UNDERSTANDING TRAINING STABILITY IN LLM REINFORCEMENT LEARNING: A SURVEY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has become a key paradigm for enhancing capabilities of large language models, enabling remarkable advances in mathematical reasoning, code generation, and long-horizon tasks. However, compared with supervised fine-tuning, RL optimization is substantially more unstable, frequently suffering from loss spikes and policy collapse. Despite extensive surveys on RL algorithms and alignment, training stability in LLM RL has not been systematically reviewed. This paper provides a unified survey of stabilization techniques from an optimization pipeline perspective. We categorize existing methods into two stages: **gradient generation**, covering data sampling, importance sampling, and credit assignment; and **gradient propagation**, covering gradient regulation and numerical/system alignment. Based on this unified perspective, we systematically review representative methods, analyze their underlying stabilization mechanisms, summarize their advantages and limitations, and identify open challenges and promising directions toward stable and scalable RL for LLMs. We hope this survey serves as both a comprehensive reference for future research.

1 INTRODUCTION

Reinforcement learning (RL) has become a key paradigm for aligning large language models (LLMs) and enhancing their reasoning capabilities (Jaech et al., 2024; Guo et al., 2025), leading to remarkable advances across a wide range of tasks, including mathematical reasoning (Guan et al., 2025), code generation (Li et al., 2022), and complex tasks (Hu et al., 2026). Unlike supervised fine-tuning (SFT), which learns from static demonstrations, RL continuously optimizes policies through interaction with reward signals, allowing models to explore novel reasoning trajectories and progressively discover behaviors that cannot be obtained from imitation alone. Consequently, RL has become an indispensable component of modern LLM post-training and is widely regarded as one of the key technologies toward more capable general-purpose AI systems.

Despite the remarkable progress of reinforcement learning for LLMs, achieving stable optimization remains significantly more challenging than SFT. The sparse reward signals (Yang et al., 2026a; Ding et al., 2026), long-horizon credit assignment (Kim et al., 2026; Gong et al., 2026), importance sampling correction (Zhang et al., 2026b), and operator errors (Qiu & Yao, 2025) collectively make RL optimization highly unstable. In practice, RL training frequently exhibits phenomena such as gradient spikes and policy collapse. As model sizes and context lengths increase and distributed mixed-precision systems become more complex, these instabilities increase training costs and degrade optimization efficiency. Consequently, understanding where optimization instability originates, how it propagates throughout the training pipeline, and how it can be systematically mitigated has become a central challenge in LLM reinforcement learning.

Although numerous surveys (Yu et al., 2026b; Liu et al., 2026b; Karydas et al., 2026; Wu et al., 2025; Zhang et al., 2025b; Kumar et al., 2025b) review reinforcement learning for LLMs from the perspectives of algorithms, reward modeling, alignment, and downstream applications, training stability has not yet been systematically surveyed as an independent research problem. Existing discussions don't reveal how these methods relate to different stages of the RL optimization process. As a result, researchers still lack a unified perspective for understanding the origins of instability and the relationships among existing stabilization techniques.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

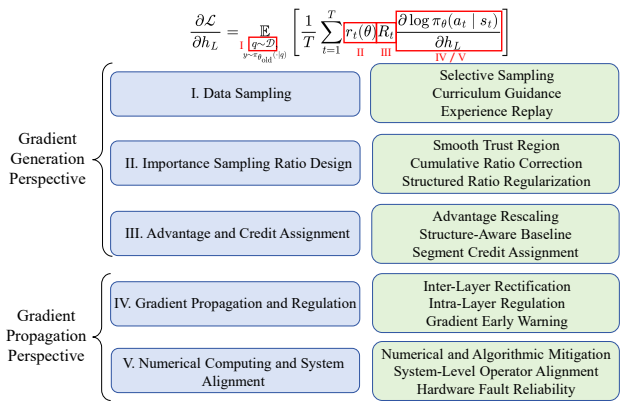

Figure 1: This constitutes the taxonomy framework of our survey. We categorize existing methods into two primary paradigms and five core sub-categories, with the underlying approaches within each sub-category systematically classified into three distinct dimensions.

To bridge this gap, this survey presents a comprehensive review of training stability in LLM reinforcement learning from a unified optimization pipeline perspective. Rather than categorizing methods by algorithmic variants, we organize existing stabilization techniques according to where instability is introduced and how it propagates during optimization. Specifically, we divide existing methods into two complementary stages: **gradient generation**, including data sampling and replay, importance sampling ratio design, and fine-grained credit assignment; and **gradient propagation**, including gradient regulation and numerical/system-level stabilization. This gradient-centric taxonomy unifies seemingly disparate approaches into a coherent optimization framework, highlights their underlying connections, and provides practical guidance for developing stable, scalable, and reliable reinforcement learning algorithms for LLMs.

2 Preliminary: Reinforcement Learning for LLMs

In this section, we first elucidate the forward propagation workflow of LLMs alongside the optimization objectives of reinforcement learning. Building upon this foundational formulation, we derive the complete optimization pipeline of policy gradients. Then we systematically categorize existing methods based on the structural characteristics delineated by these mathematical formulations.

2.1 LLM Reinforcement Learning Objective

During the alignment or post-training phase via reinforcement learning (e.g., RLVR (Guo et al., 2025), PPO (Schulman et al., 2017), or GRPO (Shao et al., 2024)), the training workflow typically alternates between two stages: *Rollout* (data collection) and *Update* (parameter optimization). Assume that a prompt q is sampled from a dataset $\mathcal{D}$. Let $s_t = (q, a_{<t})$ denote the input at timestep t, $a = (a_1, \cdots, a_T)$ denote the complete generated token sequence, where T is the generated sequence length. The verifier assigns a reward score R_t for each token. The reinforcement learning objective for LLMs is to maximize the expected return (equivalently minimize its negative):

$$\mathcal{L} = -\mathbb{E}_{q\sim\mathcal{D}, a\sim\pi_\theta^{\text{train}}(\cdot|q)}\left[\frac{1}{|a|}\sum_{t=1}^{|a|} R_t\right], \tag{1}$$

where $\pi_\theta^{\text{train}}$ denotes the trainable language model, and R_t denotes the reward of token a_t. However, to reduce rollout latency in large-scale training, trajectory sampling is typically performed on a dedicated inference engine. Under this paradigm, the rollout policy parameters $\pi_{\theta_{\text{old}}}^{\text{rollout}}$ on the inference engine are synchronized periodically after several updates to the parameters $\pi_\theta^{\text{train}}$ on the training engine. As a result, the sampling policy and training policy are not perfectly aligned due to discrepancies in parameter versions, parameter precision, and operator implementations across the

training and inference engines, resulting in the **train-rollout mismatch**. To address this discrepancy, a resampling mechanism is applied to compensate for the discrepancy between the two policies:

$$\mathcal{L} = -\mathbb{E}_{q\sim\mathcal{D},a\sim\pi^{\text{rollout}}_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{|a|}\sum_{t=1}^{|a|} r_t(\theta)R_t\right], r_t(\theta) = \frac{\pi^{\text{train}}_{\theta}(a_t|s_t)}{\pi^{\text{rollout}}_{\theta_{\text{old}}}(a_t|s_t)}. \tag{2}$$

