Novelty Assessment Report

Paper: Escaping Policy Contraction: Contraction-Aware PPO (CaPPO) for Stable Language Model Fine-Tuning

PDF URL: https://openreview.net/pdf?id=vDlkJewkDu

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Reinforcement learning from human feedback (RLHF) with proximal policy optimization (PPO) is widely used but often yields less diverse outputs than supervised fine-tuning, suggesting an effect in which the policy's support contracts during on-policy optimization. We formalize this "policy contraction" with the Support Retention Ratio (SRR)—the share of SFT completions that retain non-negligible probability under the RL policy—and additionally track token-entropy, Kullback–Leibler (KL) divergence to the reference, and repetition. We propose Contraction-Aware PPO (CaPPO), a minimum-norm multi-gradient update that co-optimizes reward, entropy, and KL, paired with a controller that steers exploration toward a target token entropy. On HH-RLHF, Summarize-from-Feedback, and UltraFeedback with Qwen2-7B, Qwen2.5-14B, Mistral-7B-Instruct, and Llama-3-8B-Instruct, CaPPO increases win rate by 2 to 4 points over PPO and improves diversity, gaining 0.2 to 0.3 higher SRR. The gains persist under decoding sweeps and are robust to reward scaling and critic variance. Treating reward, diversity, and stability as first-class objectives, CaPPO mitigates contraction without sacrificing alignment performance.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Mitigating Policy Contraction in Reinforcement Learning from Human Feedback**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Policy Stability and Collapse Prevention**
- **Reward Model Robustness and Reliability**
- **Reward Over-Optimization Control**
- **Alignment Tax and Capability Preservation**
- **Preference Learning and Human Feedback Quality**
- **Training Efficiency and Exploration**
- **Safety and Constraint Satisfaction**
- **Theoretical Foundations and Analysis**
- **Alternative Optimization Frameworks**
- **Domain-Specific Applications and Extensions**
- ... and 2 more categories

Complete Taxonomy Tree

- Mitigating Policy Contraction in Reinforcement Learning from Human Feedback Survey Taxonomy
- Policy Stability and Collapse Prevention
  - General Policy Collapse Mitigation (3 papers)
  - [1] Overcoming policy collapse in deep reinforcement learning (S Dohare, 2023) View paper
  - [14] Weight clipping for deep continual and reinforcement learning (elsayed mohamed, 2024) View paper
  - [32] Research on Policy Stability of Reinforcement Learning in Complex Dynamic Decision-Making Environments (Zhang, 2025) View paper
  - RLHF Policy Contraction and Entropy Management ★ (3 papers)
  - [0] Escaping Policy Contraction: Contraction-Aware PPO (CaPPO) for Stable Language Model Fine-Tuning (Anon et al., 2026) View paper
  - [7] The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward (Li Long, 2025) View paper
  - [42] M-GRPO: Stabilizing Self-Supervised Reinforcement Learning for Large Language Models with Momentum-Anchored Policy Optimization (Bizhe Bai, 2025) View paper
  - Auxiliary Network and Regularization Approaches (2 papers)
  - [4] Enhancing autonomous driving policy stability through auxiliary network in reinforcement learning from human feedback (Hengcong Guo, 2025) View paper
  - [35] Reward Calibration for Continual Reinforcement Learning from Human Feedback (J Lang, 2025) View paper
- Reward Model Robustness and Reliability
  - Uncertainty-Aware Reward Modeling (3 papers)
  - [24] Towards reliable alignment: Uncertainty-aware rlhf (Debangshu Banerjee, 2024) View paper
  - [26] Uncertainty-Penalized Reinforcement Learning from Human Feedback with Diverse Reward LoRA Ensembles (Zhai, 2024) View paper
  - [31] Information-Theoretic Reward Modeling for Stable RLHF: Detecting and Mitigating Reward Hacking (Miao, 2025) View paper

