Learning from Less: Measuring the Effectiveness of RLVR in Low Data and Compute Regimes

Justin Bauer, Thomas Walshe, **Derek Pham**, Harit Vishwakarma, Armin Parchami, Frederic Sala, Paroma Varma

Motivation

- RLVR — rewarding verifiable correct answers — has become standard in reasoning post-training
- Most results assume abundant data and compute (e.g., DeepMath-103K: 100K+ samples)

- Prior scaling work focuses on model size and compute (Kaplan, Hoffmann, ScaleRL, Tan et al.)
- Data composition under fixed budgets remains under-studied

How does model performance evolve when data and compute are limited, and what data characteristics impact generalization in such regimes?

Three Procedural Datasets for Controlled Study

Programmatically generated → controllable size, complexity, diversity · verifiable ground truth

Template
Parameterized schema per dataset

→

Sample params
Operators, ranges, step counts, etc.

→

Generate problem
Natural language output

→

Ground truth
Programatically compute answer

→

Verify
Validate problem; add to dataset

Counting Problems

Multi-step numerical reasoning over integer sequences

Example:
Q: Consider integers 1-100.
Keep only even numbers.
Keep only those divisible by 3.
How many remain?

A: 16

Graph Reasoning

Graph-theoretic reasoning on node/edge structures

Example:
Q: Find the maximum independent set of a graph with 5 nodes.
Nodes: [0,1,2,3,4]
Edges: [(0,2), (0,4)]

A: [1, 2, 3, 4]

Spatial Reasoning

Tracking entities in 2D spatial environments

Example:
Q: P1 at (-1.5, 2.5) faces East.
P2 at (3.5, 1.5) faces West.
P1 moves forward, P2 moves back.
Location of P1 relative to P2?

A: (-5.0, 1.0)

~1,500 problems generated per dataset

Calibrating Difficulty with Frontier Models

- Generate **~1,500 problems per dataset**
- Evaluate every problem across **10 LLMs**
- Tier each problem by **fraction of models that solve it correctly**
- Difficulty grounded in **empirical capability**, not human intuition

Easy	Medium	Hard
67–100% of models correct	34–66% of models correct	0-33% of models correct

Curated Subsets

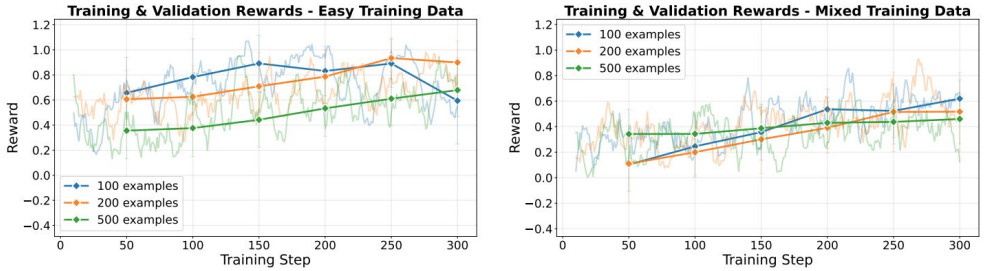
- **Training**: Easy-only and Mixed (~33% per tier), at sizes **100 / 200 / 500**
- **Test**: 200-500 examples held out, spans all tiers
- All splits strictly disjoint

Experimental Setup

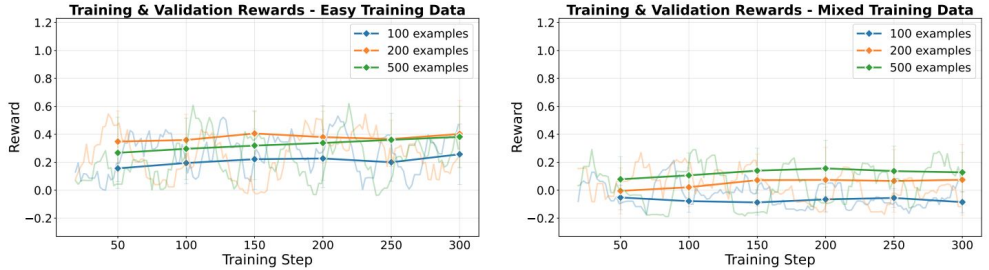
Model	Qwen3-4B with LoRA (rank 64, alpha 16) ~100M trainable parameters out of 4B
Algorithm	GRPO (Group Relative Policy Optimization) 5-8 completions per prompt (K=8 counting/graph, K=5 spatial)
Compute	4x NVIDIA A100 80GB GPUs Fixed step budget: 300 steps (counting/graph), 1000 steps (spatial) 5-12 hours training time Max completion length: 2048 tokens
Data Configs	Per dataset: Easy-{100, 200, 500} and Mixed-{100, 200, 500} Mixed = ~33% each of Easy, Medium, Hard
Evaluation	Held-out mixed-difficulty test sets (200-500 examples)