For notational simplicity, we abbreviate $\pi^{\text{train}}_{\theta}, \pi^{\text{rollout}}_{\theta_{\text{old}}}$ as $\pi_\theta, \pi_{\theta_{\text{old}}}$ unless otherwise specified. To propagate the scalar loss signal $\mathcal{L}$ back to the shallow layers of the network, we derive the gradient of the loss function with respect to the i-th layer's output activation h_i:

$$\frac{\partial\mathcal{L}}{\partial h_i} = -\mathbb{E}_{q\sim\mathcal{D},a\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{|a|}\sum_{t=1}^{|a|} r_t(\theta)R_t\frac{\partial\log\pi_\theta(a_t|s_t)}{\partial h_i}\right]. \tag{3}$$

Equation 3 reveals that the policy gradient is jointly determined by three components: the sampling distribution, the importance sampling ratio, and the reward. Any modification to these components changes the gradient estimator before backpropagation begins. Therefore, methods such as data sampling, importance weighting, and credit assignment are categorized as gradient generation methods. Once the gradient estimator has been constructed, the resulting optimization signal is propagated according to the chain rule. Consequently, gradient and layer-wise regulation, and numerical stabilization methods belong to the gradient propagation stage. The discussion about instability and each component is in Appendix A.2.

To ensure a rigorous and objective taxonomy, we classify existing methods based on their direct intervention mechanism. The details are in Appendix A. In the subsequent sections, we provide a comprehensive analysis of the methods within each category.

3 Gradient Generation Perspective

3.1 I. Data Sampling

In reinforcement learning for LLMs, optimization relies predominantly on on-policy trajectories collected from the current policy. From the perspective of gradient generation, data sampling determines which trajectories contribute to the policy-gradient estimator in Eq. 3 and therefore directly influences the quality of optimization signals before backpropagation begins.

A. Selective Sampling: Selective sampling methods aim to improve optimization stability by allocating computational resources only to trajectories that are expected to produce informative policy gradients.

History-aware utility estimation: A straightforward strategy estimates the future optimization value of a sample from its historical training statistics. GRESO (Zheng et al., 2026) maintains historical reward records for each prompt and predicts whether newly generated trajectories are likely to provide effective learning signals, thereby avoiding repeated optimization on saturated samples.

Variance-aware utility estimation: Recent studies increasingly regard reward variance as a better proxy for optimization value than the reward itself. DAPO (Yu et al., 2026a) filters informative rollouts and removes all-correct or all-incorrect samples to avoid zero-gradient updates. MMR1 (Leng et al., 2025) theoretically proves that reward variance provides a lower bound on gradient magnitude and therefore prioritizes high-variance samples during optimization. Furthermore, VCRL (Jiang et al., 2025) interprets reward variance as an indicator of model-relative difficulty and performs online filtering to maintain informative training batches throughout optimization.

B. Curriculum Guidance: Curriculum-guided methods retain all training samples but dynamically regulate when and how strongly different trajectories participate in optimization. The central idea is to match supervision difficulty with the evolving capability of the policy, thereby preventing unstable optimization caused by excessively difficult or overly trivial samples.

Reward-statistics-guided curriculum: EBPO (Han et al., 2026) recalibrates local rewards using empirical Bayes shrinkage, reducing estimation variance caused by limited rollout samples while constructing smoother optimization trajectories.

Confidence-guided curriculum: Instead of relying on external verifiers, VI-CuRL (Cai & Sugiyama, 2026) measures model confidence from token probabilities and gradually shifts optimization from high-confidence to low-confidence samples, suppressing destructive gradient variance during early optimization.

Difficulty-adaptive curriculum: Another line of work explicitly aligns sample difficulty with policy capability. GHPO (Liu et al., 2025) injects ground-truth reasoning trajectories into difficult problems to alleviate sparse-reward optimization, whereas DIVA-GRPO (Gao et al., 2026) dynamically estimates relative sample difficulty and applies adaptive data augmentation to maintain stable reward variance throughout optimization.

C. Experience Replay: Experience replay addresses the low sample efficiency of strictly on-policy optimization by allowing informative trajectories discovered during previous rollouts to participate in future policy updates. Replay methods improve stability primarily through sample reuse, differing mainly in the type of optimization knowledge they preserve.

Trajectory replay: Several methods directly replay high-value trajectories to improve data utilization. SPO (Cohen et al., 2025) integrates online and historical trajectories through a learnable soft action-value function. ExGRPO (Zhan et al., 2025) and RePO (Li et al., 2025a) further replay trajectories according to correctness, predictive uncertainty, and informational diversity, substantially improving sample efficiency while preventing catastrophic policy collapse.

Intermediate-state replay: Rather than storing complete trajectories, RRL (Dou et al., 2025) preserves promising intermediate reasoning states encountered during failed explorations. Once the policy becomes sufficiently capable, optimization resumes from these intermediate states, enabling exploration beyond reasoning paths that would otherwise be permanently abandoned.

Structured knowledge replay: Instead of replaying raw trajectories, HiExp (Hao et al., 2026) organizes successful experiences into a hierarchical knowledge repository and reuses structured reasoning patterns as regularization signals, reducing optimization variance introduced by long-tail reward distributions.

Discussion: Recent research has evolved from local trajectory selection to optimization lifecycle management, evolving from online filtering and difficulty-aware sampling to historical experience reuse. However, these methods depend on accurate trajectory valuation and may suffer from sampling bias, reduced diversity, or stale experiences. They regulate optimization signal generation but cannot address instability from policy divergence, noisy rewards, or gradient propagation.

3.2 II. Importance Sampling Ratio Design

The importance sampling (IS) ratio plays a fundamental role in reinforcement learning for LLMs by correcting the distribution mismatch between the rollout policy and the training policy. From the perspective of gradient generation, the IS ratio directly determines how each sampled trajectory contributes to the policy-gradient estimator in Eq. 3. Consequently, redesigning the IS ratio does not alter the sampled data itself, but reshapes the optimization signal before gradient propagation begins.

