	Std
LoRA	$r = 16$	28.9%	34.3%	27.3%	30.2%	3.7%
QuanTA	$d = 4$	26.6%	0.8%	21.9%	16.4%	13.7%

Table 16. **QuanTA target-module ablation on Qwen 3 8B, DeepMath-Hard.** Excluding key and value projections consistently produces the best or near-best single-seed performance for $d = 3$, and the best mean across seeds. $d = 4$ multi-seed results are not available; seed-42 values are reported for reference.

Config	Modules	Seed 42	Seed 43	Seed 44	Mean	Std
$d = 3$	target_linear	78.5%	79.3%	64.5%	74.1%	8.4%
$d = 3$	no_v_proj	80.1%	78.5%	63.7%	74.1%	9.0%
$d = 3$	no_kv	**84.0%**	76.2%	65.2%	**75.1%**	9.6%
$d = 4$	target_linear	79.7%	—	—	—	—
$d = 4$	no_kv	78.5%	—	—	—	—

The consistent advantage of `no_kv` for $d = 3$ across both seed-42 and mean performance is notable given that excluding more modules from adaptation might naively be expected to reduce expressivity. We interpret this as evidence that key and value projections encode content-retrieval representations that should remain stable during reasoning alignment: their role is to attend to the right tokens, a capability established during pre-training that aggressive adaptation may degrade rather than enhance. The primary locus of reasoning circuit modification appears to be query projections and feed-forward layers, which receive the full adaptation budget in the `no_kv` configuration. For $d = 4$, the `target_linear` configuration marginally outperforms `no_kv` on seed 42 (79.7% vs. 78.5%), though the difference is small and multi-seed validation is unavailable; we defer a definitive recommendation for the $d = 4$ configuration to future work.

G. AIME 2025–2026 Generalization

Table 17. **AIME 2025–2026 out-of-distribution generalization (Avg@32), Qwen 3 8B.** All models trained on DeepMath-Hard only ($\sim$5,400 samples). Full fine-tuning collapses to near-random performance despite achieving competitive in-distribution accuracy on DeepMath-Hard (77.3%, Table 5).

Method	Config	AIME Avg@32
LoRA	$r = 16$	9.95%
DoRA	$r = 16$	**11.04%**
QuanTA	$d = 3$ (no_kv)	10.36%
FFT	—	1.35%

FFT's near-floor AIME performance (1.35%) requires careful interpretation because it is not straightforwardly explained by in-distribution overfitting. If FFT had simply memorized DeepMath-Hard's surface features, we would expect it to score substantially *higher* than LoRA in-distribution; instead, it scores comparably (77.3% vs. LoRA $r = 8$ at 78.1%), while collapsing far more severely out-of-distribution.

The explanation lies in the effective-rank paradox documented in Section 4.4. FFT concentrates its learned updates into $\sim$400 effective singular directions spread across 8B unconstrained parameters. With no structural prior governing *which* directions those updates occupy, the RL optimizer is free to align them with the distributional idiosyncrasies of DeepMath-Hard—its graduate-level topic vocabulary, abstract algebra and differential geometry conventions, and characteristic problem phrasing—rather than with the underlying reasoning operations that would transfer across problem styles. This produces updates that are simultaneously low effective rank *and* highly dataset-specific: narrow enough to miss the general reasoning structure, unconstrained enough to track surface features that are useless on AIME's concise, elementarily-stated competition

problems.

LoRA, despite its own low effective rank ($\sim$7), is protected from this failure mode by its structural constraint. The rank-r manifold forces the optimizer to find solutions that work within a tighter geometric subspace, which inadvertently prevents the learned directions from aligning too precisely with DeepMath-Hard's idiosyncratic surface. Structured adapters (DoRA, QuanTA) avoid both failure modes simultaneously: their higher effective rank ($\sim$900 and $\sim$1580 respectively) encodes richer representations, and their structural constraints ensure those representations are distributed across directions that generalize. The AIME result is therefore the out-of-distribution analogue of the rank collapse phenomenon: in both cases, unconstrained optimization under sparse rewards finds solutions that satisfy the immediate objective while failing to learn what the task actually requires.