  - Reward Model Noise and Filtering (4 papers)
  - [20] Robust reinforcement learning from corrupted human feedback (Bukharin, 2024) View paper
  - [29] Policy Filtration in RLHF to Fine-Tune LLM for Code Generation (Zhang, 2024) View paper
  - [45] When Human Preferences Flip: An Instance-Dependent Robust Loss for RLHF (Yi-Fan Xu, 2025) View paper
  - Off-Policy Reward Correction (1 papers)
  - [28] Off-Policy Corrected Reward Modeling for Reinforcement Learning from Human Feedback (Ackermann, 2025) View paper
- Reward Over-Optimization Control
  - Regularization-Based Over-Optimization Prevention (3 papers)
  - [12] Mitigating reward over-optimization in rlhf via behavior-supported regularization (Dai Jun-tao, 2025) View paper
  - [13] The energy loss phenomenon in rlhf: A new perspective on mitigating reward hacking (Miao, 2025) View paper
  - [27] Mitigating Reward Over-optimization in Direct Alignment Algorithms with Importance Sampling (Nguyen Phuc Minh, 2025) View paper
  - Direct Alignment Over-Optimization (1 papers)
  - [16] Copr: Continual human preference learning via optimal policy regularization (Han Zhang, 2025) View paper
- Alignment Tax and Capability Preservation (2 papers)
  - [15] Mitigating the alignment tax of rlhf (Lin Yong, 2024) View paper
  - [21] Delve into PPO: Implementation matters for stable RLHF (R Zheng, 2023) View paper
- Preference Learning and Human Feedback Quality
  - Heterogeneous and Contextual Preference Modeling (3 papers)
  - [5] On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization (Xiao, 2025) View paper
  - [18] PB2: Preference Space Exploration via Population-Based Methods in Preference-Based Reinforcement Learning (Davey, 2025) View paper
  - [22] Low-Rank Contextual Reinforcement Learning from Heterogeneous Human Feedback (Lee Seong Jin, 2024) View paper
  - Strategic Feedback and Annotator Reliability (2 papers)
  - [17] Governance Challenges in Reinforcement Learning from Human Feedback: Evaluator Rationality and Reinforcement Stability (Alsagheer, 2025) View paper
  - [47] Strategyproof Reinforcement Learning from Human Feedback (Buening, 2025) View paper
  - Real-Time Feedback Integration (2 papers)
  - [30] Use of Winsome Robots for Understanding Human Feedback (UWU) (Dai, 2025) View paper
  - [33] Pref-GUIDE: Continual Policy Learning from Real-Time Human Feedback via Preference-Based Learning (Ji, 2025) View paper
- Training Efficiency and Exploration
  - Off-Policy and Asynchronous RLHF (1 papers)
  - [9] Faster, more efficient RLHF through off-policy asynchronous learning (M Noukhovitch, 2025) View paper
  - Exploration and Data Collection (1 papers)
  - [6] Towards Efficient Online Exploration for Reinforcement Learning with Human Feedback (Li Gen, 2025) View paper
  - Multi-Turn Policy Optimization (2 papers)
  - [25] Multi-turn Training with Basic Human Feedback Helps Little on LLM Reasoning (Liu Qiang, 2025) View paper
  - [38] Regressing the Relative Future: Efficient Policy Optimization for Multi-turn RLHF (Zhan WenHao, 2024) View paper
- Safety and Constraint Satisfaction (3 papers)
  - [8] Trustworthy human-ai collaboration: Reinforcement learning with human feedback and physics knowledge for safe autonomous driving (Huang Zi-lin, 2024) View paper
  - [19] Enhancing safety in reinforcement learning with human feedback via rectified policy optimization (Peng, 2024) View paper
  - [46] Directed Policy Gradient for Safe Reinforcement Learning with Human Advice (Plisnier, 2018) View paper
- Theoretical Foundations and Analysis (2 papers)
  - [2] Open problems and fundamental limitations of reinforcement learning from human feedback (Casper, 2023) View paper
  - [11] The policy cliff: A theoretical analysis of reward-policy maps in large language models (Xu, 2025) View paper
- Alternative Optimization Frameworks (2 papers)
  - [34] The Hidden Link Between RLHF and Contrastive Learning (Chen Ke-hai, 2025) View paper
  - [40] Policy-labeled Preference Learning: Is Preference Enough for RLHF? (Taehyun Cho, 2025) View paper
- Domain-Specific Applications and Extensions
  - Code Generation and Reasoning Tasks (2 papers)
  - [39] Making Qwen3 Think in Korean with Reinforcement Learning (Lee Jungyup, 2025) View paper
  - [41] A Preference-Driven Methodology for Efficient Code Generation (Y Li, 2025) View paper
  - Spiking Neural Network RL (1 papers)
  - [44] Adaptive Spiking TD3+BC for Offline-to-Online Spiking Reinforcement Learning (Xiangfei Yang, 2024) View paper
- Calibration and Output Quality (3 papers)
  - [3] The inadequacy of reinforcement learning from human feedbackâ□□radicalizing large language models via semantic vulnerabilities (Timothy R. McIntosh, 2024) View paper
  - [43] Can Reasoning Help Large Language Models Capture Human Annotator Disagreement? (Ni, 2025) View paper
  - [50] Calibrating Language Models with Adaptive Temperature Scaling (Xie, 2024) View paper
- Implementation and Practical Considerations (5 papers)
  - [10] Secrets of rlhf in large language models part i: Ppo (Zheng Rui, 2023) View paper
  - [23] Marsan at PAN 2024 TextDetox: ToxiCleanse RL and paving the way for toxicity-free online discourse (M Najafi, 2024) View paper
  - [36] Towards safe, aligned, and efficient reinforcement learning from human feedback (Daniel, 2025) View paper
  - [48] Trustworthy Reinforcement Learning for Dynamic Pricing and Large Language Model Alignment (Liu, 2025) View paper
  - [49] HIAT: Human-in-the-Loop Reinforcement Learning with Auxiliary Task (Bo Niu, 2025) View paper