A. Smooth Trust Region: Smooth trust-region methods improve optimization stability by replacing rigid clipping with continuous or adaptive policy constraints. Although different methods adopt different formulations, they all pursue the same objective: preserving optimization stability while maintaining useful exploratory gradients.

Distribution-level smoothing: Instead of directly modifying the clipping function, several methods first smooth the policy distribution itself before computing IS ratios. GR-PSPO (Dwyer et al., 2025) interpolates the current policy toward the behavior policy to suppress extreme ratios, while GMPO (Zhao et al., 2025) aggregates trajectory-level policy updates through a geometric averaging. Complementary theoretical analyses (Wu et al., 2026; Xu, 2025) further show that entropy regularization and KL-based divergence constraints provide equivalent interpretations of smooth trust-region optimization.

Continuous trust-region weighting: Another line of work replaces piecewise clipping with continuous weighting functions. Representative methods include SAPO (Gao et al., 2025), SSPO (Glazyrina et al., 2026), MHPO (Wang et al., 2026a), and TROLL (Becker et al., 2025), which respectively

employ adaptive weighting, Cauchy-style decay, differentiable bounded mappings, and localized trust-region projection. Although their formulations differ, they share the same objective of ensuring gradients decay smoothly as policy divergence increases.

Adaptive trust-region allocation: Recent studies further argue that trust regions should evolve together with optimization statistics. ESS-based optimization (Fakoor et al., 2026) adjusts trust-region size according to effective sample size. BandPO (Li et al., 2026b) adapts clipping thresholds based on action confidence, RVPO (Luo et al., 2026) minimizes ratio variance, and ESPO (Sheng et al., 2025) assigns different trust regions to high- and low-entropy tokens. These methods dynamically allocate optimization flexibility according to the current training state.

B. Cumulative Ratio Correction: Unlike smooth trust-region methods, these methods recognize that autoregressive generation naturally introduces sequential dependency between token-level IS ratios. Rather than correcting each token independently, these approaches explicitly compensate for accumulated distribution shifts along reasoning trajectories.

Prefix-aware correction: Several methods estimate current importance ratios using historical prefix information. CTPO (Zhang et al., 2026b) progressively tightens clipping thresholds according to accumulated prefix ratios, while MinPRO (Lei et al., 2026) employs the minimum prefix ratio as a variance-reduction correction factor.

Future-aware correction: Instead of relying exclusively on historical information, NFPO (Yoon et al., 2026) incorporates future cumulative ratios to estimate the downstream influence of current decisions, thereby achieving a more balanced bias-variance trade-off.

Sequence-level correction: Another line of work treats complete sequences rather than individual tokens as optimization units. Li et al. (2025b) discard entire trajectories whenever any token violates trust-region constraints, whereas GSPO (Zheng et al., 2025b) performs importance sampling correction at the sequence level to reduce the instability.

Staleness-aware correction: To alleviate severe ratio drift during off-policy optimization, μ-GRPO (Tian et al., 2026) explicitly models trajectory staleness and asymmetrically adjusts clipping boundaries according to the age of collected trajectories, mitigating severe ratio drift.

C. Structured Ratio Regularization: Instead of redesigning trust regions or sequential corrections, these methods exploit optimization structures to design task-dependent ratio regularization.

Advantage-aware optimization: Several methods distinguish positive and negative advantages during importance weighting. TOPR (Roux et al., 2025), BAPO (Xi et al., 2025), DISPO (Karaman et al., 2026), and ASPO (Wang et al., 2026b) assign asymmetric clipping or weighting strategies according to the advantage sign, preventing negative trajectories from dominating policy optimization while preserving beneficial exploration.

Structure-aware divergence regularization: Another line of work redesigns the divergence measure rather than directly modifying IS ratios. KPop (Guo et al., 2026), DPPO (Qi et al., 2026), and DRPO (Yao et al., 2026) respectively replace conventional KL divergence with binary KL divergence, total variation distance, or smooth quadratic penalties, enabling finer-grained policy constraints.

Architecture-aware correction: For Mixture-of-Experts architectures, routing inconsistency becomes an additional source of ratio instability. RSPO (Zhang et al., 2025a), MiniRL (Zheng et al., 2025a), and GSPO (Zheng et al., 2025b) explicitly model routing dynamics or eliminate unstable routing replay, substantially improving optimization robustness under dynamic expert selection.

Task-specific ratio design: Several methods further customize IS ratios for specific optimization scenarios. Chen et al. (2026) employ Hőlder means to balance variance reduction and learning signal strength, while Search-R1 (Jin et al., 2025) masks retrieved tokens to prevent noisy external knowledge from destabilizing policy optimization.

Discussion: Recent research has evolved from static ratio correction to adaptive optimization control. Despite their effectiveness, these methods are inherently limited to policy divergence and may restrict policy improvement, introduce hyperparameter sensitivity, and incur implementation cost. Moreover, they cannot address instability caused by noisy rewards, inaccurate credit assignment, or gradient propagation, making them one component of a broader stabilization framework.

3.3 III. Advantage and Credit Assignment

Credit assignment determines how reward signals are propagated by modifying the reward or advantage term in Eq. 3. Consequently, it directly affects both the bias and variance of policy-gradient estimation. Existing policy optimization algorithms (e.g., PPO and GRPO) typically assign identical sequence-level rewards to every token, neglecting the heterogeneous contributions of different generation steps.

A. Advantage Rescaling: Advantage rescaling methods improve optimization by reallocating gradient magnitudes among generated tokens instead of assigning identical optimization strength throughout an entire sequence. Existing approaches mainly differ in which optimization property is explicitly calibrated.

Gradient magnitude calibration: A straightforward strategy directly rescales token-wise advantages to balance optimization magnitudes. FlowRL (Zhu et al.) transforms rewards into a normalized target distribution through a learnable partition function, while STAPO (Liu et al., 2026a) suppresses unstable pseudo-labeled tokens whose probabilities and local entropies indicate unreliable optimization signals. Complementary methods (Yang et al., 2025b; Huang et al., 2026b) further rebalance updates between high- and low-probability tokens by controlling optimization magnitude.

Direction-aware optimization: Instead of modifying only update magnitudes, several methods explicitly decouple optimization direction from optimization strength. For example, OWPO (Yang et al., 2026a) dynamically reallocates advantages according to the relative progress between the training and rollout policies, amplifying updates when training policy lags behind while preventing unnecessary suppression once the policy has surpassed previous behavior.