H. Frozen Weight Spectral Analysis

Figure 7 plots the normalized singular value decay of frozen linear layers across all three base models, measured before any fine-tuning occurs. This analysis serves as an independent, training-free predictor of the behavioral partition documented in Section 4.3.

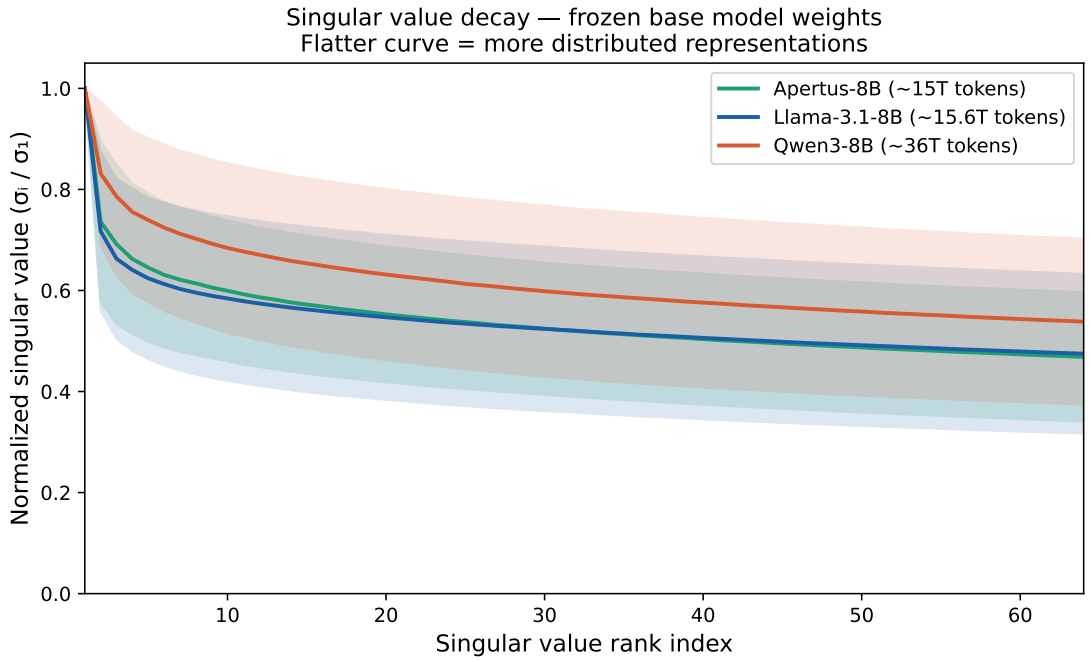

Figure 7. **Singular value decay of frozen base model weights, averaged across sampled linear layers using the top 64 singular values per layer.** Shaded regions denote one standard deviation across layers. Qwen 3 8B ($\sim$36T tokens) exhibits markedly flatter decay than both $\sim$15T models, indicating more distributed representations in its pre-trained weights. Critically, Apertus-8B and Llama 3.1-8B—architecturally independent models differing in tokenizer, hidden activation function, and training data mixture—produce nearly identical decay curves, ruling out any single family-specific confound. This pre-training spectral signature predicts the behavioral partition observed in Section 4.3: the two models whose frozen weights decay fastest are precisely the two models for which structured high-rank adaptation offers no advantage over LoRA.

Three observations follow from Figure 7.

The maturity ordering is visible in frozen weights. Qwen 3 8B's decay curve sits consistently above the two 15T models across all 64 singular value indices, indicating that its pre-trained representations are more evenly distributed across the weight spectrum. This is not a post-hoc measurement—it is observable on the base checkpoint before any gradient step is taken, making it a prospective predictor of adaptation behavior rather than a retrospective one.

Apertus and Llama are spectrally indistinguishable. Despite differing in architecture, tokenizer, and training data mixture, the two $\sim$15T models produce overlapping decay curves with nearly identical means throughout. This convergence

of independent model families to the same spectral signature at the same pre-training data volume constitutes the strongest available evidence that data volume—rather than any family-specific design choice—is the primary driver of the observed behavioral partition.