Narrative

Core task: Mitigating policy contraction in reinforcement learning from human feedback. The field addresses a fundamental challenge in RLHF: as policies are optimized against learned reward models, they often exhibit pathological behaviors including mode collapse, reduced output diversity, and degraded capabilities. The taxonomy organizes research into twelve major branches. Policy Stability and

Collapse Prevention focuses on entropy regularization and diversity maintenance techniques to prevent the policy from collapsing to narrow distributions. Reward Model Robustness and Reliability examines how to build more trustworthy reward signals that resist exploitation. Reward Over-Optimization Control studies mechanisms to prevent policies from gaming imperfect reward models, while Alignment Tax and Capability Preservation investigates trade-offs between alignment objectives and model performance. Additional branches cover preference learning quality, training efficiency, safety constraints, theoretical foundations, alternative optimization frameworks beyond standard RLHF, domain-specific applications, calibration methods, and practical implementation considerations. Representative works like Policy Collapse[1] and RLHF Open Problems[2] have documented these failure modes, while methods such as Auxiliary Network Stability[4] and Preference Collapse[5] propose targeted interventions.

Several active research directions reveal key tensions in the field. One line emphasizes explicit entropy management and diversity preservation, recognizing that standard KL-penalty approaches may be insufficient when policies contract toward high-reward but narrow behaviors, as documented in Diversity Collapse[7]. Another direction explores reward model uncertainty and robustness, with works like RLHF Semantic Vulnerabilities[3] showing how policies exploit model weaknesses. CaPPO[0] sits within the policy stability branch alongside entropy-focused methods, proposing contraction-aware mechanisms to maintain policy expressiveness during optimization. Compared to neighboring approaches like M-GRPO[42], which modifies the optimization objective itself, CaPPO[0] emphasizes direct intervention in the policy update process to preserve distributional breadth. The broader challenge remains balancing alignment quality against the risk of over-optimization, with ongoing debate about whether solutions should modify rewards, constrain policy updates, or fundamentally rethink the RLHF training loop.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward

Authors: Li Long, Long Li, Jiaran Hao, Zhou Zhi-jian, Jason Klein Liu, et al. (20 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

A central paradox in fine-tuning Large Language Models (LLMs) with Reinforcement Learning with Verifiable Reward (RLVR) is the frequent degradation of multi-attempt performance (Pass@k) despite improvements in single-attempt accuracy (Pass@1). This is often accompanied by catastrophic forgetting, where models lose previously acquired skills. While various methods have been proposed, the choice and function of the divergence term have been surprisingly unexamined as a proactive solution. We argue...

