Executive Summary

Stable AI Training: Breakthrough Algorithm for Better Language Models

The Problem

Training large AI language models using reinforcement learning has been plagued by critical stability issues. Current state-of-the-art algorithms like GRPO often cause catastrophic model collapse during training, where the model becomes permanently damaged and unusable. This instability prevents companies from pushing the boundaries of AI capabilities through continued training, limiting the development of more sophisticated AI systems that can solve complex problems in mathematics, programming, and reasoning.

The Breakthrough

Researchers developed **Group Sequence Policy Optimization (GSPO)**, a novel reinforcement learning algorithm that fundamentally redesigns how AI models learn from feedback. Unlike previous approaches that apply corrections at the individual token level, GSPO operates at the sequence level—evaluating and optimizing entire responses as complete units. This aligns the learning process with how AI models are actually evaluated and rewarded, creating a more stable and efficient training system.

How It Works

GSPO introduces a sequence-level importance ratio that measures how much a complete response deviates from what the model would typically generate. By applying optimization and corrections to entire responses rather than individual words, GSPO eliminates the high-variance noise that causes model collapse. The algorithm achieves **remarkably higher training efficiency** than previous methods, delivering better performance with the same computational resources. Testing showed GSPO maintains stable training throughout the entire process, while successfully improving model capabilities on mathematical reasoning (AIME'24), coding (LiveCodeBench), and competitive programming (CodeForces) benchmarks.

Why This Matters

This breakthrough enables the continued scaling of AI models to new levels of capability. With GSPO, companies can now train larger, more powerful AI systems without fear of catastrophic failure during training. The algorithm has already contributed to significant improvements in Qwen3 models and solves critical challenges for Mixture-of-Experts (MoE) models, which are essential for efficient AI deployment at scale. This opens the door to AI systems that can tackle increasingly sophisticated problems requiring deeper reasoning and longer responses.

The Business Opportunity

GSPO creates new possibilities for developing enterprise-grade AI systems with enhanced reasoning capabilities while reducing training costs and infrastructure complexity. The algorithm’s stability and efficiency improvements make it possible to build more reliable AI products and services that can handle complex, multi-step tasks across various industries from software development to scientific research.