

DeepSeekMath: Pushing the Limits of Mathematical Reasoning (GRPO)

2026-02-26

Paper Info

- **Authors:** Shao et al.
- **Link:** [arXiv](#)
- **Read date:** 2026-02-26

Summary

Introduces GRPO (Group Relative Policy Optimization) as a variant of PPO that avoids learning a separate value function. Instead of a critic, it estimates baselines from a *group* of sampled outputs for the same prompt. This dramatically simplifies the RL pipeline.

Key Ideas

1. **No critic needed** — sample G completions per prompt, use the group mean reward as baseline. This removes an entire model from the training loop.
2. **RLVR (RL with Verifiable Rewards)** — works best when you have binary/verifiable correctness signals (math proofs, code execution). Directly applicable to our clinical task verification setup.
3. **Outcome supervision vs process supervision** — they show outcome-level rewards work surprisingly well when paired with GRPO, even without step-level annotations.

Relevance to MedEnv

- Our FHIR action verification gives us exactly the kind of verifiable reward GRPO is designed for
- The “no critic” property is huge for us — training a value function over clinical trajectories would be extremely expensive
- Group sampling maps well to generating multiple clinical plans and rewarding the best ones

Critiques / Limitations

- Group size G is a key hyperparameter they don’t ablate thoroughly — matters a lot for sample efficiency
- Unclear how well this extends to very long trajectories (their math tasks are relatively short)
- The verifiable reward assumption is strong; many clinical tasks have ambiguous correctness

Quotes to Remember

(paraphrase) GRPO replaces the value model baseline with the group mean, reducing memory and compute by ~50% compared to PPO.