Relationship Analysis

Both papers belong to the RLHF Policy Contraction and Entropy Management category, addressing support contraction and diversity loss during RLHF optimization through multi-objective approaches. They overlap in identifying policy contraction as a critical problem where probability mass concentrates on fewer completions, and both propose methods that balance reward maximization with diversity preservation. The key difference is that the original paper (CaPPO) uses a minimum-norm multi-gradient descent combining reward, entropy, and KL objectives with entropy scheduling, while the candidate paper (DPH-RL) focuses specifically on replacing the standard reverse-KL divergence with mass-covering f-divergences (forward-KL, JS-divergence) to function as a rehearsal mechanism, particularly in verifiable reward settings.

2. M-GRPO: Stabilizing Self-Supervised Reinforcement Learning for Large Language Models with Momentum-Anchored Policy Optimization

Authors: Bizhe Bai, Hongming Wu, Peng Ye, Tao Chen | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Self-supervised reinforcement learning (RL) presents a promising approach for enhancing the reasoning capabilities of Large Language Models (LLMs) without reliance on expensive human-annotated data. However, we find that existing methods suffer from a critical failure mode under long-horizon training: a "policy collapse" where performance precipitously degrades. We diagnose this instability and demonstrate that simply scaling the number of rollouts -- a common strategy to improve performance -- ...

Relationship Analysis

Both papers belong to the RLHF Policy Contraction and Entropy Management category, addressing support contraction and entropy loss during reinforcement learning optimization. They overlap in identifying policy collapse/contraction as a critical failure mode where entropy declines and probability mass concentrates on fewer outputs, and both propose multi-objective approaches that explicitly manage entropy alongside reward. The key difference is that the original paper (CaPPO) uses a minimum-norm multi-gradient descent with Pareto optimization to balance reward, entropy, and KL objectives in standard RLHF settings, while the candidate paper (M-GRPO) focuses on self-supervised RLVR without ground-truth labels and employs a momentum-anchored policy model with IQR-based trajectory filtering to stabilize training in label-free reasoning tasks.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Support Retention Ratio (SRR) metric

Description: The authors formalize policy contraction by introducing the Support Retention Ratio (SRR), which measures the fraction of supervised fine-tuning completions that retain non-negligible probability under the RL policy. This metric is independent of decoding and comparable across prompts, providing a direct way to quantify support loss during on-policy optimization.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. The invisible leash: Why rlvr may not escape its origin

URL: View paper

Brief Assessment

Invisible Leash[61] introduces 'empirical support' concepts and perplexity-based analysis for RLVR, not the SRR metric for RLHF/PPO policy contraction measurement.

2. The Invisible Leash: Why RLVR May or May Not Escape Its Origin

URL: View paper

Brief Assessment

Invisible Leash[62] introduces different support-based metrics (support retention rate, net discovery rate, support dynamic score) that measure preservation of base model solutions under RLVR, but these focus on verifiable reward settings and solution accessibility rather than the ORIGINAL paper's SRR metric for measuring policy contraction during RLHF with PPO. The contexts and objectives differ fundamentally.

Contribution 2: Contraction-Aware PPO (CaPPO) algorithm

Description: The authors propose CaPPO, a minimum-norm multi-gradient update method that treats reward, entropy, and KL divergence as peer objectives rather than using fixed scalarization. It computes parameter updates that approximate Pareto-improving steps, avoiding brittle trade-offs and ensuring progress on reward does not collapse entropy or cause uncontrolled KL drift.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Fast policy extragradient methods for competitive games with entropy regularization

URL: View paper

Brief Assessment

Policy Extragradient[64] focuses on competitive zero-sum games with entropy regularization using extragradient methods, not on multi-objective RLHF optimization balancing reward, KL divergence, and entropy as peer objectives in policy gradient methods.

2. Rethinking kl regularization in rlhf: From value estimation to gradient optimization

URL: View paper

Brief Assessment

Rethinking KL Regularization[72] focuses on the mathematical formulation and gradient properties of KL divergence terms in RLHF, not on multi-objective optimization for balancing reward, entropy, and KL as peer objectives to address policy contraction.