Training

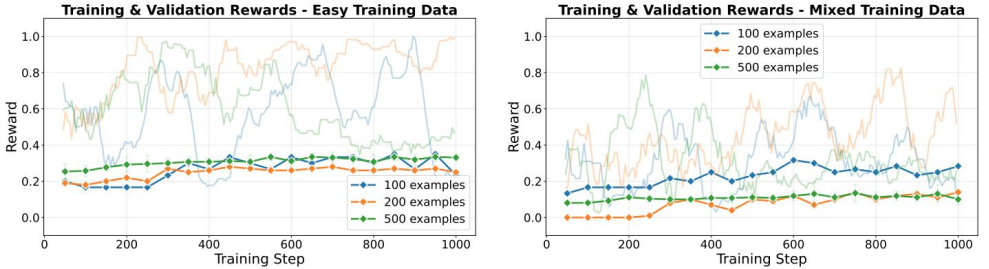
- **One model** per (dataset, composition, size) — **18 fine-tuning runs** total
- Train and test always **within the same dataset**
- Why one at a time: 5–12 hours per run on 4× A100s; cross-dataset transfer is a separate experimental design



(a) Counting: training and validation reward over 300 steps. Note the instability in the easy 100-example model (collapse after step 200).

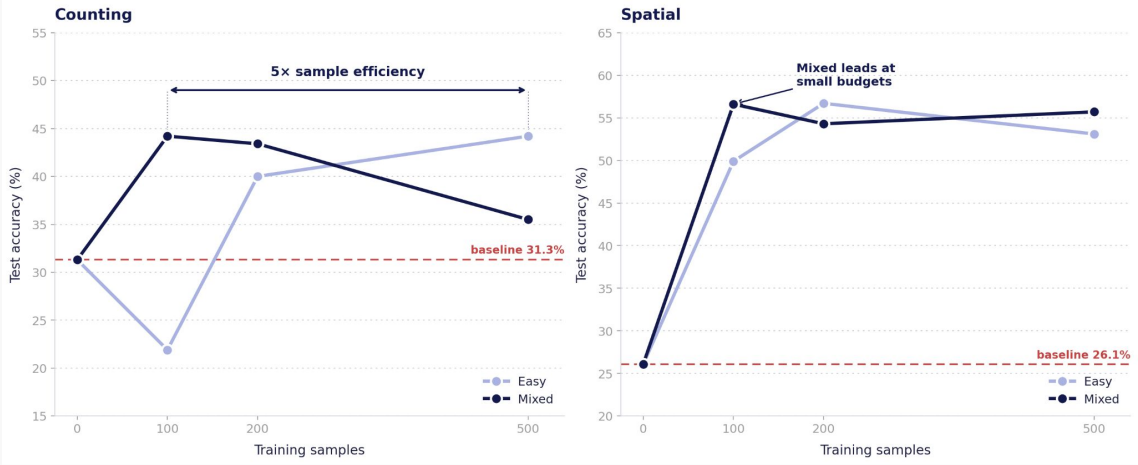


(b) Graph: training and validation reward over 300 steps. Easy-only training (left) achieves positive validation rewards; mixed-difficulty (right) shows consistently negative rewards due to incomplete rollouts under token constraints.



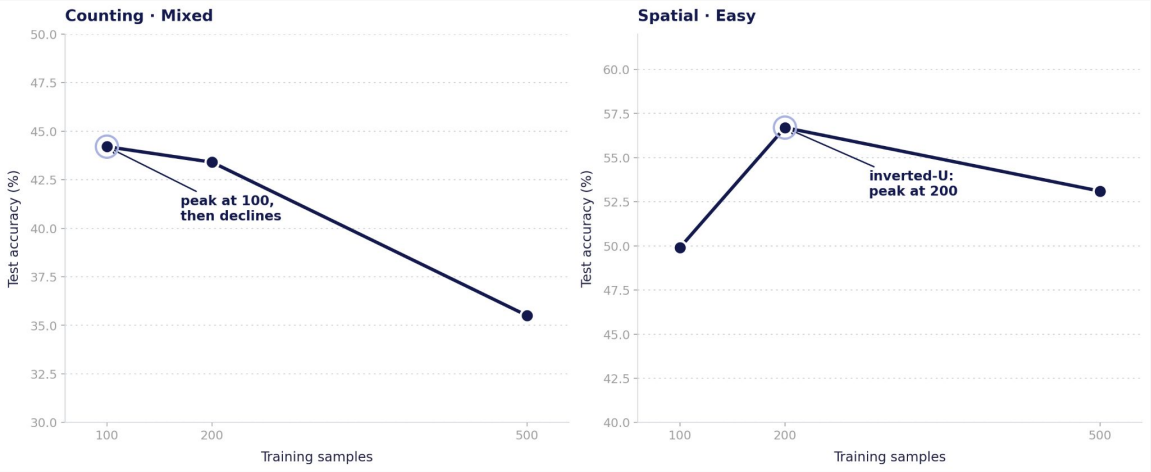
(c) Spatial: training and validation reward over 1000 steps. Binary reward ($r \in \{0, 1\}$) creates discrete performance levels. Both regimes show steady improvement.

