Executive Summary

Stable Training for AI Supermodels: The GSPO Breakthrough

The Problem

Training today's largest AI models through reinforcement learning (RL) has hit a critical stability wall. Current state-of-the-art algorithms like GRPO cause catastrophic model collapse when training massive models like those with 30+ billion parameters, making continued improvement impossible and wasting enormous computational resources.

The Breakthrough

Researchers developed **Group Sequence Policy Optimization (GSPO)**, a fundamentally new approach that shifts from token-level to sequence-level optimization. Instead of weighting individual words differently, GSPO treats entire responses as unified units, eliminating the unstable gradient estimates that plague current methods.

How It Works

GSPO aligns the unit of optimization with the unit of reward—both apply to complete responses rather than individual words. This elegant match removes high-variance noise that accumulates during training, enabling **stable and continuous improvement** even for massive 30-billion parameter models. The algorithm also naturally resolves training instabilities in Mixture-of-Experts (MoE) architectures without requiring complex workarounds.

Why This Matters

This breakthrough enables reliable training of next-generation AI models that can solve competition-level mathematics and complex programming challenges. It removes a fundamental barrier preventing AI models from becoming more capable through continued learning, opening the door to AI systems that can tackle increasingly sophisticated reasoning tasks.

The Business Opportunity

GSPO provides the stable foundation needed for scaling AI capabilities reliably, enabling companies to invest confidently in large-scale RL training without fear of catastrophic failure. This accelerates the development of advanced AI products and services while reducing wasted computational resources and training costs.