3. Flow density control: Generative optimization beyond entropy-regularized fine-tuning

URL: View paper

Brief Assessment

Flow Density Control[68] addresses a different problem: optimizing general utilities beyond expected rewards in generative flow/diffusion models. CaPPO focuses on multi-objective RL for language model fine-tuning with reward, entropy, and KL as peer objectives using minimum-norm multi-gradient updates.

4. Controlled decoding from language models

URL: View paper

Brief Assessment

Controlled Decoding[69] addresses controlled generation from language models using value functions for inference-time control, not multi-objective policy gradient optimization during training. The candidate focuses on inference-time add-on solutions with prefix scorers, while CaPPO is a training-time multi-gradient update method.

5. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization

URL: View paper

Brief Assessment

Potential Games[65] focuses on independent natural policy gradient methods for potential games with entropy regularization in multi-agent settings, not on multi-objective optimization for single-agent RLHF with KL divergence and entropy trade-offs as in CaPPO.

6. Fast policy learning for linear quadratic control with entropy regularization

URL: View paper

Brief Assessment

Linear Quadratic Control[66] addresses entropy-regularized linear-quadratic control problems with Gaussian policies, not the multi-objective RLHF framework for language model fine-tuning that balances reward, entropy, and KL divergence as peer objectives in CaPPO.

7. EnTRPO: trust region policy optimization method with entropy regularization

URL: View paper

Brief Assessment

EnTRPO[71] adds entropy regularization to TRPO's advantage function for cart-pole control, whereas CaPPO addresses policy contraction in RLHF by treating reward, entropy, and KL as peer objectives through minimum-norm multi-gradient updates with adaptive entropy scheduling—fundamentally different algorithmic approaches and application domains.

8. The entropy mechanism of reinforcement learning for reasoning language models

URL: View paper

Brief Assessment

Entropy Mechanism Reasoning[67] focuses on entropy collapse in RL for reasoning tasks and proposes covariance-based regularization (clip-cov, kl-cov) rather than multi-objective Pareto optimization. The candidate does not address minimum-norm multi-gradient updates or treat reward, entropy, and KL as peer objectives through Pareto-improving steps.

9. Fast global convergence of natural policy gradient methods with entropy regularization

URL: View paper

Brief Assessment

Natural Policy Gradient[63] focuses on theoretical convergence guarantees for entropy-regularized NPG in tabular MDPs with exact policy evaluation, not on multi-objective optimization balancing reward, entropy, and KL divergence in language model fine-tuning.

10. Uncertainty-aware multi-objective reinforcement learning-guided diffusion models for 3D de novo molecular design

URL: View paper

Brief Assessment

Multi-Objective Molecular Design[70] applies multi-objective RL to molecular generation with diffusion models, not language model fine-tuning. The technical domain (3D molecular design vs. text generation) and application context differ fundamentally from CaPPO's focus on policy contraction in RLHF.

Contribution 3: Entropy-scheduling controller

Description: The authors develop an adaptive controller that tracks per-token entropy and dynamically adjusts the entropy coefficient during training. This controller steers exploration toward a target token entropy, complementing the multi-objective update by stabilizing entropy and preventing policy contraction without manual tuning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Off-policy asymptotic and adaptive maximum entropy deep reinforcement learning

URL: View paper

Brief Assessment

Asymptotic Maximum Entropy[52] focuses on off-policy SAC with a meta-objective for entropy temperature tuning, while the original paper develops an on-policy PPO controller that tracks per-token entropy in language model fine-tuning. The technical contexts and application domains differ fundamentally.

2. Adaptive joint entropy reward: a mechanism to efficient exploration in reinforcement learning

URL: View paper

Brief Assessment

Adaptive Joint Entropy[60] focuses on adaptive reward fusion through time-dependent coefficients for exploration, not on tracking per-token entropy to dynamically adjust entropy coefficients during policy optimization as in the original paper's controller.

3. An information entropy-driven evolutionary algorithm based on reinforcement learning for many-objective optimization

URL: View paper

Brief Assessment

Information Entropy Evolution[51] applies information entropy to evolutionary algorithms for many-objective optimization, not to reinforcement learning policy optimization or language model fine-tuning.