Finding 1 — Composition outweighs volume



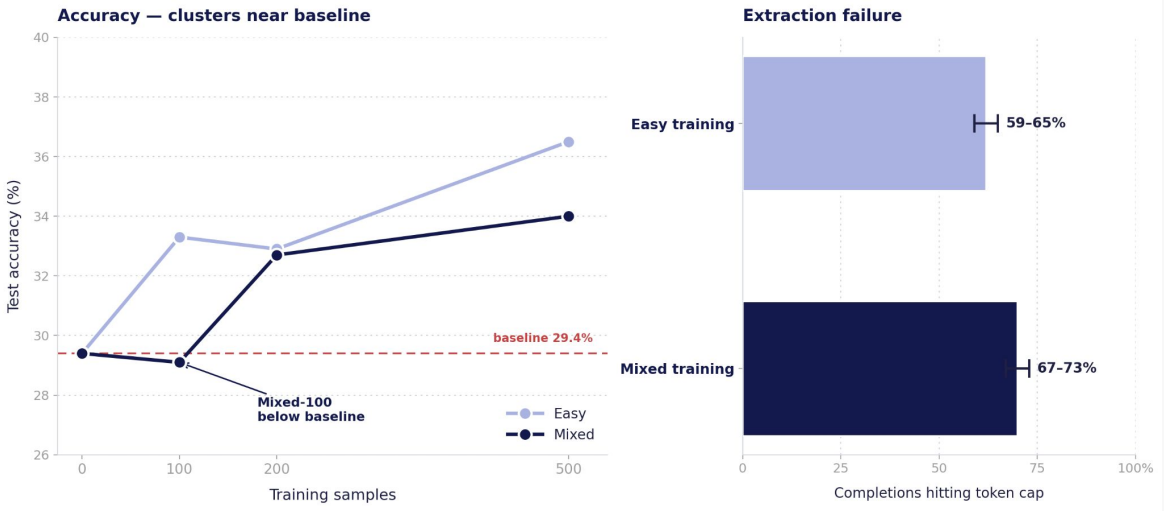
- **Counting**: 100 mixed examples match 500 easy-only → **5× sample efficiency**
- **Spatial**: Mixed configs hold steady (~50–57%), exceeding Easy-only at small budgets
- **Caveat**: on Counting, Mixed degrades past 100 — over-mixing can hurt under fixed step budget

Finding 2 — Scaling data alone can be ineffective



- **Fixed step budget**: larger datasets receive fewer optimization updates per example
- **Counting**: Mixed accuracy declines past 100 — validation reward still climbing at final step
- **Spatial**: Easy peaks at 200, then declines (inverted-U); Mixed plateaus past 100
- **Implication**: jointly tune dataset size and training duration — scaling data alone is not enough

Finding 3 — Token limits can bind before data does



- **Near baseline**: all configs cluster near 29.4% — scaling has little impact on accuracy
- **Token cap binds**: majority of completions exhaust budget before producing parseable output
- **Verbose domains**: token budget allocation is more binding than data size

Three Findings

1. Composition can outweigh volume

- Counting: 100 mixed match 500 easy-only — 5× sample efficiency
- Spatial: Mixed configs hold steady, exceeding Easy-only at small budgets
- Evidence: Counting + Spatial

2. Scaling data alone can be ineffective

- Fixed step budget: larger datasets receive fewer updates per example
- Counting-Mixed declines past 100; Spatial-Easy inverted-U at 200
- Evidence: Counting + Spatial

3. Token limits can bind before data does

- Verbose domains exhaust rollouts before producing parseable output
- Suppressed reward signal masks any composition effect
- Evidence: Graph Reasoning

Interacting constraints, not a single "complexity" knob

Limitations & Future Work

Limitations

- Single base model (Qwen3-4B) with LoRA only
- No multi-seed repetitions — robustness inferred from consistency across 18 configs
- Procedural data does not capture full natural-language complexity
- Hypothesis-generating, not universal scaling laws

Future Work

- Validate data composition trends at larger model and compute scales
- Test transfer to natural-language reasoning benchmarks
- Develop budget-aware RLVR theory — joint treatment of compute, token limits, and reward sparsity
- Mix existing datasets and/or extend procedural data generators to new domains

Summary

- **Procedural datasets** enable controlled study of RLVR scaling
- **Composition matters under fixed budgets** — but interacts with token limits and step counts
- Three findings: composition > volume · scaling data alone can be ineffective · token limits can bind first

Paper: Learning from Less: Measuring the Effectiveness of RLVR in Low Data and Compute Regimes

MLSys 2026

Thank you!