Entropy-aware advantage weighting: Recent studies further argue that policy entropy provides a better indicator of optimization reliability. EMPG (Wang et al., 2025) strengthens confident correct predictions while suppressing uncertain exploratory tokens. EntroPIC (Yang et al., 2025a) dynamically adjusts coefficients for positive and negative samples according to entropy deviation from a target distribution, whereas DynaMO (Fang et al., 2026) monitors policy entropy and compensates gradients for informative high-confidence samples while suppressing updates that induce excessive entropy fluctuations.

B. Structure-Aware Baseline: Unlike advantage rescaling, structure-aware baseline methods reduce policy-gradient variance through more accurate reward baseline estimation. Existing approaches mainly differ in how token contributions are modeled.

Uniform statistical baselines: This method constructs simplified variance-reduction baselines based on statistical assumptions. OPO (Hao et al., 2025) assumes approximately orthogonal token gradients with similar magnitudes, leading to a length-aware average reward estimator that effectively suppresses optimization variance. GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024) reduce advantage variance by constructing statistical baselines from uniformly sampled responses within the same prompt.

Variance-optimal baselines: Recent studies abandon uniform assumptions and explicitly optimize baseline estimation for variance reduction. OBLR-PO (Huang et al., 2025b) derives the theoretically optimal non-uniform baseline that minimizes gradient variance while simultaneously providing adaptive learning-rate guidance. OTB (Li et al., 2026a) further estimates token importance from cumulative gradient variance and efficiently approximates gradient norms using output probabilities.

C. Segment Credit Assignment: Instead of assigning rewards independently to every token, these methods propagate sparse rewards according to higher-level reasoning structures. Their common objective is to shorten effective optimization horizons by assigning supervision to semantically meaningful reasoning components.

Hierarchical temporal decomposition: This method divides long reasoning trajectories into shorter semantic segments. Kim et al. (2026) decompose long-horizon objectives into intermediate subgoals, substantially reducing credit-assignment difficulty by shortening optimization horizons.

Counterfactual credit assignment: IBPO (Ding et al., 2026) estimates process-level contributions by comparing counterfactual reasoning trajectories, enabling more informative reward allocation based on causal differences between reasoning paths.

Semantic-segment aggregation: Instead of estimating advantages token by token, SAE (Gong et al., 2026) aggregates advantages only at semantic segment boundaries, filtering noisy intermediate value estimates and substantially reducing the estimation bias accumulated by token-level GAE.

Discussion: Recent research has evolved from statistical variance reduction to reasoning-aware supervision, shifting from advantage and baseline design to process-level credit assignment. Despite their effectiveness, these methods remain constrained by supervision quality. Outcome-level rewards cannot fully capture intermediate reasoning, whereas process-level supervision often requires costly annotations. Moreover, inaccurate reward redistribution may introduce optimization bias, making reward design complementary to policy divergence control and gradient regulation.

4 Gradient Propagation Perspective

4.1 IV. Gradient Propagation and Regulation

Once the policy-gradient estimator has been constructed through data sampling, importance weighting, and reward design, optimization enters the gradient propagation stage, where the generated optimization signal is propagated backward through the Transformer according to the chain rule. Methods in this stage focus on how optimization signals evolve during backpropagation.

A. Inter-Layer Rectification: Inter-layer rectification methods improve optimization stability by correcting abnormal gradient propagation across layers. Existing approaches mainly differ in which propagation property is explicitly regulated.

Noise-assisted propagation stabilization: A straightforward strategy introduces perturbations into hidden representations to strengthen the model's ability to resist noise. ALP (Ye et al., 2026) injects learnable perturbations into intermediate representations, effectively suppressing long-tail importance-ratio explosions during long-horizon reasoning.

Gradient-direction rectification: Rather than modifying activation signals, another line of work directly regulates gradient directions across multiple layers. GAC (Xu et al., 2026) monitors the similarity between consecutive policy gradients and projects out stale gradient components to alleviate overshooting during asynchronous RL. Similarly, FPA (Oh et al., 2026) estimates future policy updates through logit extrapolation and adjusts current gradient weights to preserve constructive tokens shared across successful and unsuccessful trajectories.

Low-rank optimization dynamics: A recent study further reveals that policy optimization often evolves within extremely low-dimensional parameter subspaces. Wei et al. (2026) demonstrate that downstream improvements can be approximated by Rank-1 parameter updates, enabling stable optimization through low-rank extrapolation of future parameter trajectories.

B. Intra-Layer Regulation: Unlike inter-layer rectification, intra-layer regulation focuses on maintaining stable gradient statistics within individual network modules. These methods regulate activation scales and gradient statistics inside attention, feed-forward, and residual branches.

Activation-scale regulation: Early studies stabilize optimization by directly controlling activation magnitudes within attention and feed-forward modules while strengthening residual signal propagation, thereby preventing residual collapse in large Transformer models (Takase et al., 2023).

Adaptive gradient clipping: Instead of applying fixed global clipping thresholds, recent methods dynamically estimate appropriate clipping ranges from historical gradient statistics (Huang et al., 2026a). AGGC (Li et al., 2026c) independently traces gradient behavior for different parameter groups through exponential moving averages, whereas ZClip (Kumar et al., 2025a) identifies abnormal gradients using statistical Z-scores and computes temporary clipping thresholds accordingly.

Historical gradient normalization: Rather than clipping abnormal gradients directly, GradientStabilizer (Huang et al., 2025a) preserves instantaneous gradient directions while replacing noisy gradient norms with stable estimates computed from historical optimization statistics, effectively reducing optimization fluctuations without sacrificing directional information.

C. Gradient Early Warning: Instead of correcting unstable gradients after they emerge, gradient early-warning methods proactively monitor optimization dynamics and identify unstable optimization trajectories before training collapse.

Curvature-aware monitoring: Melo et al. (2025) approximates second-order curvature information using only final-layer representations and gradients, enabling real-time identification of unstable optimization trajectories with relatively low computational overhead.

Optimization-state monitoring: Another line of work models optimization as a dynamical system. Zhang et al. (2026c) constructs a low-dimensional meta-state by jointly modeling gradients, losses, and optimizer states, showing that optimization collapse is consistently preceded by coordinated state shifts. Zhang et al. (2026a) monitor response length anomalies and trigger dynamic learning rate decay during gradient noise spikes to prevent policy collapse proactively.

Theoretical instability indicators: Recent theoretical studies further characterize optimization stability from different mathematical perspectives. Ghosh (2026) introduces non-normal operator theory and proposes an eigenvector-condition-number criterion for instability prediction, while Suk & Duan (2025) derive a gradient-gap-based stability condition. These studies establish principled foundations for understanding optimization dynamics beyond empirical observations.