4. Learning implicit credit assignment for cooperative multi-agent reinforcement learning

URL: View paper

Brief Assessment

Implicit Credit Assignment[56] proposes adaptive entropy regularization for multi-agent RL that dynamically rescales entropy gradients based on current policy stochasticity, whereas the original paper's entropy-scheduling controller tracks per-token entropy and adjusts the entropy coefficient to steer toward a target token entropy in language model fine-tuning. These are distinct mechanisms applied to different problem domains (multi-agent coordination vs. LLM alignment).

5. On entropy control in llm-rl algorithms

URL: View paper

Prior Art Analysis

Entropy Control[55] demonstrates prior work on adaptive entropy coefficient control in LLM-RL. The candidate paper proposes an adaptive entropy coefficient adjustment mechanism that monitors per-token entropy and dynamically adjusts the entropy coefficient during training, similar to the original paper's entropy-scheduling controller. Both papers address the same core problem: preventing entropy collapse during RL training by automatically adjusting entropy coefficients rather than using fixed values. The candidate explicitly states their method 'automatically adjusts entropy coefficient according to the clamped entropy value' and uses a proportional update scheme to maintain entropy within target bounds, which directly parallels the original paper's controller that 'monitors per-token entropy and dynamically adjusts the effective entropy coefficient.'

Evidence

Evidence 1 - **Rationale**: Both papers describe automatic adjustment of entropy coefficients based on monitoring entropy values during training, addressing the same fundamental problem of entropy control in RL. - **Original**: cappo introduces an entropy-scheduling controller that monitors per-token entropy and dynamically adjusts the effective entropy coefficient: injecting exploration pressure when entropy collapses and relaxing it when entropy is sufficient. - **Candidate**: the algorithm automatically adjusts entropy coefficient according to the clamped entropy value, effectively controlling the entropy-induced bias while leveraging the entropy's benefits.

Evidence 2 - **Rationale**: Both papers identify the same problem: fixed entropy coefficients are insufficient for LLM-RL training, motivating the need for adaptive entropy control mechanisms. - **Original**: entropy scheduling complements cappo by stabilizing exploration through a simple feedback controller on the entropy weight. we track the length-normalized sequence entropy - **Candidate**: for entropy-regularized rl, a constant entropy coefficient λ is often sufficient to properly control the policy entropy in robotic and games rl (mnih et al., 2016; haarnoja et al., 2018). however, we observe in figure 2 that this assumption does not necessarily hold in llm-rl training as the entropy...

6. Rediscovering entropy regularization: Adaptive coefficient unlocks its potential for llm reinforcement learning

URL: View paper

Prior Art Analysis

Adaptive Entropy Coefficient[54] demonstrates that adaptive entropy coefficient scheduling for exploration control in RL policy optimization was proposed prior to the original paper. Both papers develop controllers that dynamically adjust entropy coefficients during training by tracking per-token entropy. The candidate paper presents a comprehensive framework with three components including 'dynamic global coefficient adjustment' that adaptively adjusts entropy coefficients based on current policy entropy to maintain target entropy levels. The original paper's entropy-scheduling controller follows a similar design pattern of tracking entropy via exponential moving average and adjusting coefficients proportionally to maintain target entropy, as evidenced by nearly identical mathematical formulations and control mechanisms.

Evidence

Evidence 1 - **Rationale**: Both papers propose adaptive entropy coefficient mechanisms that monitor policy entropy and dynamically adjust coefficients to maintain target entropy levels during training, demonstrating prior work on this contribution. - **Original**: we introduce an entropy-scheduling controller that monitors per-token entropy and dynamically adjusts the effective entropy coefficient: injecting exploration pressure when entropy collapses and relaxing it when entropy is sufficient. - **Candidate**: we proposeadaptive entropy regularization (aer)as shown in figure 1, which dynamically balances exploration and exploitation through adaptive coefficients, including three components: (i)difficulty-aware coefficient allocation estimates task difficulty relative to the current policy and assigns samp...

7. An adaptive entropy-regularization framework for multi-agent reinforcement learning

URL: View paper

Brief Assessment

Adaptive Entropy Multi-Agent[53] focuses on multi-agent RL with per-agent target entropy adaptation across agents, not per-token entropy tracking in language model fine-tuning. The controller in [53] adjusts target entropy for each agent based on their contribution to joint return, whereas the original paper's controller tracks per-token entropy during LLM policy optimization to prevent policy contraction.