The proposed mechanism. A model whose frozen weights exhibit steep singular value decay has concentrated its pre-trained representations into a small number of high-variance directions. Adapting such a model is well-suited to low-rank methods: the useful update subspace is itself low-dimensional, and LoRA's rank constraint aligns with the geometry of the task. Conversely, a model with flatter decay—such as Qwen 3 8B—encodes information more uniformly across its weight spectrum. Activating the latent reasoning circuits in such a model requires updates distributed across many singular directions simultaneously; a low-rank adapter projects this requirement onto a subspace that is too narrow to capture it faithfully, while QuanTA's MPO structure provides the distributed update geometry the task demands. This mechanism connects the pre-training spectral signature directly to the performance gap observed between LoRA and structured adapters on Qwen 3 8B, and explains why no such gap appears on the 15T models.

I. Qualitative Analysis: The Rank Bottleneck in Action

To understand the qualitative difference between high-rank and low-rank adaptations, we analyze the Chain-of-Thought (CoT) traces generated by Qwen 3 8B on a complex counting problem from the **Skywork-Hard** subset. This problem requires combinatorial reasoning with inclusion-exclusion principles and state tracking (i.e., tracking which sprinters remain available).

We compare the output of the best **LoRA** checkpoint (Rank 16) against the **QuanTA** checkpoint ($d = 3$) sampled from the final 5% of the training run.

I.1. Case Study: Combinatorial Counting

Problem: *In an Olympic 100-meter final, there are 10 sprinters competing, among which 4 are Americans. The gold, silver, and bronze medals are awarded to first, second, and third place, respectively. Calculate the number of ways the medals can be awarded if at most two Americans are to receive medals.*

Correct Answer: 588.

LoRA ($r = 16$) **Trace [FAILURE]**	**QuanTA** ($d = 3$) **Trace [SUCCESS]**
... (Case 1: No Americans is correct) ...	*... (Case 1: No Americans is correct) ...*
Case 2: Exactly one American. "We have 4 choices for which American receives a medal. For that American, we have 4 choices (since they can get any of the 4 positions). [ERROR: There are only 3 medal positions, not 4] ... The number of ways is $4 \times 4 \times 6 \times 5 = 480$."	**Case 2: Exactly one American.** "In this case, we have to choose which of the three medals the American will receive, then choose which one of the 4 Americans will receive that medal... The number of ways is: $3 \times 4 \times P(6, 2) = 3 \times 4 \times 30 = \mathbf{360}$."
Case 3: Exactly two Americans. "...For the remaining medal, we have 6 choices. The number of ways to award the medals in this case is: $6 \times 6 \times 6 = 216$." [ERROR: Failed to account for permutations of the specific Americans selected]	**Case 3: Exactly two Americans.** "...Choose which two of the three medals the Americans will receive, then choose which two of the 4 Americans... The number of ways is: $3 \times \binom{4}{2} \times 6 = 3 \times 6 \times 6 = \mathbf{108}$."
Final Calculation: $120 + 480 + 216 = \mathbf{816}$	**Final Calculation:** $120 + 360 + 108 = \mathbf{588}$

Table 18. **Reasoning Failure Mode.** LoRA (left) correctly identifies the cases but fails to model the state-dependent permutations within Case 2 and Case 3, defaulting to erroneous multiplication heuristics. QuanTA (right) correctly applies the permutation formula $P(n, k)$ and binomial coefficients $\binom{n}{k}$, demonstrating a robust capability to maintain high-dimensional state information (available slots vs. available candidates) throughout the reasoning chain.

I.2. Analysis of the Failure

The LoRA failure is characteristic of a "Rank Bottleneck." The model successfully retrieved the high-level strategy (Case Analysis) but failed to execute the granular logic required to link the sub-steps.

- **Loss of Precision:** In Case 2, LoRA hallucinated a 4th medal position ("any of the 4 positions"), a semantic drift that suggests the representation of the problem constraints (3 medals) was lost during the intermediate computation.
- **Heuristic Collapse:** In Case 3, LoRA defaulted to a simple 6^3 calculation, ignoring the dependency between slots. This suggests the model reverted to a lower-complexity heuristic when the reasoning load exceeded its expressive capacity.

QuanTA, utilizing a high-rank tensor manifold, maintained the precise state of the problem constraints throughout the generation, correctly deriving the combinatorial factors for all three cases.

