Detailed Breakdown

The Problem

Reinforcement learning training of large language models faces fundamental stability challenges that prevent scaling to more capable systems. Current state-of-the-art algorithms like GRPO exhibit severe instability when training massive models, often resulting in catastrophic and irreversible model collapse. The core issue stems from GRPO's misapplication of importance sampling weights at the token level, which introduces high-variance training noise that accumulates over longer sequences and is amplified by the clipping mechanism. This creates a compounding effect where small errors at individual token positions accumulate throughout the response, ultimately leading to complete model failure. The problem becomes particularly acute with larger models, sparse architectures like Mixture-of-Experts (MoE), and longer response lengths—exactly the conditions needed for advanced AI capabilities. Once model collapse occurs, it's often impossible to resume training even by reverting to previous checkpoints and carefully tuning hyperparameters.

The training instability has practical consequences for AI development. Models require massive rollout batch sizes to utilize hardware efficiently, necessitating partitioning data into mini-batches for gradient updates. This creates an off-policy learning setting where corrections become essential, but the current approach of applying these corrections at the token level is fundamentally flawed. As response lengths increase to tackle more complex problems, the instability worsens, creating a barrier to developing AI systems capable of sophisticated reasoning and problem-solving.

The Innovation

Group Sequence Policy Optimization (GSPO) represents a paradigm shift in reinforcement learning for language models by moving from token-level to sequence-level optimization. The key insight is that since rewards are granted to entire sequences, off-policy correction should also occur at the sequence level to maintain theoretical consistency.

- **Sequence-level importance ratios**: GSPO defines importance ratios based on complete sequence likelihood rather than individual token probabilities, aligning with the fundamental principle of importance sampling theory

- **Length-normalized optimization**: The algorithm normalizes importance ratios by response length to control variance and maintain consistent numerical ranges across different response lengths
- **Theoretical grounding**: Unlike GRPO's ill-posed objective, GSPO's approach is mathematically sound and eliminates the high-variance noise that causes model collapse

The fundamental difference lies in how gradients are weighted. In GRPO, tokens receive unequal weights based on their individual importance ratios, which can vary dramatically and accumulate unpredictably. GSPO weights all tokens in a response equally, using a sequence-level importance factor that provides stable, consistent training signals. This approach matches the unit of optimization (sequences) with the unit of reward (sequences), creating a coherent learning process.

How It Works

GSPO implements a sequence-level reinforcement learning objective through several key components:

1. **Sequence-level importance ratio calculation**: The algorithm computes the importance ratio as the geometric mean of token-level importance ratios across the entire response:

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}\right)^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \log\frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x,y_{i,<t})}\right)$$

2. **Group-based advantage estimation**: Similar to GRPO, GSPO computes advantages by normalizing rewards within groups of responses to the same query:

$$\widehat{A}_i = \frac{r(x,y_i) - \text{mean}(\{r(x,y_i)\}_{i=1}^G)}{\text{std}(\{r(x,y_i)\}_{i=1}^G)}$$

3. **Sequence-level clipping**: Instead of clipping individual tokens, GSPO applies clipping to entire responses, excluding overly "off-policy" samples from gradient estimation:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^G \min\left(s_i(\theta)\widehat{A}_i, \text{clip}(s_i(\theta),1-\varepsilon,1+\varepsilon)\widehat{A}_i\right)\right]$$

4. **Gradient computation**: The gradient weights all tokens in a response equally by the sequence-level importance ratio, eliminating the instability caused by unequal token weights in GRPO.

5. **GSPO-token variant**: For multi-turn RL scenarios requiring finer-grained control, GSPO-token allows token-wise advantage

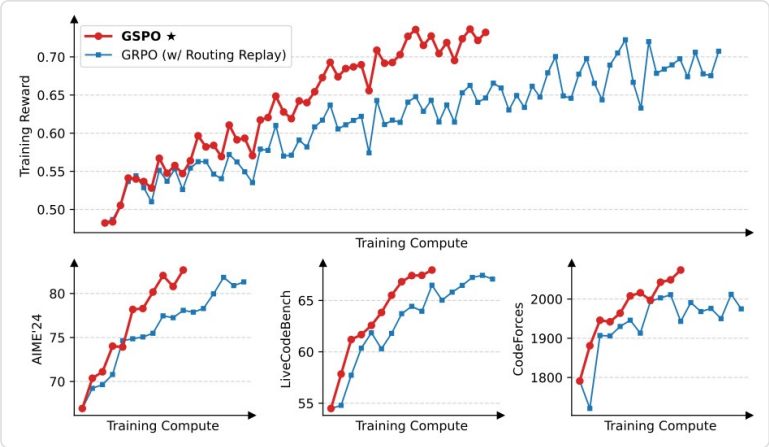
customization while maintaining the same sequence-level importance weighting.

The length normalization in the importance ratio calculation prevents dramatic fluctuations from a few tokens and ensures consistent behavior across responses of different lengths. The clipping ranges typically differ by orders of magnitude from GRPO due to the distinct importance ratio definitions.