8. Entropy-regularized diffusion policy with q-ensembles for offline reinforcement learning

URL: View paper

Brief Assessment

Entropy Regularized Diffusion[58] uses entropy regularization in diffusion policies for offline RL with fixed or automated temperature coefficients, not an adaptive per-token entropy tracking controller that dynamically adjusts coefficients during training as in the original paper's CAPPO method for online RLHF.

9. CE-GPPO: Coordinating Entropy via Gradient-Preserving Clipping Policy Optimization in Reinforcement Learning

URL: View paper

Brief Assessment

CE-GPPO[57] addresses entropy dynamics through gradient-preserving mechanisms for clipped tokens in PPO, not through an adaptive controller that tracks per-token entropy and dynamically adjusts coefficients. The technical approaches differ fundamentally in implementation.

10. MARL-MambaContour: Unleashing Multi-Agent Deep Reinforcement Learning for Active Contour Optimization in Medical Image Segmentation

URL: View paper

Brief Assessment

MambaContour[59] proposes an entropy regularization adjustment mechanism (ERAM) for contour-based medical image segmentation, which adjusts entropy based on contour consistency rather than per-token entropy in language models. The technical context and application domain differ fundamentally from the original paper's RLHF setting.

Appendix: Text Similarity Detection

Textual similarity detection checked 24 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. The invisible leash: Why rlvr may not escape its origin

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Escaping Policy Contraction: Contraction-Aware PPO (CaPPO) for Stable Language Model Fine-Tuning View paper
- [1] Overcoming policy collapse in deep reinforcement learning View paper
- [2] Open problems and fundamental limitations of reinforcement learning from human feedback View paper
- [3] The inadequacy of reinforcement learning from human feedbackâ□□radicalizing large language models via semantic vulnerabilities View paper
- [4] Enhancing autonomous driving policy stability through auxiliary network in reinforcement learning from human feedback View paper
- [5] On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization View paper
- [6] Towards Efficient Online Exploration for Reinforcement Learning with Human Feedback View paper
- [7] The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward View paper
- [8] Trustworthy human-ai collaboration: Reinforcement learning with human feedback and physics knowledge for safe autonomous driving View paper
- [9] Faster, more efficient RLHF through off-policy asynchronous learning View paper
- [10] Secrets of rlhf in large language models part i: Ppo View paper
- [11] The policy cliff: A theoretical analysis of reward-policy maps in large language models View paper
- [12] Mitigating reward over-optimization in rlhf via behavior-supported regularization View paper
- [13] The energy loss phenomenon in rlhf: A new perspective on mitigating reward hacking View paper
- [14] Weight clipping for deep continual and reinforcement learning View paper
- [15] Mitigating the alignment tax of rlhf View paper
- [16] Copr: Continual human preference learning via optimal policy regularization View paper
- [17] Governance Challenges in Reinforcement Learning from Human Feedback: Evaluator Rationality and Reinforcement Stability View paper
- [18] PB2: Preference Space Exploration via Population-Based Methods in Preference-Based Reinforcement Learning View paper
- [19] Enhancing safety in reinforcement learning with human feedback via rectified policy optimization View paper
- [20] Robust reinforcement learning from corrupted human feedback View paper
- [21] Delve into PPO: Implementation matters for stable RLHF View paper
- [22] Low-Rank Contextual Reinforcement Learning from Heterogeneous Human Feedback View paper
- [23] Marsan at PAN 2024 TextDetox: ToxiCleanse RL and paving the way for toxicity-free online discourse View paper
- [24] Towards reliable alignment: Uncertainty-aware rlhf View paper
- [25] Multi-turn Training with Basic Human Feedback Helps Little on LLM Reasoning View paper
- [26] Uncertainty-Penalized Reinforcement Learning from Human Feedback with Diverse Reward LoRA Ensembles View paper