Discussion: Recent research has shifted from reactive stabilization toward optimization-aware regulation, evolving from clipping and normalization to layer-aware propagation and adaptive monitoring. However, these methods operate only during gradient propagation and cannot compensate for errors from gradient generation. Moreover, overly restrictive regulation may suppress useful learning signals, while monitoring increases computational overhead. Therefore, gradient propagation should be combined with improvements in sampling, IS ratio, and reward design for end-to-end stable RL optimization.

4.2 V. Numerical Computing and System Alignment

After policy gradients have been generated and propagated through the network, the final optimization performance ultimately depends on whether these gradients are faithfully executed by the underlying numerical computing system. In modern LLM reinforcement learning, training typically spans heterogeneous hardware platforms, mixed-precision arithmetic, distributed communication, and decoupled rollout-training architectures. Consequently, even theoretically correct optimization algorithms may become unstable due to numerical inconsistencies introduced during implementation.

A. Numerical and Algorithmic Mitigation: Rather than modifying the optimization objective, numerical mitigation methods directly reduce the numerical errors introduced during low-precision computation. Existing approaches mainly differ in which source of numerical bias is explicitly corrected.

Precision-aware numerical computation: A straightforward strategy improves the numerical behavior of existing operators under low precision. Qiu & Yao (2025) analyze the systematic bias introduced by BF16 accumulation and online softmax computation, proposing lightweight numerical corrections within FlashAttention to suppress low-rank numerical degeneration. Similarly, Qi et al. (2025) demonstrate that replacing BF16 with FP16 substantially reduces mantissa-induced rounding errors, thereby narrowing the mismatch between training and inference significantly.

Numerical characterization: Beyond directly increasing numerical precision, recent work has investigated the underlying arithmetic behavior of modern accelerators. MMA-Sim (Xie et al., 2025) develops a bit-accurate reference model for Tensor Cores and Matrix Cores, revealing undocumented arithmetic behaviors that can introduce numerical inconsistencies across GPU architectures and potentially affect training stability and reproducibility.

Quantization-aware optimization: Instead of increasing numerical precision, another line of work redesigns RL optimization for low-bit computation. Jet-RL (Xi et al., 2026) establishes a unified FP8 training and rollout framework through fine-grained block-wise weight quantization together with group-wise activation and gradient quantization. AIS (Zhou et al., 2026) treats the quantization error as dynamic optimization noise and adaptively regulates its impact on gradient updates through importance sampling.

B. System-Level Operator Alignment: Unlike numerical mitigation methods, system-level alignment methods aim to eliminate inconsistencies introduced by heterogeneous execution environ-

ments. Existing approaches mainly focus on aligning numerical behaviors across different stages of training and inference.

Training-inference alignment: Several methods explicitly enforce consistency between rollout sampling and policy optimization. QaRL (Gu et al., 2026) reproduces inference-time quantized computation during training and uses sequence-level dual clipping to suppress corrupted signals. VeXact (Zhong et al., 2026) diagnoses policy drift under identical execution environments, while recent work shows that parallel execution can also introduce mismatches despite identical weights and precision. Zhang et al. (2025c) and GDN (Wang & the TorchTitan team, 2026) address such discrepancies through invariant reduction and batch-invariant operators, respectively.

Batch- and parallelism-invariant execution: Another line of work removes numerical discrepancies by designing execution-invariant operators. He & Lab (2025) develops batch-invariant inference kernels, while SGLang (Team, 2025) integrates deterministic and batch-invariant kernels for key operators. LLM-42 (Gond et al., 2026) instead uses verified speculation to detect and correct nondeterministic outcomes through rollback.

Kernel-level numerical equivalence: Numerical inconsistency can also originate from GPU kernel implementations. Differences in reduction order, accumulation precision, fused operations, or rounding can produce different results despite identical mathematical operations. Yang et al. (2026b) introduce a descriptor for capturing the reduction structure of GEMM operations and reconstruct a bitwise-equivalent Triton implementation of cuBLAS GEMM. Similarly, Cankaya (2026) explores mechanisms for validating numerical equivalence across inference implementations, providing another layer of verification between operator-level computation and end-to-end system alignment.

C. Hardware Fault Reliability: Even with aligned numerical computation, hardware failures can still destabilize optimization. Altenbernd et al. (2026) show that high-order floating-point bit flips during backpropagation can severely corrupt policy gradients, and propose targeted fault detection with selective recomputation of the latest optimization step for efficient fault tolerance.

Discussion: Recent works reflect a shift from numerical correction toward execution consistency, recognizing that small discrepancies from kernels, quantization, and distributed systems can accumulate into optimization bias. However, operator-level methods mainly improve execution fidelity rather than optimization quality. Precision enhancement may increase cost, while strict alignment reduces flexibility, and these methods cannot address instability from poor sampling, noisy rewards, or improper policy updates. They therefore serve as a reliability layer complementary to algorithmic stabilization.

5 Conclusion

This survey provides a comprehensive review of training stability in reinforcement learning for large language models. Rather than organizing prior work by algorithms or applications, we adopt a unified optimization-pipeline perspective, categorizing stabilization methods by where instability arises and how it propagates. Our taxonomy connects representative approaches across gradient generation and propagation, offering a coherent view of the rapidly evolving literature. We hope this survey serves as a comprehensive reference for researchers and practitioners, and inspires future research toward more stable, scalable, and reliable RL training frameworks.

AI use statement

In this work, we used generative AI tools for: 1. design or provide feedback on research methodology or experiments, 2. assist with translation, 3. help develop theoretical models or conceptual frameworks. And 1. support qualitative and thematic data analysis, 2. interpret results, 3. implement methods, 4. generate synthetic data sets, 5. formulate mathematical claims, provide critical ingredients for proving mathematical claims, 6. assist in the writing of proofs, 7. propose or refine hypotheses, 8. clean and reformat datasets do not apply to this work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Anton Altenbernd, Philipp Wiesner, and Odej Kao. Exploring silent data corruption as a reliability challenge in llm training. *arXiv preprint arXiv:2604.00726*, 2026.

Philipp Becker, Niklas Freymuth, Serge Thilges, Fabian Otto, and Gerhard Neumann. Troll: Trust regions improve reinforcement learning for large language models. *arXiv preprint arXiv:2510.03817*, 2025.

Xin-Qiang Cai and Masashi Sugiyama. Vi-curl: Stabilizing verifier-independent rl reasoning via confidence-guided variance reduction. *arXiv preprint arXiv:2602.12579*, 2026.

Naci Cankaya. Bit-exact ai inference verification without performance tradeoffs. *arXiv preprint arXiv:2606.00279*, 2026.