J. PEFT Theoretical Comparison

Our spectral analysis identifies two properties jointly necessary for strong performance in mature-model RLVR: **(a)** high effective rank in the learned ΔW, and **(b)** structural prevention of degenerate reward-satisfying solutions. We analyze each method through this lens, with particular focus on why DoRA and QuanTA satisfy both criteria via structurally distinct mechanisms while arriving at the same empirical outcome, and why BOFT and MoRA are geometrically precluded from doing so.

LoRA. Constraining $\Delta W = BA$ to the rank-r determinantal variety $\mathcal{M}_r$ bounds rank from above but not from below: nothing prevents the optimizer from concentrating all update energy into a handful of singular directions. In SFT this is benign; in RLVR it enables the catastrophic collapse we document at $r \geq 128$, where near-zero effective rank persists despite high nominal rank. LoRA satisfies neither criterion.

DoRA. Column-wise decomposition of each weight into an independent magnitude scalar and unit-norm direction, with LoRA applied only to the directional component, prevents the optimizer from satisfying sparse binary rewards purely by rescaling dominant feature directions. This inadvertently redistributes gradient energy across singular directions—not by design, but as a consequence of the coupling introduced between magnitude and direction updates. The result is an effective rank of $\sim$900 at r=16, versus $\sim$7 for vanilla LoRA at identical nominal rank. DoRA satisfies both criteria: the decomposition raises effective rank while simultaneously constraining which reward-satisfying solutions are reachable by gradient descent.

QuanTA. Parameterizing ΔW as an MPO tensor network constrains updates to a tensor manifold $\mathcal{T}_\chi$ whose elements are generically full-rank as matrices—unlike $\mathcal{M}_r$. Tensor contraction distributes update energy across all mode indices by construction, yielding the flat singular value spectra we measure ($\sim$1580 effective rank for d=3). Simultaneously, not all full-rank matrices lie on $\mathcal{T}_\chi$, so the degenerate reward-hacking solutions available to unconstrained optimizers are structurally excluded. QuanTA satisfies both criteria via a mechanism entirely distinct from DoRA's: where DoRA achieves high effective rank through magnitude-direction decoupling, QuanTA achieves it through the tensor contraction structure itself. That two methods arrive at the same spectral signature and the same task performance through independent geometric mechanisms is the strongest evidence that effective rank produced by structural constraint—not any property specific to either parameterization—is the operative variable.

BOFT. Restricting updates to the orthogonal group $\mathrm{O}(d)$ via butterfly-factored rotation matrices produces norm-preserving updates by construction: $\|Wx\|_2 = \|W_0 x\|_2$ for all inputs. While this satisfies criterion (b)—orthogonal updates cannot produce degenerate reward-hacking rescalings—it structurally fails criterion (a). Activating underutilized reasoning circuits in a mature model likely requires non-trivial rescaling of feature magnitudes, which orthogonal updates cannot perform by definition. The update manifold of BOFT is therefore structurally incompatible with the distributed, magnitude-modifying updates that our spectral analysis identifies as necessary for mature-model RLVR.

MoRA. Sandwiching a learnable square matrix between fixed non-learnable compression operators constrains updates to the image of a linear subspace of $\mathbb{R}^{d\times k}$ determined before any gradient information is observed. While this can in principle

achieve high effective rank within that subspace— satisfying criterion (a) approximately—it structurally fails criterion (b): the fixed compression operators cannot adapt to exclude the specific degenerate solutions that emerge during RLVR training, since their structure is set a priori. In cold-start GRPO, where the optimizer must simultaneously discover directions encoding output formatting, reasoning strategies, and domain knowledge, the inability to learn the compression structure itself is particularly limiting. QuanTA's learnable core tensors allow the optimal entanglement structure to emerge from training rather than being fixed before it begins.

Summary.

Method	(a) High effective rank	(b) Degenerate solution exclusion
LoRA	×	×
DoRA	✓	✓
QuanTA	✓	✓
BOFT	×	✓
MoRA	∼	×
FFT	×	×

The geometric prediction that follows is testable: any PEFT method satisfying both criteria should outperform LoRA and FFT in cold-start GRPO on mature models, regardless of the specific structural form used. DoRA and QuanTA constitute two independent confirmations of this prediction; methods satisfying only one criterion are expected to occupy an intermediate performance tier.