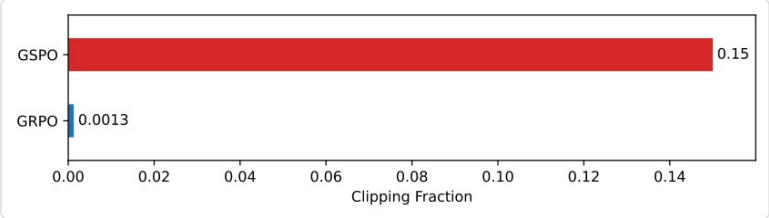
Key Results

Experimental evaluation demonstrates GSPO's significant superiority over GRPO across multiple dimensions:

- **Training stability**: GSPO maintains stable training throughout entire training cycles, while GRPO often experiences catastrophic collapse requiring intervention
- **Training efficiency**: GSPO achieves **remarkably higher training efficiency** than GRPO, delivering better performance with the same computational resources and query consumption
- **MoE model training**: GSPO **eliminates the need for complex Routing Replay strategies** in Mixture-of-Experts training, fundamentally resolving expert-activation volatility issues
- **Benchmark performance**: GSPO-trained models show continuous improvement on mathematical reasoning (AIME'24), coding (LiveCodeBench 202410-202502), and competitive programming (CodeForces Elo Rating) benchmarks
- **Infrastructure simplification**: GSPO enables **direct use of inference engine likelihoods** for optimization, avoiding the need for training engine recomputation
- **Clipping efficiency**: Despite clipping **two orders of magnitude more tokens** than GRPO, GSPO still achieves superior training efficiency, indicating more reliable learning signals



Training curves showing GSPO's superior performance and stability compared to GRPO



Clipping fractions comparison showing GSPO clips significantly more tokens but achieves better efficiency

The evaluation used a cold-start model fine-tuned from Qwen3-30B-A3B-Base, with training data partitioned into four minibatches for gradient updates. GSPO used clipping ranges of 3e-4 and 4e-4, while GRPO required carefully tuned ranges of 0.2 and 0.27.

Practical Applications

Advanced Mathematical Reasoning

GSPO enables stable training of AI models capable of solving competition-level mathematics problems. The algorithm's stability allows for longer reasoning chains and more complex mathematical proofs, making it valuable for educational platforms, mathematical research tools, and automated theorem proving systems.

Code Generation and Programming

The improved training stability directly benefits code generation models, enabling them to handle more complex programming tasks and longer code sequences. This applies to software development tools, automated programming assistants, and coding education platforms that require

sophisticated reasoning about algorithms and data structures.

Enterprise AI Deployment

For businesses deploying AI systems at scale, GSPO's ability to stabilize MoE model training without additional strategies like Routing Replay reduces infrastructure complexity and costs. This makes it easier to deploy efficient, large-scale AI systems in enterprise environments.

Multi-turn Conversational AI

The GSPO-token variant provides the fine-grained control needed for sophisticated conversational AI systems that must maintain context and coherence across multiple turns of dialogue. This applies to customer service chatbots, personal assistants, and interactive learning systems.

Research and Development

GSPO provides a more reliable foundation for AI research, allowing researchers to push the boundaries of model capabilities without worrying about training instability. This accelerates progress in developing more intelligent and capable AI systems.

Limitations & Considerations

- **Clipping range sensitivity**: GSPO requires different clipping ranges (typically 3e-4 to 4e-4) compared to GRPO, necessitating careful hyperparameter tuning for optimal performance
- **Length normalization requirements**: The sequence-level importance ratios require length normalization to prevent variance issues, adding computational overhead
- **MoE model compatibility**: While GSPO eliminates the need for Routing Replay, it still requires careful implementation to ensure proper handling of sparse model architectures
- **Implementation complexity**: The algorithm requires careful implementation of sequence-level computations, which may be more complex than token-level approaches
- **Benchmark dependency**: Performance gains may vary depending on the specific benchmarks and evaluation metrics used, requiring validation for target applications

What This Means for Builders

Immediate Opportunities

Developers can immediately leverage GSPO for more stable reinforcement learning training of language models, particularly for applications requiring long responses or complex reasoning. The algorithm’s stability makes it possible to train models for sophisticated tasks like mathematical problem-solving, code generation, and multi-turn dialogue without the risk of catastrophic failure. Companies working on MoE models can eliminate complex stabilization strategies, reducing development complexity and infrastructure costs.

Implementation Pathway

GSPO can be implemented as a drop-in replacement for GRPO in existing RL training pipelines, with modifications to compute sequence-level rather than token-level importance ratios. The algorithm is already integrated into the latest Qwen3 models, demonstrating its practical applicability at scale. Implementation requires careful attention to clipping range tuning and length normalization, but the core algorithmic changes are straightforward. The availability of both sequence-level and token-level variants provides flexibility for different use cases.

Strategic Implications

GSPO represents a fundamental advance in AI training methodology that could reshape how large language models are developed and deployed. The algorithm’s stability and efficiency improvements make it possible to continue scaling AI models to new levels of capability, potentially enabling breakthroughs in areas requiring sophisticated reasoning and long-form content generation. This could accelerate progress toward more general artificial intelligence capabilities and open new applications for AI in scientific research, education, and complex problem-solving domains.

Cost Optimization

By eliminating the need for complex stabilization strategies like Routing Replay and enabling more efficient use of computational resources, GSPO can significantly reduce the cost of training large language models. The algorithm’s superior sample efficiency means better results with less training data and computation, while the infrastructure simplification reduces engineering overhead. These cost benefits make advanced AI capabilities more accessible to organizations with limited resources, potentially democratizing access to cutting-edge AI technology.