Yuxiang Chen, Dingli Liang, Yihang Chen, Ziqin Gong, Chenyang Le, Zhaokai Wang, Jiachen Zhu, Lingyu Yang, Jianghao Lin, Weinan Zhang, and Jun Wang. Holder policy optimisation. *arXiv preprint arXiv:2605.12058*, 2026.

Taco Cohen, David W Zhang, Kunhao Zheng, Yunhao Tang, Remi Munos, and Gabriel Synnaeve. Soft policy optimization: Online off-policy rl for sequence models. *arXiv preprint arXiv:2503.05453*, 2025.

Fei Ding, Yongkang Zhang, Yeling Peng, Youwei Wang, Guoxiong Zhou, and Zijian Zeng. Reducing credit assignment variance via counterfactual reasoning paths. *arXiv preprint arXiv:2605.16302*, 2026.

Shihan Dou, Muling Wu, Jingwen Xu, Rui Zheng, Tao Gui, and Qi Zhang. Improving rl exploration for llm reasoning through retrospective replay. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 594–606. Springer, 2025.

Madeleine Dwyer, Adam Sobey, and Adriane Chapman. It's not you, it's clipping: A soft trust-region via probability smoothing for llm rl. *arXiv preprint arXiv:2509.21282*, 2025.

Rasool Fakoor, Murdock Aubry, Nicholas Stranges, and Alexander J Smola. Trust the batch, on-or off-policy: Adaptive policy optimization for rl post-training. *arXiv preprint arXiv:2605.12380*, 2026.

Yangyi Fang, Jiaye Lin, Xiaoliang Fu, Cong Qin, Haolin Shi, Chaowen Hu, Lu Pan, Ke Zeng, and Xunliang Cai. How to allocate, how to learn? dynamic rollout allocation and advantage modulation for policy optimization. *arXiv preprint arXiv:2602.19208*, 2026.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

Haowen Gao, Zhenyu Zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, Shaoguo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. Diva-grpo: Enhancing multimodal reasoning through difficulty-adaptive variant advantage. *arXiv preprint arXiv:2603.01106*, 2026.

Souvik Ghosh. Non-normal spectral signatures of instability in neural network training dynamics. *arXiv preprint arXiv:2605.23476*, 2026.

Svetlana Glazyrina, Maksim Kryzhanovskiy, and Roman Ischenko. Soft sequence policy optimization. *arXiv preprint arXiv:2602.19327*, 2026.

Raja Gond, Aditya K Kamath, Ramachandran Ramjee, and Ashish Panwar. Llm-42: Enabling determinism in llm inference with verified speculation. *arXiv preprint arXiv:2601.17768*, 2026.

Xue Gong, Qi Yi, Ziyuan Nan, Guanhua Huang, Kejiao Li, Yuhao Jiang, Ruibin Xiong, Zenan Xu, Jiaming Guo, Shaohui Peng, et al. Segmental advantage estimation: Enhancing ppo for long-context llm training. *arXiv preprint arXiv:2601.07320*, 2026.

Hao Gu, Hao Wang, Jiacheng Liu, Lujun Li, Qiyuan Zhu, Bei Liu, Binxing Xu, Lei Wang, Xintong Yang, Sida Lin, et al. Qarl: Rollout-aligned quantization-aware rl for fast and stable training under training–inference mismatch. *arXiv preprint arXiv:2604.07853*, 2026.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. Rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Jia Guo, Yan Sun, Zhenyu Huang, Zihao Wang, Zujie Wen, Zhiqiang Zhang, Jun Zhou, and Stanley Kok. Kpop: Taming training–inference mismatch in reinforcement learning with adaptive masking regions, May 2026. URL `https://ringtech.notion.site/kpop`.

Kevin Han, Yuhang Zhou, Mingze Gao, Gedi Zhou, Serena Li, Abhishek Kumar, Xiangjun Fan, Weiwei Li, and Lizhu Zhang. Ebpo: Empirical bayes shrinkage for stabilizing group-relative policy optimization. *arXiv preprint arXiv:2602.05165*, 2026.

Chuzhan Hao, Wenfeng Feng, Guochao Jiang, Guofeng Quan, Guohua Liu, and Yuewei Zhang. Beyond stochastic exploration: What makes training data valuable for agentic search. *arXiv preprint arXiv:2604.08124*, 2026.

Yaru Hao, Li Dong, Xun Wu, Shaohan Huang, Zewen Chi, and Furu Wei. On-policy rl with optimal reward baseline. *arXiv preprint arXiv:2505.23585*, 2025.

Horace He and Thinking Machines Lab. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250910. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *Advances in Neural Information Processing Systems*, 38:162239–162262, 2026.

Guanhua Huang, Tingqiang Xu, Jinbo Wang, Guangming Sheng, Siheng Li, Evander Yang, Kejiao Li, Yunxiang Li, Zenan Xu, Qi Yi, Xue Gong, Ziyuan Nan, Yuhao Jiang, Chenchen Zhang, Taiqiang Wu, Feiyuan Zhang, Junhao Wang, Bo Zhou, Alex Chen, Di Wang, and Shunyu Yao. Stabilizing rlvr via token-level gradient diagnosis and layerwise clipping, 2026a. URL `https://hy.tencent.com/research/100015`.

Kexin Huang, Haoming Meng, Junkang Wu, Jinda Lu, Chiyu Ma, Ziqian Chen, Xue Wang, Bolin Ding, Jiancan Wu, Xiang Wang, et al. Beyond magnitude: Leveraging direction of rlvr updates for llm reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Tianjin Huang, Zhangyang Wang, Haotian Hu, Zhenyu Zhang, Gaojie Jin, Xiang Li, Li Shen, Jiaxing Shang, Tianlong Chen, Ke Li, Lu Liu, Qingsong Wen, and Shiwei Liu. Gradientstabilizer: Fix the norm, not the gradient. *arXiv preprint arXiv:2502.17055*, 2025a.

Zixun Huang, Jiayi Sheng, and Zeyu Zheng. Oblr-po: A theoretical framework for stable reinforcement learning. *arXiv preprint arXiv:2511.23310*, 2025b.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. Vcrl: Variance-based curriculum reinforcement learning for large language models. *arXiv preprint arXiv:2509.19803*, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Batuhan K Karaman, Aditya Rawal, Suhaila Shakiah, Mohammad Ghavamzadeh, Mingyi Hong, Arijit Biswas, and Ruida Zhou. Dispo: Enhancing training efficiency and stability in reinforcement learning for large language model mathematical reasoning. *arXiv preprint arXiv:2602.00983*, 2026.