- [27] Mitigating Reward Over-optimization in Direct Alignment Algorithms with Importance Sampling View paper
- [28] Off-Policy Corrected Reward Modeling for Reinforcement Learning from Human Feedback View paper
- [29] Policy Filtration in RLHF to Fine-Tune LLM for Code Generation View paper
- [30] Use of Winsome Robots for Understanding Human Feedback (UWU) View paper
- [31] Information-Theoretic Reward Modeling for Stable RLHF: Detecting and Mitigating Reward Hacking View paper
- [32] Research on Policy Stability of Reinforcement Learning in Complex Dynamic Decision-Making Environments View paper
- [33] Pref-GUIDE: Continual Policy Learning from Real-Time Human Feedback via Preference-Based Learning View paper
- [34] The Hidden Link Between RLHF and Contrastive Learning View paper
- [35] Reward Calibration for Continual Reinforcement Learning from Human Feedback View paper
- [36] Towards safe, aligned, and efficient reinforcement learning from human feedback View paper
- [37] Policy Filtration for RLHF to Mitigate Noise in Reward Models View paper
- [38] Regressing the Relative Future: Efficient Policy Optimization for Multi-turn RLHF View paper
- [39] Making Qwen3 Think in Korean with Reinforcement Learning View paper
- [40] Policy-labeled Preference Learning: Is Preference Enough for RLHF? View paper
- [41] A Preference-Driven Methodology for Efficient Code Generation View paper
- [42] M-GRPO: Stabilizing Self-Supervised Reinforcement Learning for Large Language Models with Momentum-Anchored Policy Optimization View paper
- [43] Can Reasoning Help Large Language Models Capture Human Annotator Disagreement? View paper
- [44] Adaptive Spiking TD3+BC for Offline-to-Online Spiking Reinforcement Learning View paper
- [45] When Human Preferences Flip: An Instance-Dependent Robust Loss for RLHF View paper
- [46] Directed Policy Gradient for Safe Reinforcement Learning with Human Advice View paper
- [47] Strategyproof Reinforcement Learning from Human Feedback View paper
- [48] Trustworthy Reinforcement Learning for Dynamic Pricing and Large Language Model Alignment View paper
- [49] HIAT: Human-in-the-Loop Reinforcement Learning with Auxiliary Task View paper
- [50] Calibrating Language Models with Adaptive Temperature Scaling View paper
- [51] An information entropy-driven evolutionary algorithm based on reinforcement learning for many-objective optimization View paper
- [52] Off-policy asymptotic and adaptive maximum entropy deep reinforcement learning View paper
- [53] An adaptive entropy-regularization framework for multi-agent reinforcement learning View paper
- [54] Rediscovering entropy regularization: Adaptive coefficient unlocks its potential for llm reinforcement learning View paper
- [55] On entropy control in llm-rl algorithms View paper
- [56] Learning implicit credit assignment for cooperative multi-agent reinforcement learning View paper
- [57] CE-GPPO: Coordinating Entropy via Gradient-Preserving Clipping Policy Optimization in Reinforcement Learning View paper
- [58] Entropy-regularized diffusion policy with q-ensembles for offline reinforcement learning View paper
- [59] MARL-MambaContour: Unleashing Multi-Agent Deep Reinforcement Learning for Active Contour Optimization in Medical Image Segmentation View paper
- [60] Adaptive joint entropy reward: a mechanism to efficient exploration in reinforcement learning View paper
- [61] The invisible leash: Why rlvr may not escape its origin View paper
- [62] The Invisible Leash: Why RLVR May or May Not Escape Its Origin View paper
- [63] Fast global convergence of natural policy gradient methods with entropy regularization View paper
- [64] Fast policy extragradient methods for competitive games with entropy regularization View paper
- [65] Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization View paper
- [66] Fast policy learning for linear quadratic control with entropy regularization View paper
- [67] The entropy mechanism of reinforcement learning for reasoning language models View paper
- [68] Flow density control: Generative optimization beyond entropy-regularized fine-tuning View paper
- [69] Controlled decoding from language models View paper
- [70] Uncertainty-aware multi-objective reinforcement learning-guided diffusion models for 3D de novo molecular design View paper
- [71] EnTRPO: trust region policy optimization method with entropy regularization View paper
- [72] Rethinking kl regularization in rlhf: From value estimation to gradient optimization View paper