Dimitris Karydas, Dimosthenis Margaritis, and Helen C Leligou. Training methods for large language models: Current approaches and challenges. *Technologies*, 14(2):133, 2026.

Sunghwan Kim, Junhee Cho, Beong-woo Kwak, Taeyoon Kwon, Liang Wang, Nan Yang, Xingxing Zhang, Furu Wei, and Jinyoung Yeo. On training large language models for long-horizon tasks: An empirical study of horizon length. *arXiv preprint arXiv:2605.02572*, 2026.

Abhay Kumar, Louis Owen, Nilabhra Roy Chowdhury, and Fabian Güra. Zclip: Adaptive spike mitigation for llm pre-training. *arXiv preprint arXiv:2504.02507*, 2025a.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025b.

Shiye Lei, Zhihao Cheng, and Dacheng Tao. A step back: Prefix importance ratio stabilizes policy optimization. *arXiv preprint arXiv:2601.22718*, 2026.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, et al. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *arXiv preprint arXiv:2509.21268*, 2025.

Siheng Li, Zhanhui Zhou, Wai Lam, Chao Yang, and Chaochao Lu. Repo: Replay-enhanced policy optimization. *arXiv preprint arXiv:2506.09340*, 2025a.

Yingru Li, Jiacai Liu, Jiawei Xu, Yuxuan Tong, Ziniu Li, Qian Liu, and Baoxiang Wang. Trust region masking for long-horizon llm reinforcement learning. *arXiv preprint arXiv:2512.23075*, 2025b.

Yingru Li, Jiawei Xu, Ziniu Li, Jiacai Liu, Wei Liu, Yuxuan Tong, Longtao Zheng, Zhenghai Xue, Yaxiang Zhang, Tianle Cai, et al. The optimal token baseline: Variance reduction for long-horizon llm-rl. *arXiv preprint arXiv:2602.07078*, 2026a.

Yuan Li, Bo Wang, Yufei Gao, Yuqian Yao, Xinyuan Wang, Zhangyue Yin, and Xipeng Qiu. Bandpo: Bridging trust regions and ratio clipping via probability-aware bounds for llm reinforcement learning. *arXiv preprint arXiv:2603.04918*, 2026b.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Zhiyuan Li, Yuan Wu, and Yi Chang. Aggc: Adaptive group gradient clipping for stabilizing large language model training. *arXiv preprint arXiv:2601.11864*, 2026c.

Shiqi Liu, Zeyu He, Guojian Zhan, Letian Tao, Zhilong Zheng, Jiang Wu, Yinuo Wang, Yang Guan, Kehua Sheng, Bo Zhang, et al. Stapo: Stabilizing reinforcement learning for llms by silencing rare spurious tokens. *arXiv preprint arXiv:2602.15620*, 2026a.

Zexiang Liu, Xianglong He, and Yangguang Li. Advances in grpo for generation models: A survey. *arXiv preprint arXiv:2603.06623*, 2026b.

Ziru Liu, Cheng Gong, Xinyu Fu, Yaofang Liu, Ran Chen, Shoubo Hu, Suiyun Zhang, Rui Liu, Qingfu Zhang, and Dandan Tu. Ghpo: Adaptive guidance for stable and efficient llm reinforcement learning. *arXiv preprint arXiv:2507.10628*, 2025.

Yu Luo, Shuo Han, Yihan Hu, Lei Lv, Huaping Liu, Fuchun Sun, Jianye Hao, and Dong Li. Ratio-variance regularized policy optimization. *arXiv preprint arXiv:2605.26784*, 2026.

Luckeciano C Melo, Alessandro Abate, and Yarin Gal. Stabilizing policy gradients for sample-efficient reinforcement learning in llm reasoning. *arXiv preprint arXiv:2510.00819*, 2025.

Minjae Oh, Yunho Choi, Dongmin Choi, and Yohan Jo. Future policy approximation for offline reinforcement learning improves mathematical reasoning. *arXiv preprint arXiv:2509.19893*, 2026.

Penghui Qi, Zichen Liu, Xiangxin Zhou, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Defeating the training-inference mismatch via fp16. *arXiv preprint arXiv:2510.26788*, 2025.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in llm reinforcement learning. *arXiv preprint arXiv:2602.04879*, 2026.

Haiquan Qiu and Quanming Yao. Why low-precision transformer training fails: An analysis on flash attention. *arXiv preprint arXiv:2510.04212*, 2025.

Nicolas Le Roux, Marc G Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy reinforce: Stable and efficient reinforcement learning for llms. *arXiv preprint arXiv:2503.14286*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yuepeng Sheng, Yuwei Huang, Shuman Liu, Anxiang Zeng, and Haibo Zhang. Espo: Entropy importance sampling policy optimization. *arXiv preprint arXiv:2512.00499*, 2025.

Joe Suk and Yaqi Duan. On the optimization dynamics of rlvr: Gradient gap and step size thresholds. *arXiv preprint arXiv:2510.08539*, 2025.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike no more: Stabilizing the pre-training of large language models. *arXiv preprint arXiv:2312.16903*, 2023.

SGLang Team. Towards deterministic inference in sglang and reproducible rl training. *LMSYS Org blog*, 2025.

Minghao Tian, Yunfei Xie, and Chen Wei. How off-policy can grpo be? mu-grpo for efficient llm reinforcement learning. *arXiv preprint arXiv:2605.17570*, 2026.

Hongjun Wang, Wei Liu, Weibo Gu, Xing Sun, and Kai Han. Mhpo: Modulated hazard-aware policy optimization for stable reinforcement learning. *arXiv preprint arXiv:2603.16929*, 2026a.

Jiakang Wang, Runze Liu, Qingpeng Cai, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, Kun Gai, and Ling Pan. When importance sampling misallocates credit: Asymmetric ratios for outcome-supervised rl. *arXiv preprint arXiv:2510.06062*, 2026b.

Jiawei Wang, Jiacai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265*, 2025.

Yichuan Wang and the TorchTitan team. Defending against the training–inference numeric mismatch in rl (especially linear attention) — and whether it helps async rl, August 2026. URL `https://yichuan-w.github.io/blog/GDN-train-inference-mismatch-asyncRL/`.

Zhepei Wei, Xinyu Zhu, Wei-Lin Chen, Chengsong Huang, Jiaxin Huang, and Yu Meng. You only need minimal rlvr training: Extrapolating llms via rank-1 trajectories. *arXiv preprint arXiv:2605.21468*, 2026.

Qingyuan Wu, Yuhui Wang, Simon Sinong Zhan, Yanning Dai, Shilong Deng, Sarra Habchi, Qi Zhu, Matthias Gallé, and Chao Huang. A unified framework for rethinking policy divergence measures in grpo. *arXiv preprint arXiv:2602.05494*, 2026.

Weijia Wu, Chen Gao, Joya Chen, Kevin Qinghong Lin, Qingwei Meng, Yiming Zhang, Yuke Qiu, Hong Zhou, and Mike Zheng Shou. Reinforcement learning for large model: A survey. *arXiv preprint arXiv:2508.08189*, 2025.

Haocheng Xi, Charlie Ruan, Peiyuan Liao, Yujun Lin, Han Cai, Yilong Zhao, Shuo Yang, Kurt Keutzer, Song Han, and Ligeng Zhu. Jet-rl: Enabling on-policy fp8 reinforcement learning with unified training and rollout precision flow. *arXiv preprint arXiv:2601.14243*, 2026.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, et al. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*, 2025.

Peichen Xie, Yang Wang, Fan Yang, and Mao Yang. Mma-sim: Bit-accurate reference model of tensor cores and matrix cores. *arXiv e-prints*, pp. arXiv–2511, 2025.

Haofeng Xu, Junwei Su, Yukun Tian, Lansong Diao, Zhengping Qian, and Chuan Wu. Gac: Stabilizing asynchronous rl training for llms via gradient alignment control. *arXiv preprint arXiv:2603.01501*, 2026.

Xingcheng Xu. The policy cliff: A theoretical analysis of reward-policy maps in large language models. *arXiv preprint arXiv:2507.20150*, 2025.

Kai Yang, Xin Xu, Yangkun Chen, Weijie Liu, Jiafei Lyu, Zichuan Lin, Deheng Ye, and Saiyong Yang. Entropic: Towards stable long-term training of llms via entropy stabilization with proportional-integral control. *arXiv preprint arXiv:2511.15248*, 2025a.

Shuo Yang, Jinda Lu, Kexin Huang, Chiyu Ma, Shaohang Wei, Yuyang Liu, Guoyin Wang, Jingren Zhou, and Li Yuan. One-way policy optimization for self-evolving llms. *arXiv preprint arXiv:2605.22156*, 2026a.

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. Do not let low-probability tokens over-dominate in rl for llms. *arXiv preprint arXiv:2505.12929*, 2025b.

Ziteng Yang, Nicholas J Riasanovsky, Warren Deng, and Vivek Sarkar. Taming bitwise behavior in gpu kernels with tensor core: Black-box reconstruction, compiler enforcement, and static verification. *arXiv preprint arXiv:2609.11356*, 2026b.

Jiarui Yao, Xiangxin Zhou, Penghui Qi, Wee Sun Lee, Liefeng Bo, and Tianyu Pang. Rethinking the divergence regularization in llm rl. *arXiv preprint arXiv:2606.09821*, 2026.

Chenlu Ye, Xuanchang Zhang, Yifan Hao, Zhou Yu, Ziji Zhang, Abhinav Gullapalli, Hao Chen, Jing Huang, and Tong Zhang. Adaptive layerwise perturbation: Unifying off-policy corrections for llm rl. *arXiv preprint arXiv:2603.19470*, 2026.

Deokgyu Yoon, Hyungkyu Kang, Joongkyu Lee, Byeongchan Kim, Gyungin Shin, Sungrae Park, and Min-hwan Oh. Multi-step likelihood-ratio correction for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2605.20865*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Zhiyin Yu, Yuchen Mou, Juncheng Yan, Junyu Luo, Chunchun Chen, Xing Wei, Yunhui Liu, Hongru Sun, Yuxing Zhang, Jun Xu, et al. A survey of reinforcement learning for large language models under data scarcity: Challenges and solutions. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22823–22846, 2026b.

Runzhe Zhan, Yafu Li, Zhi Wang, Xiaoye Qu, Dongrui Liu, Jing Shao, Derek F Wong, and Yu Cheng. Exgrpo: Learning to reason from experience. *arXiv preprint arXiv:2510.02245*, 2025.

Di Zhang, Xun Wu, Shaohan Huang, Lingjie Jiang, Yaru Hao, Li Dong, Zewen Chi, Zhifang Sui, and Furu Wei. Towards stable and effective reinforcement learning for mixture-of-experts. *arXiv preprint arXiv:2510.23027*, 2025a.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025b.

Yaxiang Zhang, Yingru Li, Jiacai Liu, Jiawei Xu, Ziniu Li, Qian Liu, and Haoyuan Li. Beyond precision: Training-inference mismatch is an optimization problem and simple lr scheduling fixes it. *arXiv preprint arXiv:2602.01826*, 2026a.

Yuheng Zhang, Chenlu Ye, Shuowei Jin, Changlong Yu, Wei Xiong, Saurabh Sahu, and Nan Jiang. Rethinking importance sampling in llm policy optimization: A cumulative token perspective. *arXiv preprint arXiv:2605.07331*, 2026b.

Zhipeng Zhang, Zhenjie Yao, Kai Li, and Lei Yang. Training instability in deep learning follows low-dimensional dynamical principles. *arXiv preprint arXiv:2601.13160*, 2026c.

Ziyang Zhang, Xinheng Ding, Jiayi Yuan, Rixin Liu, Huizi Mao, Jiarong Xing, and Zirui Liu. Deterministic inference across tensor parallel sizes that eliminates training-inference mismatch. *arXiv preprint arXiv:2511.17826*, 2025c.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025.

Chujie Zheng, Kai Dang, Bowen Yu, Mingze Li, Huiqiang Jiang, Junrong Lin, Yuqiong Liu, Hao Lin, Chencan Wu, Feng Hu, et al. Stabilizing reinforcement learning with llms: Formulation and practices. *arXiv preprint arXiv:2512.01374*, 2025a.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025b.

Haizhong Zheng, Yang Zhou, Brian Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. *Advances in Neural Information Processing Systems*, 38:124321–124346, 2026.

Tianle Zhong, Neiwen Ling, Yifan Pi, Zijun Wei, Tianshu Yu, Geoffrey Fox, Peng Wu, and Xiao Yu. Diagnosing training inference mismatch in llm reinforcement learning. *arXiv preprint arXiv:2605.14220*, 2026.

Jiajun Zhou, Wei Shao, Lingchao Zheng, Yuwei Fan, and Ngai Wong. Ais: Adaptive importance sampling for quantized rl. *arXiv preprint arXiv:2605.13907*, 2026.

Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Hengli Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, et al. Flowrl: Matching reward distributions for llm reasoning, 2025. *URL https://arxiv.org/abs/2509.15207*.